Advanced Audio Identification Using MPEG-7 Content Description

Oliver Hellmuth[1], Eric Allamanche[1], Jürgen Herre[1], Thorsten Kastner[1],
Markus Cremer[2], Wolfgang Hirsch[2]

Fraunhofer Institute for Integrated Circuits IIS-A
91058 Erlangen, Germany[1]
98693 Ilmenau, Germany[2]
hel@iis.fhg.de, alm@iis.fhg.de, hrr@iis.fhg.de, ksr@iis.fhg.de,
cre@emt.iis.fhg.de, hrh@emt.iis.fhg.de

ABSTRACT

Driven by an increasing need for characterizing multimedia material, much research effort has been spent in the field of content-based classification recently. This paper presents a system for automatic identification of audio material from a database of registered works. The system is designed to allow reliable, fast and robust detection of audio material with the resources provided by today's standard computing platforms. Based on low level signal features standardized within the MPEG-7 framework, the underlying audio fingerprint format bears the potential for worldwide interoperability. Particular attention is given to issues of robustness to common signal distortions, providing good performance not only under laboratory conditions, but also in real-world applications. Improvements in discrimination, speed of search and scalability are discussed.

1 Introduction

The ever-increasing amount of multimedia data available over networks, such as the Internet, requires efficient means of identifying and indexing content for search and retrieval. In the case of audio data, a number of solutions for attaching additional descriptive data (so-called *meta-data*, have been proposed. Examples are the CD-ID and the CD-Text formats for CD media or the ID3 tags for MP3 compressed audio files [1] [2]. In the general case, however, descriptive information is not necessarily stored together with the content on the same physical medium and thus reliable association / linking of meta-data to the audiovisual content often becomes a challenge. To address this need, means for generating unique content identifiers, sometimes referred to as "fingerprints", have appeared recently [3][4][5][6][7] . However, these approaches are proprietary and thus do not allow for any interoperability between different systems and signature databases. Furthermore, very limited knowledge is available about the performance of such systems under real-world application conditions.

In an effort to facilitate managing, indexing and retrieval of audiovisual content, the MPEG working group (ISO/IEC JTC1/SC29/WG11) started a new work item, formally called the MPEG-7 "Multimedia Content Description Interface" [8][9] around 1996. In contrast to previous successful MPEG coding standards (MPEG-1, MPEG-2, MPEG-4), this standardization process aims at providing rich and interoperable descriptions for audiovisual content. This ranges from typical "meta-data" type labels (such as title, composer, and year of recording) to complex semantic descriptions and signal-derived characteristics ("low level descriptors"), all of which are defined independently of the way the described content is coded or stored. The availability of an international standard opens a wide area of possible applications and enables worldwide interoperability which also made previous MPEG standards both attractive and popular.

This paper addresses the topic of automatic identification of audio signals, based on a database of registered works (audio signals that are known to the system) and thus is closely related to the issue of "fingerprinting", as mentioned above. Furthermore, since this application is well within the scope of the MPEG-7 framework, MPEG-7 has adopted descriptive elements supporting the functionality of robust audio identification. This paper provides some background on the corresponding MPEG-7 element which emerged from prior work investigating the suitability of low level descriptors for this task and has led to the addition of a dedicated new element for this purpose [10][11]. The core issues behind the descriptor design are described and supported by experimental results.

For purposes of experimental evaluation, this paper follows up on prior work on a system for content-based identification of audio material [11]. Besides exploiting MPEG-7 compliant low level descriptors for audio signals, improvements of the audio identification system with respect to robustness against (intended or unintended) alterations of the original audio signal have been examined. Robustness in this context refers to all types of changes / manipulations of the original audio signal which do not significantly deteriorate the subjective sound quality, as perceived by a human listener. Such distortions are to be considered part of any realistic application scenarios for audio identification systems. An extensive review of possible distortions and manipulations can be found in [11].

First, the paper will present the basic setup of the audio identification system together with the underlying principles. A brief introduction into concepts of the upcoming MPEG-7 standard is given next with emphasis on the aspects relevant to audio description and audio identification. Furthermore, the adoption of MPEG-7 elements and further significant improvements of the presented system are discussed. Experimental results using a corpus of 30,000 audio items demonstrate the overall performance and usability of the system. Finally, a summary is given and ideas for further improvement are discussed.

2 System Overview

Following a standard pattern recognition paradigm, the audio identification system presented here consists of four active components and a database, as shown in Figure 1.

