MPEG-7 Scalable Robust Audio Fingerprinting

Thorsten Kastner[1], Eric Allamanche[1] , Jürgen Herre[1], Oliver Hellmuth[1],
Markus Cremer[2], Holger Grossmann[2]

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

ABSTRACT

Much interest has recently been received by systems for audio fingerprinting which enable automatic identification of audio content by extracting unique signatures from the signal. Requirements for such systems include robustness to a wide range of signal distortions and availability of fast search methods, even for large fingerprint databases. This paper describes the provisions of the MPEG-7 standard for audio fingerprinting which allow for interoperability of fingerprint information generated according to the open standardized specification for extraction. In particular, it discusses the ability to generate scalable fingerprints providing different trade-offs between fingerprint compactness, temporal coverage and robustness of recognition, and gives experimental results for various system configurations.

1 Introduction

Currently, the topic of "audio fingerprinting" has been receiving more and more interest. The basic idea is to identify a piece of audio content by extracting a unique signature from it. This signature, commonly called a *fingerprint* (or sometimes a "robust hash" [1]), is extracted from a set of known audio material and stored in a *fingerprint database*. Unknown content can then be identified by comparing its signature to the signatures contained in the database. Based on this simple concept of *content-based identification*, many application scenarios in different domains, such as music sales and content related information services, are emerging. As an example, a traditional application relates to comprehensive monitoring of radio broadcast content for purposes of statistical analysis or proper royalty assessment. While this was partially conducted by human listeners in the past, the steadily growing number of broadcast stations has created a need for the automation of this task. A more recent development is the convenient distribution of music via networks, such as the Internet, and has led the music industries to investigate new business models involving the use of fingerprinting technologies ([2], [3]). Moreover, other fields of service to the potential customer, such as the identification of audio over Internet ([4], [5]) or even cell phones present a promising opportunity for future businesses. The obvious demand for audio fingerprinting technology has entailed the development of a variety of systems suited for one or another of the deployment scenarios described above ([6], [7], [5], [8], [1]) and inspired research work ([9], [1], [10], [11], [12], [13]). A well-designed fingerprinting scheme has to satisfy several and somehow partially conflicting requirements which have been discussed in more detail in [14]:

- Computational complexity of the fingerprint extraction process
- Size/compactness of the fingerprint
- Robustness to alterations of the original recording and associated recognition performance
- Database size and speed/complexity of the identification process

Clearly, these aspects of system performance are of varying importance in the context of different applications. As an example, a broadcast monitoring service does not necessarily need to yield low processing demands or a search performance that is faster than real time. On the other hand, the recognition performance and the database size are key factors for this field of application. For an Internet service the extraction of the fingerprint should not create excessive computational load on the user's computer. Here the compactness of the audio fingerprint, the database size and the search speed play a very important role. For identification of audio using cell phones the enormous degradation of the audio quality, sometimes further increased by interferences or disruption of the transmission channel, poses a strong challenge to the robustness of the recognition process. Both fingerprint size and processing expenses may be, however, of minor relevance in this context. Nonetheless, database size and identification speed are crucial factors in this scenario. Little is known generally about the recognition performance and robustness or the underlying technology of proprietary systems that are available for one or the other application described previously. An algorithm suitable for all scenarios, though, would have to feature a large amount of flexibility with regard to the trade-off of performance aspects to be discussed. This goal can be achieved by designing a *scalable* fingerprinting system that can be adapted to the demands of the service. At the same time, the compatibility to and interoperability with differently scaled systems can be maintained, thus providing the possibility of exchanging fingerprint data or even whole databases across different applications. International standardization of the fingerprint format creates the necessary foundation for multiple systems to interact and share metadata information.

This paper presents the concepts behind such a universally scalable audio identification system which was designed with these goals in mind. The following sections describe the technological ideas of the first openly standardized fingerprinting framework [15] which has been defined within the context of the recent ISO/MPEG-7 Audio standard [16]. First experimental results will be given to illustrate the powerful concept of fingerprint scalability and its use in different application scenarios.

2 The MPEG-7 Audio Standard

For more than ten years, the name MPEG (Moving Pictures Experts Group) has been synonymous with successful standardization in the field of coding of audiovisual material. The well-known standards, MPEG-1, MPEG-2 and MPEG-4, have defined the state of the art in the perceptual coding of multimedia content. More recently, the MPEG standards group (ISO/IEC JTC1/SC29/WG11) has extended its traditional scope by initiating the MPEG-7 standardization process which aims to define a unified interface for description and characterization of multimedia content [17].