First, the audio signal is converted to a standard format (monophonic signal sampled at 44.1 kHz) by the *Signal Preprocessing* stage. The subsequent *Feature Extraction* unit calculates the features on a block by block basis by means of a time-to-frequency mapping and some further computation (e.g. a psychoacoustic model in the case of the *loudness* feature). At this point, the extracted features may correspond directly to characteristics defined by the MPEG-7 audio standard (see next section for a detailed description). Subsequently, an increase in recognition performance and a decrease in data size can be achieved in the *Feature Processing* stage using transformation techniques and statistical data summarization. The resulting processed features are used as an input for both the *training phase* and the *recognition phase*.

During training phase, a *Class Generator* performs a

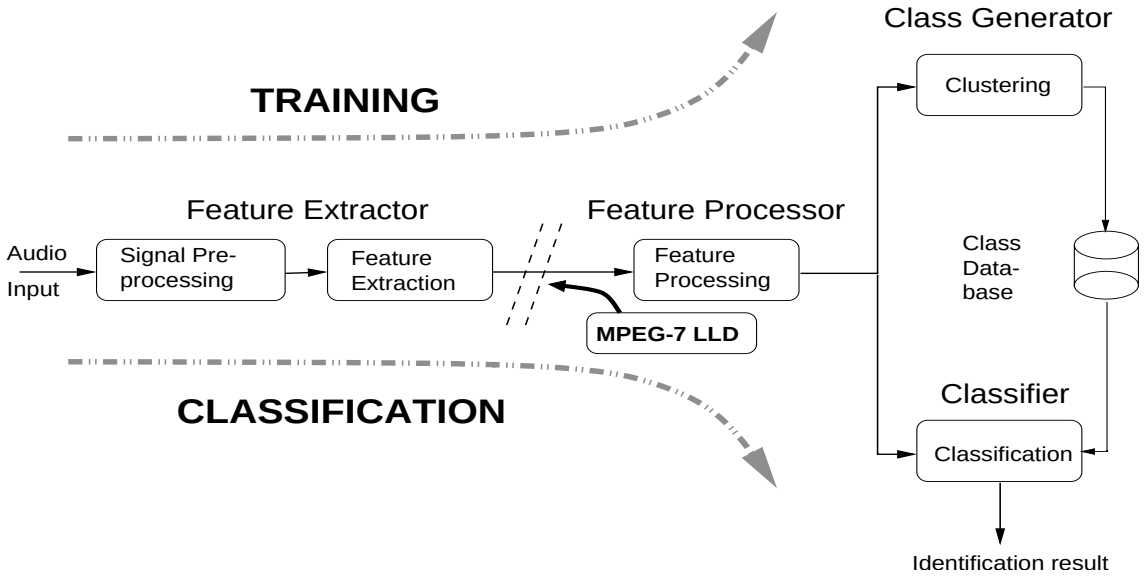

Fig. 1: Audio identification system overview

clustering of the training data (e.g. LBG vector quantization) for each training item. The resulting Vector Quantization (VQ) codebooks are then stored in a database and can be viewed as the condensed essence (fingerprint) of the reference item (class).

During recognition phase in standard mode, a VQ-based *Classifier* is used for classifying the incoming feature stream using the information from the database to find the best matching class for an unknown test excerpt. This is done by vector quantization of the incoming processed features using the trained VQ codebooks. A more sophisticated attempt is used for the advanced mode of the *Classifier*: The best matching class is determined through exploiting the temporal structure of the feature sequence. In both cases the reference item (class) with the lowest accumulated quantization error is returned.

A more extensive description of the core system can be found in [11].

3 The MPEG-7 Multimedia Content Description Interface

3.1 General MPEG-7 Concepts

Due to the ever-increasing amount of multimedia material which is available to users, efficient management of such material by means of so-called content-related techniques is of growing importance. This can be achieved by using pre-computed descriptive data ("meta-data") which is associated with the actual content, thus enabling e.g. efficient search/retrieval on a high semantic level. A particular example for the forthcoming metadata standards for audiovisual data is the MPEG-7 [8] process which is planned to be finalized in a first version in late 2001. MPEG-7 defines a wide framework for the description of audio, visual and generic properties of multimedia content, covering both high level semantic concepts as well as low level features (the latter can be extracted directly from the signal itself) [12].

- The basic descriptive entities in MPEG-7 are called Descriptors (D) and represent specific content properties or attributes by means of a defined syntax and semantics.
- Description Schemes (DS) are intended to combine components with view towards application and may comprise both Descriptors and other Description Schemes.
- Both Descriptors and Description Schemes are syntactically defined by a so-called Description Definition Language (DDL) which also provides the ability for future extension / modification of existing elements. The MPEG-7 DDL is based on XML Schema as the language of choice for the textual representation of content description and for allowing extensibility of description tools.

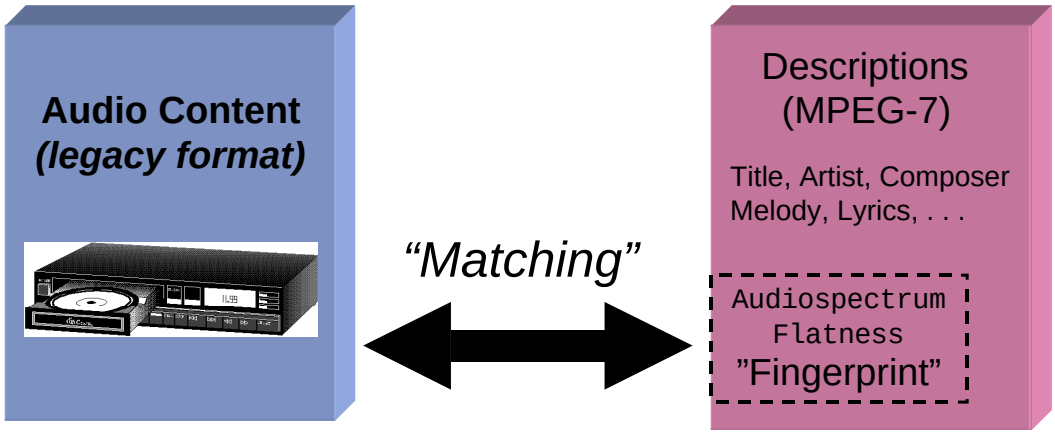

Fig. 2: Recognition of audio content using MPEG-7 descriptions

Generally, the MPEG-7 standard contains two parts defining descriptive information which is relevant to audio applications: Firstly, a major number of typical "meta-data" descriptors, such as title, composer, and year of recording are covered in the so-called Multimedia Description Scheme (MDS) part [13] of the MPEG-7 standard. This part covers also generic aspects of audiovisual descriptions, including the definition of content "Segments". Secondly, the actual audio part of the standard [14] defines descriptive data which can be extracted from the signal itself, the so-called Low Level Descriptors.

3.2 MPEG-7 Audio Low Level Descriptors (LLDs)

In the area of audio signal description, MPEG-7 Audio provides a set of Low Level Descriptors (LLDs) which are defined in terms of both syntactic format and semantics of the extraction process. While these descriptors can be considered to form a universal toolbox for many future applications, a number of concrete functionalities have already been envisaged during the development process of the standard [8]. These include "Query by humming"-type search for music, sound effects recognition, musical instrument timbre description, annotation of spoken content and robust matching of audio signals.

Specifically, the functionality of content-related identification of audio signals is supported within MPEG-7 audio by means of the AudioSpectrumFlatness low level descriptor which is designed to support robust matching of a pair of audio signals, namely the unknown signal and the known reference signal. The AudioSpectrumFlatness descriptor specifies the flatness property of the signal's power spectrum within a certain number of frequency bands, i.e. the underlying feature of the recognition system, as described previously. Using the Scalable Series concept, this data can be delivered with varying temporal granularity to achieve different tradeoffs between descriptive accuracy and compactness.

This standardized descriptor design forms the basis for achieving an open, interoperable platform for automatic audio identification:

- Identification relies on a published, open feature format rather than proprietary solutions. This allows all potential users to easily produce descriptive data for the audio works of interest (e.g. descriptions of newly released songs).
- Due to the exact standardized specification of

the descriptor, interoperability is guaranteed on a worldwide basis, i.e. every search engine relying on the MPEG-7 specification will be able to use compliant descriptions, wherever they may have been produced.

- In this sense, MPEG-7 provides a point of interoperability for these applications at the feature level. Since textual descriptions based on an XML representation are not designed to provide extremely compact representations, applications may choose to transcode the MPEG-7 compliant description into a smaller, compressed representation for storage in an internal database ("fingerprint", "signature"). Still, the "un-packed" representation will remain to be available as a point of interoperability with other schemes.

Within MPEG-7 Audio, the descriptive element for audio matching plays a double role (see Figure 2): On one hand the element is a standard element of the MPEG-7 audio toolbox, as discussed previously. At the same time, however, its matching functionality provides a bridge between MPEG-7 descriptions of audio content and the corresponding content itself which today is likely not to be available in a format including MPEG-7 meta-data. By automatic matching of legacy content with MPEG-7 compliant reference "signatures", the corresponding meta-data can be found even for legacy format content and thus is "linked" to the content.

4 Feature Extraction and Representation

4.1 Low Level Descriptors for Audio Identification

During the standardization process of the audio part of MPEG-7, several LLDs have been adopted. While each of these descriptors has been designed with specific purposes in mind (power measures, spectral envelope estimation, spectral flatness measure, ...), it was still felt desirable to follow a toolbox approach with a number of general purpose LLDs (such as the LLD for spectral envelope).