While MPEG-7 provides a large number of concepts that describe multimedia data in general, the audio part of the MPEG-7 standard [16] specifically contains descrip-

tive elements that characterize the underlying audio signal itself rather than merely "labelling" it with high-level tags (as is frequently associated with the name "metadata").

The MPEG-7 Audio standard comprises structures that can be divided into two classes, namely the generic audio description framework and application-specific tools [18]. The audio description framework is the basic platform ("toolbox") upon which generic descriptions and unique applications may be built for any signal. On the other hand, a number of tools are targetted at specific application domains, and thus belong to the set of the application-specific elements [19]. The latter includes tools for the following functionalities:

- Sound recognition: Enables sound-effect recognition in complex sound mixtures (e.g. gun shots, dog barks, laughter, etc.)
- Spoken content: Reliable annotation, search and retrieval of databases containing spoken content
- Musical instrument timbre: Search and retrieval, query by example, sound classification of musical instrument sounds
- Melody description: Search and retrieval according to melodic features, "Query by humming / singing / playing"
- Robust matching: Robust and efficient matching between pairs of audio signals

The tool for robust audio matching addresses the core problem of fingerprinting – how to determine that two audio items are essentially identical even though the signals may have undergone many types of linear or non-linear distortions. Efficient and robust matching between an unknown audio signal and entries in a database of reference items provides the capability of automatic identification of audio material, seeking to emulate the human ability of recognizing sound based on the listener's recollection of these items. More importantly in the context of MPEG-7, it establishes a method for locating content description data (e.g. title, artist, etc.) for a given piece of audio content which may be available only in an existing legacy data format, i.e. a Compact Disc provided without any link to the corresponding description database entry. In this sense, the provisions for robust matching ("MPEG-7 Fingerprinting") enable the use of MPEG-7 descriptions even in today's legacy content world, providing a universal media-independent mechanism for linking descriptive data ("metadata") to audio content.

3 MPEG-7 Audio Provisions for Fingerprinting

The audio subpart of the MPEG-7 standard [16] essentially provides two structual elements supporting the functionality of robust audio fingerprinting. At first, the so-called *AudioSpectrumFlatness Descriptor* defines a procedure for extracting pertinent feature information from the audio signal which has been shown to be both robust with respect to a wide range of distortions [14] and can be represented in a very compact fashion [15]. This element is part of the general MPEG-7 Audio "toolbox" of Low Level Descriptors (LLDs) and is meant to be universal in its application, just like many other descriptors in this framework (e.g. spectral envelope, temporal envelope, pitch ...). Secondly, the *AudioSignature Description Scheme*, like other MPEG-7 description schemes (DSs), describes how LLDs are used for specific applications, i.e. for audio fingerprinting in this case. More specifically, the AudioSignature element employs the AudioSpectrumFlatness descriptor as a "payload" of the fingerprint, but adds a number of instantiation requirements which ensure the interoperability of differently proportioned/scaled fingerprints. As a result, a fingerprint representation according to AudioSignature DS is scalable with regard to the following parameters:

- Temporal scope of fingerprint
- Temporal resolution of fingerprint
- Spectral coverage / bandwidth of fingerprint

The *temporal scope* of the fingerprint represents a first degree of freedom and relates to the start position and the length of the audio item for which the feature extraction is carried out. Therefore, only queries located within this interval can be matched accurately. It is, however, conceivable that a successful identification might also be achieved outside this interval due to structural similarities or repetitive musical patterns like a chorus. The choice of the start and stop positions of the fingerprint generation depends on the type of the envisaged application. As an example, a broadcast monitoring system using as little processing and memory resources as possible may choose to hold only small fingerprints corresponding to an audio segment of a few seconds duration which is taken from a characteristic part of the content.

The *temporal resolution* of the fingerprint is an important parameter which can - unlike in other systems - be set to choose the trade-off between fingerprint compactness and its descriptive power (recognition strength). According to the MPEG-7 standard in which the basic frame size is settled to 30ms, a single feature vector of AudioSpectralFlatness in different frequency bands is gen-

erated at this rate. Using the general MPEG-7 Audio concept of the *Scalable Series* [16], subsequent individual feature values are grouped ("downsampled") and the relevant statistical fields (mean and variance) are calculated. According to the standards specification, the group size may assume powers of two in the range of 2 ... 128, corresponding to a temporal granularity range between 60ms and 3.84s. The default value has been defined as 32 corresponding to a 960ms granularity. As will be seen in the context of experimental results, the concept of temporal fingerprint scalability is an extremely efficient means of balancing fingerprint compactness, computational effort for fingerprint matching and speed of identification.

The third degree of freedom regarding the scalability of audio fingerprints resides in its *spectral coverage / bandwidth*. The AudioSpectrumFlatness descriptor provides a vector of feature values with each value corresponding to a certain frequency range (band). These bands are spaced in a quarter octave fashion with slightly overlapping borders with respect to the neighboring bands. In order to introduce a degree of scalability across frequency the number of frequency bands above a fixed base frequency (250Hz) was chosen as the parameter of scalability. A typical fingerprint might contain information for 16 bands, roughly corresponding to a frequency range from 250Hz to 4kHz. According to the standards specification, a minimum number of 4 frequency bands is required. Similarly to the temporal scalability discussed in the preceding paragraph, the scalability of the fingerprint's spectral coverage allows balancing of the mentioned performance aspects. Moreover, it provides the means to adapt the fingerprint representation to the properties of the transmission channel, such as its audio bandwidth (e.g. telephone channels).

In addition to the issues of temporal resolution and spectral coverage explained above, another interesting aspect of scalabilty lies in the numeric representation of feature values. A first obvious choice is to use a floating point representation, such as the IEEE-754 floating point format. However, though not specified in the MPEG-7 standard, other coded representations are conceivable. Further analysis of the sensitivity of the relevant values to quantization demonstrated that an appropriate 8 bit representation (rather than 32 bits consumed by the floating point representation) and even less yielded the same recognition performance. While this type of representation scalability is not yet covered in the Version 1 of the recent MPEG-7 Audio standard, it carries a high potential for further data rate reduction and may be addressed in an upcoming version. Some investigations on the performance of quantized features can be found in [15].

4 Experimental Setup

This section describes the system architecture of a recognition system based on the MPEG-7 Audio framework as well as the experimental conditions chosen for the experiments reported subsequently.

4.1 System architecture

A system for robust identity matching between pairs of audio signals adheres to a general pattern recognition paradigm. Therefore, two basic modes of operation can be distinguished (see Figure 1). In the *training phase* known audio content is processed by the system. For every item a fingerprint is calculated and stored in the system's database. The *classification phase* consists of a *Robust Matching* [15] approach in which the smallest distance between the fingerprint of the unknown input signal and the ones stored in the reference database is calculated. The reference item with the smallest distance is - with a certain probability - considered as the item to be identified.

The first steps of the signal processing chain are the same for training and classification: The audio signal is converted to a standard format (monophonic PCM) in the pre-processing stage of the *Feature Extractor*. This is followed by the actual feature extraction. Using a time-to-frequency mapping and some further computation on a block by block basis, an MPEG-7 compliant fingerprint is generated (see "AudioSpectrumFlatness" LLD [16]). The fingerprint is based on the calculation of the Spectral Flatness Measure (SFM) which relates to the distinction between a more tone-like or noise-like signal quality. The task of the *Feature Processor* is to increase the recognition performance and decrease the fingerprint size by means of statistical data summarization. The processed data can again be interpreted as a fingerprint, corresponding to the MPEG-7 "Audio Signature"Description Scheme [16], see Section 2).

Based on this representation, matching between fingerprints can be attempted by means of numerous different approaches. Since the choice of the matching approach or distance metric does not affect the interoperability of different applications using such a fingerprint, this choice is beyond the scope of the MPEG-7 standardization and left to the individual technology provider. For the sake of simplicity, the matching process is illustrated based on a simple VQ/Nearest Neighbor approach in this paper (although clearly more sophisticated techniques may be used to increase both matching robustness and speed). During the training phase the *Class Generator* may perform a clustering algorithm (e.g. LBG vector