In order to support the functionality of *robust audio matching* based on signal features (see Section 2) within MPEG-7 Audio, the usability of a number of available spectral features was tested with regard to system recognition performance and robustness [15].

The tests have shown that estimated values for spectral features, such as loudness and SFM (Spectral Flatness Measure), are indeed suitable for audio identification, as described in this paper. In the evaluations, approximate loudness values were calculated from the MPEG-7 AudioSpectrumEnvelope LLD. During the evaluation phase, however, SFM has evinced to be a feature that is easy to extract with respect to algorithmic expense, yields a very high identification reliability and is significantly more robust against commonly occurring alterations of audio material than the loudness-based approach. Similarly, a feature closely related to SFM showed even slightly higher performance and thus was adopted an additional low level descriptor for the MPEG-7 Audio standard. This AudioSpectralFlatness LLD is defined as the ratio of the largest value of the power spectral density (PSD) and its average value. In analogy to the notion of the *Crest Factor*, expressing the ratio of a signal's maximum spectral magnitude and its RMS value, this feature will be referred to as *Spectral Crest Factor* (SCF) in this paper. Note that both features, SFM and SCF, express the "flatness" / "un-flatness" of a signal's spectrum and thus relate to the perceptual quality of "tone-likeness" (as opposed to noise-likeness). Nonetheless they are not claimed to be exact models of tonality, as would be used (and required) for the calculation of masking models in audio coding.

In accordance with the recommend method for extracting the MPEG-7 AudioSpectrumEnvelope descriptor, the generation of the different features for audio identification was based on a short time frequency analysis, using a windowed Discrete Fourier Transform (DFT). For an audio sampling rate of 44.1 kHz, a Hamming window of size 1323 samples and an DFT (FFT) length of 2048 values is used.

In order to increase the discriminant capabilities of an identification system, it is desirable to calculate several values for each block rather one single value. In the case of spectral features, such as SFM and other tonality-related measures, this can be achieved by dividing the frequency axis into several intervals which may or may not be overlapping. There are many possible ways of partitioning the frequency interval, but only two different rules were applied and tested here. Firstly, the most obvious configuration consists of a linear division of the target frequency range into equally spaced regions. The second configuration employs a logarithmic partitioning of the frequency range instead. This kind of partitioning mimics the natural frequency selectivity of the human middle ear to some extent [16] and is used for the computation of the MPEG-7 AudioSpectrumEnvelope descriptor.

The recognition performance values achieved with these two partitioning rules are presented and discussed in the

results section both for the N-band SFM and the N-band SCF features. In these evaluations, the logarithmic division shows a superior behavior compared to the linear one.

4.2 Quantization of features

So far, the discussion about feature extraction has mainly focused on the selection of appropriate features in order to achieve maximum robustness and recognition performances. It was assumed that these features were represented, manipulated and stored as real numbers by means of a standard floating point encoding. However, by examining the range of values of these features, it can be seen from their mathematical definitions that these ranges are well defined and bounded. For example, the SFM feature can only accept values in the interval [0,1]. The same holds for the associated short term statistics. Thus, only a fraction of the possible range of number encoding offered by floating point representations is used. On the other hand, it is interesting to investigate the required accuracy to maintain recognition performance comparable to a reference (floating point) implementation. Some preliminary investigations have been carried out using a reduced-precision floating point representation [17] and by applying a uniform quantizer with varying word length. The main benefits of reducing the precision of the representation lies in the saving of data to be stored and transmitted. See Section 5.4 for a presentation and discussion of the experiments. Please note that a compact representation of audio features in binary format is currently outside the scope of the MPEG-7 Audio standard. However, each application is free to use a custom-taylored internal representation if required. In this way, the unreduced MPEG-7 representation (based on XML) acts as a point of interoperability between different internal compact representations.

5 Experimental Results

5.1 Robustness Requirements

In order to evaluate the performance of an audio identification system, meaningful robustness requirements have to be defined. Certainly, human listeners exhibit a very high level of skill in recognizing previously trained music items, often requiring just a sound excerpt of a few seconds to identify an item. The excerpt of the song is usually recognized correctly, even if it is distorted by additive noise, coding artefacts or other distortions which do not degrade subjective sound quality by an unacceptable degree. To cover a wide variety of robustness requirements for real-world application scenarios, the following signal modifications have been carried out as a test of the recognition engine's robustness:

- Cropping. Taking only a sub-segment from the trained item. Care has to be taken to ensure that the new start offset into the item lies outside the series of offsets for the time/frequency transform used at the time of training.
- Amplitude change: Scaling the input signal by a constant factor (level change) or a slowly time varying factor (dynamic range processing)
- Resampling: Slight deviations in sampling rate (+5/-5%)
- Filtering: Linear distortion resulting from equalization, band limiting or other non-flat frequency responses of reasonable amount
- Perceptual audio coding: The effects of perceptual audio coding (within an acceptable quality range, such as 96kbps for an MPEG-1/2 Layer-3 coded stereo signal)
- Background noise: Analog or digital background noise with a reasonable SNR (e.g. 20-25 dB)
- Loudspeaker/microphone chain: The imperfections caused by acoustic playback should be tolerated under moderate acoustic conditions. This includes an A/D – D/A conversion and turned out to be one of the most challenging types of distortions.

5.2 Test Setup Description

All experimental results published in this paper were determined using a similar test setup: The test items were chosen from a database of songs of the genre rock/pop. This results in a densely populated feature space and thus a rather demanding task concerning class separation. Training was limited to 30 seconds (starting from the beginning of the item) while test items had a length of 20 seconds ("Cropping") or 30 seconds as well (all other cases). The recognition performance is characterized by a pair of numbers, where the first one stands for the percentage of items correctly identified ("Top 1"), while the second one describes the percentage for the item to be within the first ten best matches ("Top 10"). The left column of the recognition performance table lists the distortions which the test items were subjected to. Processing times are measured on a standard PC with a Pentium III CPU @ 750 MHz.

5.3 Accelerating the Classification Process

Beyond delivering high accuracy recognition results, a good recognition system should be capable of handling user requests fast and efficiently. For an audio identification system realized as an Internet service application, this could mean processing hundreds of requests per second in parallel. In order to increase efficiency, a client/server architecture is particularly useful. The server application can load all classes and retain them in memory until the next client request arrives. This avoids the initial delay of loading the reference classes every time an item needs to be classified. For huge database, this type of architecture is essential and speeds up the classification process by orders of magnitude.

The classifier calculates a distance metric between the test item and the stored reference to determine the class with the smallest error. With multidimensional features several multiplications and summations are carried out for each processed feature vector. Optimization of the core parts of the VQ algorithm greatly improves the overall speed of the system. Using byte values instead of a floating point representation enables the use of Single Instruction Multiple Data (SIMD) techniques (e.g. Intel MMX technology - available in most standard PCs today).

Table 1 shows a comparison of the recognition performance and speed between a standard (reference) classification process using floating point (32 bit) representation and an enhanced integer classification process using a byte (8 bit) representation. The test signals were both cropped and sent over a "Loudspeaker/Microphone Chain". Note that the enhanced classification provides both a higher recognition performance and decreases the process duration per item by a factor of 6-7.

5.4 Compact Representation

Many possible applications for audio identification are conceivable using small hand-held Personal Digital Assistants (e.g. Palm or iPAQ PDAs) or mobile phones. Therefore a compact representation of the extracted feature stream is desirable or even essential sometimes (e.g. GSM channel). If the classification process will be carried out on the hand-held PC, storage considerations have to be taken into account as well.

In order to further compress the feature data without unacceptable loss in recognition performance and robustness, a number of approaches are conceivable.

- Firstly, short-time statistics of the MPEG-7 features can be used to summarize the frame-based feature data. A scalable temporal granularity can be achieved using MPEG-7 constructs, such as the *Scalable Series* [14] [18].
- Furthermore, entropy coding techniques could be applied to achieve lossless compression of the feature data.
- Finally, the precision / number of bits used for representing the MPEG-7 features (see Section 4.2) can be reduced from floating point down to the actually required level.

To explore the option of reduced precision, a test (standard classification) using a database of 1000 audio items was conducted with different numerical formats for both reference classes and extracted test feature bitstream. In order to achieve significant results, two test setups were examined. With the first test, files were used that were encoded in MP3 format at 96kbps for a stereo signal. In the second test, short excerpts (20s) of the same MP3 files were used ("cropping").

The statistical summarization of the features have been subjected to a uniform quantizer with a varying number of bits. The achieved recognition performance depending on the quantizer word length is shown in Table 2. As can be seen from these numbers, a coarse quantization with word lengths from 32 bit down to 8 bit does not affect the recognition performance at all. Furthermore, the system performance still remains high when the values are quantized with only 4 bits (corresponding to 16 levels). These results clearly demonstrate that a substantial saving in disk and memory space can be achieved through an adequate numerical representation of the feature statistics.

5.5 Experimental Setup: 1,000 items

To examine the principal behaviour of an audio identification system, no large databases of items are necessary. Smaller sized databases enable efficient testing of candidate parameter settings. The best combinations can then be evaluated with larger setups. A 1000 items setup was used to determine the recognition performance while varying the number of bands and the partitioning of the frequency interval (see Section 4). The results are presented in Table 3

It can be observed that using a more detailed band representation (16 bands, feature dimension doubled) increases the performance. This results in a different tradeoff between feature size and robustness. Without a drawback a logarithmic partitioning enhanced the recognition rate by up to 25% in this test setup.