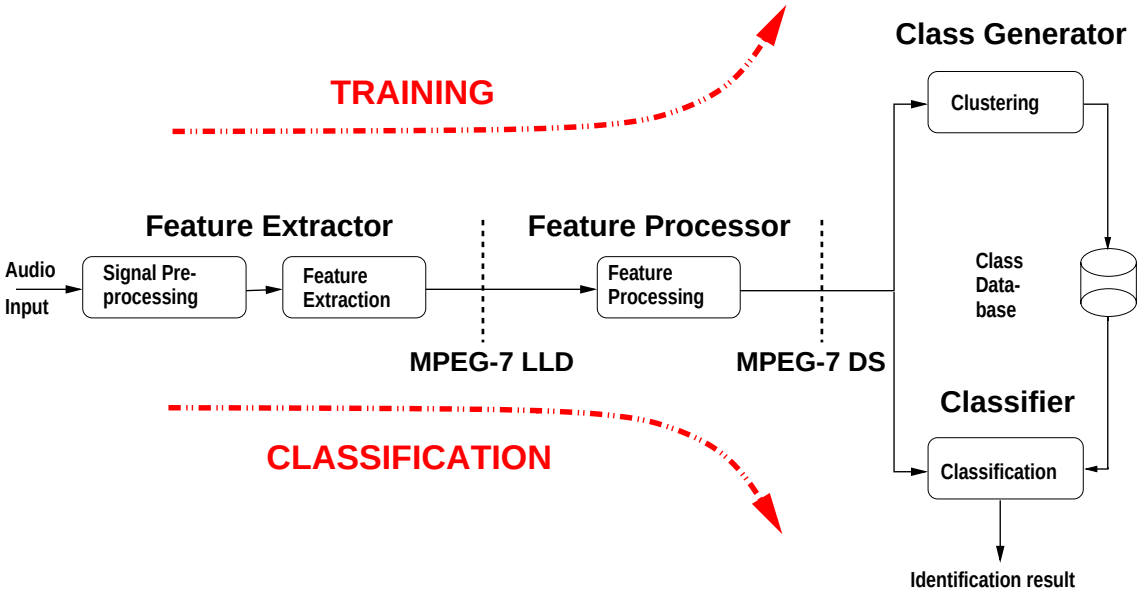

Fig. 1: Audio identification system overview

quantization) on each training item. The resulting reference codebooks are then stored in the system's class database. During the classification phase the signal at the output of the Feature Processor is compared to the codebooks stored in the reference database by the classifier. The item with the lowest distance (matching error) is presented at the system's output as a result.

4.2 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 heard 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 artifacts or other distortion types which do not degrade subjective sound quality to 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 robustness test of the recognition engine [15]. The abbreviations for the different distortions used in the result tables are parenthesized:

- Cropping. Taking only a sub-segment of the trained item for recognition. Care is 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 to emulate worst-case conditions. Cropping was used for all classification tests in addition to the other distortions.
- Amplitude change: Scaling the input signal by a certain factor, e.g. level change or a slowly time varying factor, such as dynamic range processing (DynComp)
- Resampling: Slight deviations of +5%/-5% in sampling rate. Typical for radio broadcast (Resampling)
- Filtering: Linear distortion resulted from equalization, band limiting or other non-flat frequency responses of reasonable amount (Equalizer)
- Perceptual audio coding: The effects of perceptual audio coding evaluated within an acceptable quality range, such as 96kbps Stereo (MP3@96kbs) for an MPEG-1/2 Layer-3 coded signal
- Loudspeaker/microphone chain: The imperfections caused by acoustic playback under moderate acoustic conditions. This includes an A/D – D/A conversion and turns out to be a rather challenging type of distortion (Microphone).
- Speech Codecs: Severe distortions using speech codecs, e.g. GSM Enhanced Full Rate and Full Rate codec in mobile phones (EFR, FR)

4.3 Test Setup Description

All experimental results published in this paper were determined by using a similar test setup. The test items were chosen from a database of songs of the genre rock/pop. Unless otherwise stated, the reference database for training consists of 15.000 music items, while the test database consists of a subset of 1.000 items. Training time was limited to 30 seconds (starting from the beginning of the item) while the length of the test items was variable (4s or 20s) according to the operating requirements of the different application scenarios. An advanced matching algorithm was used. The recognition performance is characterized by two numbers, while one stands for the percentage of items correctly identified ("Top1"), while the other one describes the percentage for the item to be within the first 2% ("Top2%") of the ranking list. The left column of the recognition performance table lists the distortions which the test items were subjected to. Processing times were measured on a standard PC with an Athlon CPU running at a clock rate of 1.2 GHz.

5 Experimental Results

In order to investigate the potential behind the concept of fingerprint scalability, this section presents experimental performance data for different "operating points" within the range of freedom as discussed previously. The achieved recognition rates will be reported for each choice of the scalability parameters corresponding to configurations for some typical application scenarios, including

- General purpose recognition (e.g. music on the Internet)
- Recognition of short sequences (e.g. advertisement spots)
- Very robust recognition (e.g. speech codecs)
- Efficient search of large fingerprint databases

5.1 General purpose recognition

To examine the basic behavior of the identification system, a setup for general recognition tasks was built according to the standard specifications and the default parameters provided by MPEG-7. Specifically, 16 frequency bands were used and 32 feature vectors were grouped together over time. In Table 1 recognition results of two different distortions are presented. The reference database consists of 85.000 fingerprints each of which is 30 seconds in duration. As a test database, all test items were MP3 encoded at a datarate of 96 kbit/s stereo and a subset of 15.000 items was subjected to the acoustic transmission distortion. In both cases, the duration of the test items was curtailed to 20 seconds. It can be clearly observed that the results of both distortions show excellent recognition behavior. This may serve as an indication for the high performance of a "baseline system" (i.e. using the default scalability settings provided by the standard) for standard recognition tasks.

Grouping	Top1	Top2%
MP3@96kbit	99.97%	99.98%
Microphone	99.71%	99.83%

Tab. 1: Classification results for general purpose recognition setup, test length 20s, reference database: 85.000 items

With regard to the relatively large amount of reference items in the database and the duration of the test items (20s), the standard setup covers a wide range of possible recognition tasks. Nonetheless, for certain classification tasks with specific requirements the standard system configuration may not be well-adapted to provide the necessary recognition performance. Therefore, a number of examples for how to adapt the system to specific applications will be discussed next. These adaptations make use of the concept of scalable fingerprinting and thus ensure interoperability with the "baseline system" and amongst each other.

5.2 Recognition of short sequences

Background: The ability of identifying short audio sequences is a requirement for several applications. If, for example, the occurrence of advertisement spots should be monitored, it is necessary to analyze small segments of the audiostream in order to detect the presence of very short spots (e.g. duration of 5 seconds) reliably. This may be achieved by dividing the stream into a series of short segments and classifying these sequences separately. The single classification results may then be combined by a post-processing step for final analysis of the audio stream.

Adapting the Scheme: Clearly, using only a very short excerpt of the test items for detection increases the difficulty of the recognition task. A loss in detection reliability can be expected when the duration of the test item is reduced dramatically. In Table 2 some experimental results are shown for such a scenario. In contrast to the

Grouping	4	8	16	32
MP3@96kbit	99.9%	99.8%	99.4%	96.6%
Microphone	97.5%	97.2%	96.0%	90.8%
Resampling	95.3%	96.4%	92.8%	88.6%
Equalizer	100.0%	99.3%	96.2%	69.7%
DynComp	99.9%	99.8%	99.7%	99.3%

Tab. 2: Recognition results of short audio sequences (4s), Top1, varying the temporal resolution of the fingerprint

standard setup, the duration of the 1.000 test items was reduced from 20 seconds down to only 4 seconds each, using the reference database of 15.000 items of a 30-second duration each.

The table contains results of both the "baseline" configuration (grouping of 32) and configurations with enhanced temporal resolutions (grouping of 16, 8 and 4). As it can be seen, the key for coping with the task can be found in the adaptation of the fingerprint's temporal resolution. While the performance of the baseline configuration drops significantly below the usual level and may be inadequate for some cases, a subsequent reduction of the grouping size (corresponding to a finer granularity) improves the performance again.

Thus, according to these results the benefit of the temporal scalability of the fingerprint becomes quite evident and permits to custom-tailor the amount of information needed to address a certain application. As a further note, it should be mentioned that in an actual application which monitors audio stream, each of the individual classification results is only one of many subsequent results. By combining these single results appropiately, it is possible to outperform the presented classification results significantly.