	SFM	SCF	Speed
Floating Point (32 bit)	98.0% / 99.0%	98.8% / 99.5%	2.0 seconds
Byte (8 bit)	99.7% / 99.8%	99.9% / 99.9%	0.3 seconds

Tab. 1: Comparing recognition performance and speed using byte and floating point representation

Word length in bits	32 Bit	16 Bit	10 Bit	8 Bit	6 Bit	4 Bit	3 Bit
MPEG-1/2 Layer 3 @ 96 kbit/s	96.9%	96.9%	96.8%	96.8%	95.9%	94.6%	69.5%
MPEG-1/2 Layer 3 @ 96 kbit/s & Cropping	94.3%	94.3%	94.4%	94.4%	93.2%	88.7%	53.0%

Tab. 2: Recognition performance depending on the word length of the uniform quantizer (1.000 items)

Bands:	Linear	
	8 Bands	16 Bands
Feature:	SFM	
MPEG-1/2 Layer 3 @ 96 kbit/s & Cropping	94.3% / 97.6%	94.3% / 97.5%
Loudspeaker / Microphone chain & Cropping	77.2% / 94.2%	97.2% / 99.2%
Feature:	SCF	
MPEG-1/2 Layer 3 @ 96 kbit/s & Cropping	99.2% / 99.7 %	99.0% / 99.8%
Loudspeaker / Microphone chain & Cropping	75.2% / 90.1 %	93.6% / 98.1%

Bands:	Logarithmic	
	8 Bands	16 Bands
Feature:	SFM	
MPEG-1/2 Layer 3 @ 96 kbit/s & Cropping	99.7% / 100.0%	99.6% / 100.0%
Loudspeaker / Microphone chain & Cropping	83.0% / 96.2%	99.2% / 99.7%
Feature:	SCF	
MPEG-1/2 Layer 3 @ 96 kbit/s & Cropping	99.6% / 100.0%	99.6% / 100.0%
Loudspeaker / Microphone chain & Cropping	93.9% / 99.3%	99.0% / 99.8%

Tab. 3: Recognition performance using linear and logarithmic band scaling (1.000 items)

Feature:	SFM	
Matching:	Standard	Advanced
No distortion	100.0% / 100.0%	100.0% / 100.0%
Cropping	100.0% / 100.0%	100.0% / 100.0%
MPEG-1/2 Layer 3 @ 96 kbit/s	96.1% / 97.2%	100.0% / 100.0%
MPEG-1/2 Layer 3 @ 96 kbit/s & Cropping	92.2% / 94.7%	100.0% / 100.0%

Tab. 4: Recognition performance of SFM features using standard and advanced matching (15.000 items)

Feature:	SCF	
Matching:	Standard	Advanced
No distortion	100.0% / 100.0%	100.0% / 100.0%
Cropping	100.0% / 100.0%	100.0% / 100.0%
MPEG-1/2 Layer 3 @ 96 kbit/s	99.4% / 99.6 %	100.0% / 100.0%
MPEG-1/2 Layer 3 @ 96 kbit/s & Cropping	98.8% / 99.3%	100.0% / 100.0%

Tab. 5: Recognition performance of SCF features using standard and advanced matching (15.000 items)

5.6 Standard Setup: 15,000 items

Two different classification methods are implemented in the audio identification system: standard and advanced classification (see Sec. 2). In contrast to the standard classification, the latter is designed to exploit the temporal structure of the feature stream. Especially SFM benefits most when using advanced classification.

Table 4 and Table 5 illustrate the potential for improvement by advanced classification methods by showing some results for the two features (SFM and SCF) for a setup of 15,000 items. The first column shows result values for standard, the second column for advanced classification.

5.7 Scaling Up the Database: 30,000 items

In order to be able to simulate real-world scenarios for a certain type of applications, scaling up the database of test and reference items is essential. In particular, it is necessary to test whether a reliable discrimination between the test items (classes) is still possible, even though the feature space is populated more densely with the class representatives.

Table 6 reports recognition performance for a setup with 30,000 items for SFM and SCF as the most promising features using advanced matching with accelerated classification ("byte"-representation, see above). The first robustness tests show that the system performance remains at a remarkably high level (above 99%) for SFM as well as for SCF. In order to identify the "premium" feature out of the chosen ones further testing with larger databases and other robustness tests is required. Classification time for this 30,000 items setup is 0.2 seconds per item.

5.8 Handling of Unknown Material

In real-world applications the audio identification system will be exposed to unknown music items, i.e., items which are not registered in the database. Even in such cases a reasonable result should be presented to the user, i.e. the recognition system needs to determine that an unknown item has been presented. This can be achieved by determining a suitable threshold criterion.