5.3 Very Robust Recognition

Background: Many of the application scenarios for content-based audio identification rely on the flexible use of the recognition engine in arbitrary environments - audio identification systems which run on small handheld devices could be used almost everywhere. A recognition system could, e.g., run completely on a "Personal Digital Assistant" (PDA) or, alternatively, the audio signal to be identified could be transmitted to a remote server through a mobile phone. For the latter scenario a very severe signal distortion has to be handled by the system. Current digital mobile phone systems use speech codecs for audio transmission which exploit the properties of the human speech generation process. As a consequence, a music signal transmitted by a speech coder will in general be dramatically degraded since the coder was not designed for this type of signal, thus increasing the difficulty of the recognition task.

Grouping	8	16	32	64
EFR	98.7%	97.4%	74.5%	44.9%
FR	99.9%	99.9%	92.0%	66.0%

Tab. 3: Classification results of GSM Enhanced Full Rate and Full Rate Coding, test length 10s, Top1

Adapting the Scheme: To find out how the system performs with these demanding signal distortions a series of tests using GSM Enhanced Full Rate (EFR) and Full Rate (FR) speech codecs (see [20]) were carried out. Table 3 shows the classification performance for a setup where the test items are heavily distorted by the GSM EFR and GSM FR speech codecs. The length of these items is 10 seconds whereas the training items last for 30 seconds. The fingerprints cover a frequency range of 8 bands. Each column shows the recognition rate for a different grouping size starting from 8 to 64. Again, it can be observed that the performance increases enormously if the scaling capability of an MPEG-7 compliant fingerprint is exploited. For the EFR test items this means a recognition rate above 98% with a high temporal resolution (scaling ratio = 8) compared to 74.5% with a standard granularity (scaling ratio = 32). The results for the FR speech codec are similar at a higher performance level.

5.4 Efficient search of large fingerprint databases

Background: If the size of the reference database increases and the number of classification requests grows, the time taken for classification will become a significant cost factor. The following section delivers an approach addressing this isssue. It aims to present a fast search which excludes 98% of the database entries and, as a result, provides only the 2% which are classified most similar to the test item ("pre-search"). In a subsequent analysis stage, only these 2% have to be taken into account in order to arrive at a final result.

Adapting the Scheme: Again, the scalability of the MPEG-7 based fingerprint offers the prerequisites to implement an efficient pre-search process. Data can be scaled over the frequency range as well as in its tempo-

No. of Bands	4	8	12	16
MP3@96kbit	100%	100%	100%	100%
Microphone	100%	99.9%	99.9%	99.9%
Resampling	99.4%	100%	100%	100%
Equalizer	97.3%	100%	100%	100%
DynComp	100%	100%	100%	100%

Tab. 4: Classification results using frequency range scaling, test length 20s, Top2%

Grouping	32	64	128	256
MP3@96kbit	100%	100%	100%	99.1%
Microphone	99.9%	99.7%	99.6%	98.7%
Resampling	100%	100%	100%	98.4%
Equalizer	100%	100%	100%	96.9%
DynComp	100%	99.9%	99.9%	99.5%

Tab. 5: Recoginition results of different temporal resolutions of the fingerprint, test length 20s, Top1

ral resolution. While it can be expected that running the matching process with lower resolution fingerprints will deteriorate the recognition accurracy, this 'low footprint' matching process can be designed to be computationally very inexpensive compared to the original process. Thus, the goal is to achieve a sufficiently accurate pre-search process with minimal computation.

5.4.1 Scalability over frequency range

To analyze how frequency scaling of the fingerprint can be used for a search of large databases, classification tests for a varying number of frequency bands were carried out. Table 4 gives some experimental results of the first 2% hit rate (i.e. the rate at which the correct item is contained within the first 2% of the best matching items) for different frequency ranges. Down to 8 frequency bands nearly all of the test items lie within the first 2% of the ranking list. When only 4 bands for classification are used, there is a slight decrease in the results noticeable for some distortions.

For most types of matching techniques it can be assumed that the time needed for the classification is approximately of direct proportion to the number of frequency bands used. Consequently, the reduction from 16 (default value) down to 4 processed bands saves computation by 4-fold. Taken into consideration the results in Table 4 and the classification time, the advantages of a scalable fingerprint, adaptable to precision and speed of the classification, are remarkable.