In the VQ process, a distance metric (e.g. Root Mean Squared Error, RMSE) between the test and the reference items is calculated, smaller distance values indicating a higher probability for correct identification. Thus, especially unregistered items are likely to cause higher errors than most of the trained items. Thus, recognition of untrained material can be achieved by finding a threshold for the distance metric beyond which an item is assumed to be outside the set of trained items. The goal is to reject as many unknown items as possible while rejecting as few registered items as possible.

A statistical analysis of the distance metric of registered items leads to the so-called false rejection rate (FRR) which describes the percentage of the items with a distance metric exceeding a certain threshold.

The so-called false acceptance rate (FAR) is determined by an analysis of the distance metric of unknown items. In contrast to FRR, the FAR describes the percentage of the unknown items with a distance metric smaller than a certain threshold.

The rate where FRR is equal to FAR is known as the *Equal Error Rate* (EER). The EER describes the quality of the system and is a possibility to compare the performance of different recognition systems. The EER can be used to determine a threshold criterion to distinguish between trained and untrained test items. However, system requirements may suggest setting a different threshold, corresponding to a different trade-off between higher safety against false classification or higher recognition rates.

An experimental setup of 15,000 items was examined. First, classification results were calculated using trained items. In contrast to this, the second test was carried out with untrained items. In both scenarios, all test items were distorted (using"loudspeaker/microphone chain"

Feature:	SFM	SCF
No distortion	100.0% / 100.0%	100.0% / 100.0%
Cropping	100.0% / 100.0%	100.0% / 100.0%
MPEG-1/2 Layer 3 @ 96 kbit/s	100.0% / 100.0%	100.0% / 100.0%
MPEG-1/2 Layer 3 @ 96 kbit/s & Cropping	99.9% / 100.0%	100.0% / 100.0%
Loudspeaker / Microphone chain	99.8% / 99.9%	99.9% / 100.0%
Loudspeaker / Microphone chain & Cropping	99.7% / 99.8%	99.9% / 99.9%

Tab. 6: Recognition performance of SFM and SCF features (30,000 items)

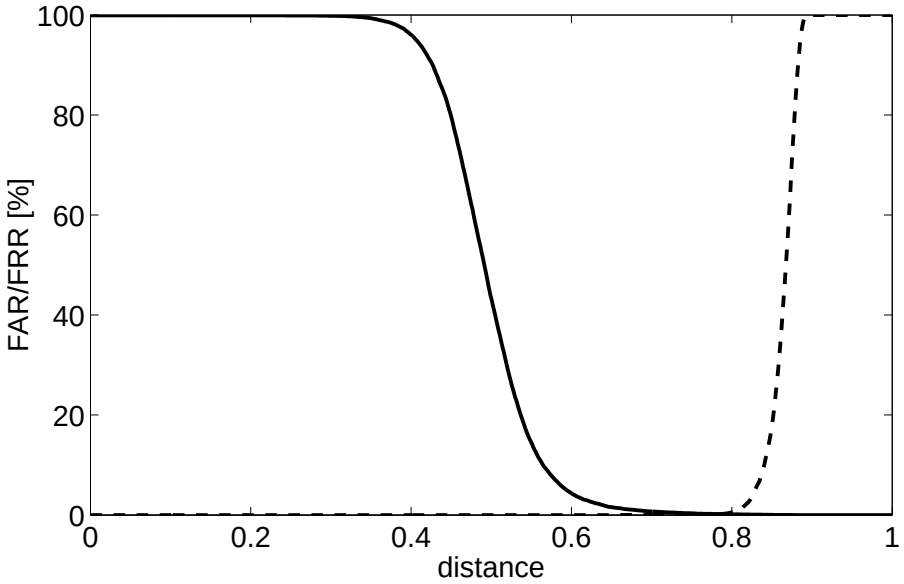

Fig. 3: FAR(dashed) and FRR(solid) curves for a 15,000 item setup

and "cropping" distortions).

In Fig. 3 the solid curve shows the FRR of the registered items, the dashed curve the FAR of the unknown items. The arrangement of the curves describes the capability of the system to separate trained from untrained items: A deep and wide "valley" formed by the two curves means high separability. It can be clearly observed that the audio identification system fulfills this criterion.

Without using a threshold criterion, the system achieves a recognition rate of 99.9% for trained items, not rejecting any untrained items. Introducing a threshold at the EER point (0.79) only 0.2% of the untrained items are assigned to registered items while the recognition rate decreases slightly to 99.7%. In order to reject all unknown test items, the threshold criterion has to be changed to 0.75. Then the recognition rate almost unnoticeably decreases to 99.6%.

6 Conclusions

This paper discussed the topic of automatic identification of audio signals based on MPEG-7 audio content description. From a basic point of view, this task can be addressed using generic pattern recognition technologies, including feature extraction and classification techniques. Naturally, the performance of such a system crucially depends on the choice of the underlying features which are employed for extracting the essence of the audio signal in a robust way. The MPEG-7 standard is the first among a number of upcoming "meta-data" standards to provide tools for this task ("robust matching of audio signals"). Specifically, a feature-based descriptor is defined which captures the audio signal's spectral unflatness property and thus relates to its "tone-likeness". The feature was shown to provide excellent recognition performance even for degraded signals. The influence of the design parameters of this MPEG-7 descriptor on the recognition performance was explored.

Finally, experimental results for recognition of larger sets of music material were presented to illustrate the potential of the technology. Particular emphasis was given to aspects of robust recognition under a wide range of common real-world signal conditions, including distortions, such as acoustic transmission between a loudspeaker and a microphone. Optimized classification schemes allow for a fast classification on today's standard computing platforms (100x faster than real-time on a standard PC (with a Pentium III CPU @ 750 MHz) for a database size of 30,000 items).

Due to the exact specification of the underlying descriptor in an open standard, all potential users can easily produce descriptive data for the audio works of interest (e.g. descriptions of newly released songs). Furthermore, worldwide interoperability is guaranteed for all compliant audio identification systems / MPEG-7 search engines, wherever they may have been produced. While the standardized representation is not designed to provide extremely compact representations, applications are free to transcode the MPEG-7 compliant description into custom-tailored compressed representations for storage in an internal database ("fingerprint", "signature"). Still, the equivalent "un-packed" MPEG-7 representation provides a point of interoperability between all compliant schemes.

A major number of applications for this technology is anticipated to benefit from this open technology, including the linking of meta-data to unknown audio content, broadcast monitoring, searching for specific audio content (e.g. on the Internet) and music sales (find unknown songs by taking a sound fingerprint of the song and querying a terminal at the music store).

References

[1] Red Book. Philips, Sony, May 1999. http://www.licensing.philips.com/cdsystems.

[2] S. Hacker. *MP3: The Definitive Guide*. O'Reilly, 2000.

[3] Frank Kurth and Michael Clausen. Full-text indexing of very large audio data bases. In *110th AES-Convention*, Amsterdam, 2001. Convention Paper 5347.

[4] Relatable homepage. http://www.relatable.com.

[5] Shazam entertainment ltd. http://www.shazam.tv.

[6] Tuneprint. robust psychoacoustic fingerprinting. http://www.tuneprint.com.

[7] etantrum. etantrum music id. http://www.etantrum.com.

[8] ISO/IEC JTC1/SC29/WG11 (MPEG). Information technology - multimedia content description interface. Final Committee Draft 15938, ISO/IEC, 2001.

[9] ISO/IEC JTC1/SC29/WG11 (MPEG). MPEG-7 webpage. http://www.cselt.it/mpeg/standards/mpeg-7/mpeg-7.htm.

[10] Eric Allamanche, Jürgen Herre, and Oliver Hellmuth. MPEG-7 audio low level descriptors for audio identification. Proposal 6832, ISO/IEC JTC1/SC29/WG11 (MPEG), 2001.

[11] Eric Allamanche, Jürgen Herre, Oliver Hellmuth, Bernhard Fröba, and Markus Cremer. Audioid: Towards content-based identification of audio material. In *110th AES-Convention*, Amsterdam, 2001. Convention Paper 5380.

[12] ISO/IEC JTC1/SC29/WG11 (MPEG). Introduction to MPEG-7. http://www.cselt.it/mpeg.

[13] ISO/IEC JTC1/SC29/WG11 (MPEG). Multimedia content description interface - part 5: Multimedia description schemes. Final Committee Draft 15938-5, ISO/IEC, 2001.

[14] ISO/IEC JTC1/SC29/WG11 (MPEG). Multimedia content description interface - part 4: Audio. Final Committee Draft 15938-4, ISO/IEC, 2001.

[15] Jürgen Herre, Christian Uhle, and Oliver Hellmuth. Descriptors for content-based audio identification. Proposal 7120, ISO/IEC JTC1/SC29/WG11 (MPEG), 2001.

[16] Eberhard Zwicker and Hugo Fastl. *Psychoacoustics*. Springer, Berlin, 2nd edition edition, 1999.

[17] Jürgen Herre, Christian Uhle, and Wolfgang Hirsch. Compactness / precision considerations for audio llds. Proposal 7121, ISO/IEC JTC1/SC29/WG11 (MPEG), 2001.

[18] Adam Lindsay and Jürgen Herre. MPEG-7 and MPEG-7 audio: An overview. *AES*, June/July 2001.