5.4.2 Temporal scalability

In this section the advantages of temporal scalability for speeding up the search on large databases are examined. The analysis is carried out in the same manner as in the preceding section. Results of different temporal resolutions are shown. Table 5 gives some experimental results of the first 2% hit rate (i.e. the rate for which the correct item is contained within the first 2% of the best matching items) for successive increased temporal resolution from 32 (default value) up to 256. It can be observed that there is indeed a graceful degradation when reducing temporal resolution. A 4-fold reduction (i.e. a statistical grouping of 128) still appears to deliver adequate performance for a fuzzy search. If it is known that the items to be tested are not heavily distorted an even coarser granularity may be sufficient.

Taking into account the required computational effort for a search, the benefit of temporal fingerprint scalability becomes most obvious. For many matching techniques it can be assumed that the computational complexity grows with the square of the number of data points in the reference and test fingerprints. Consequently, a resolution reduction by a factor of 2 accelerates the search 4 times. As a further example, if the finest (32) grouping in Table 5 is compared with a grouping of 128, there is a difference in computational complexity for the search of a factor of 16.

In summary, the benefits of a scalable fingerprint for searching a large database are quite evident. By means of a coarse fingerprint resolution, similar fingerprints can be searched very efficiently from the databases. This allows for a balancing between classification speed and accuracy of the classification results. Furthermore, it should be mentioned that the scaling possibilities in time and frequency can also naturally be combined in order to achieve an adapted speed-optimized search according to the operating requirements of the application.

6 Conclusions

The MPEG-7 Audio standard provides a generic framework for the descriptive annotation of audio data which is able to summarize essential features of the signal. Two of these descriptive elements, namely the AudioSpectrumFlatness Descriptor and the AudioSignature Description Scheme, are designed to extract a robust and compact-sized unique signature of an audio signal which

can be used as "fingerprint". Due to the properties of these descriptive elements the MPEG-7-based fingerprinting framework provides a number of unique aspects, including scalability of the fingerprint data. On a technical level, this is achieved by the ability to vary extraction parameters, such as temporal scope, temporal resolution and number of spectral bands. In this way, a flexible trade-off between the compactness of the fingerprint and its recognition robustness can be reached. From an application point of view this is a powerful concept which helps satisfy the needs of a wide range of applications by a single framework. Most importantly, the fingerprint representation still maintains interoperability so that fingerprints extracted for one application can still be compared to a database set up for a different purpose.

By its very nature, the standardized and open specification for the extraction method guarantees worldwide compatibility between all standards compliant applications. Fingerprints may, for example, become a standard ingredient of future content description packages, helping associate these descriptions with legacy format audio content. The experimental evaluation of the MPEG-7 audio identification framework confirms the excellent performance of a system based on the principles described above. Similar to other MPEG technologies, more optimizations will be done over time. Numerous different applications, such as broadcast monitoring, internet services or mobile audio identification services using cellular phones are currently developed. Similar to the success of previous MPEG standards, a wide deployment of this technology is anticipated in the near future.

References

[1] J. Haitsma, Ton Kalker, and J. Oostveen. Robust audio hashing for content identification. September 2001.

[2] Napster. http://www.napster.com.

[3] Recording Industry Association of America. http://www.riaa.org.

[4] Gracenote. http://www.gracenote.com.

[5] Moodlogic, Inc. http://www.moodlogic.com.

[6] Auditude. http://www.auditude.com.

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

[8] Relatable. http://www.relatable.com.

[9] C. Papaodysseus, G. Roussopoulos, D. Fragoulis, Th. Panagopoulos, and C. Alexiou. A new approach to the automatic recognition of musical recordings. Technical report, National Technical University of Athens, Department of Electrical and Computer Engineering, GR-15773 Athens, Greece, 2000.

[10] Erling Wold, Thom Blum, Douglas Keislar, and James Wheaton. Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, 3(3):27–36, 1996.

[11] Cheng Yang. Macs: Music audio characteristic sequence indexing for similarity retrieval. Technical report, Signal Processing Laboratory, Tampere University of Technology, October 2001.

[12] Helmut Neuschmied, Harald Mayer, and Eloi Batlle. Content-based identification of audio titles on the internet. *Wedelmusic 2001, Florence*, 2001.

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

[14] 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.

[15] Oliver Hellmuth, Eric Allamanche, Jürgen Herre, Thorsten Kastner, Markus Cremer, and Wolfgang Hirsch. Advanced Audio Identification using MPEG-7 Content Description. In *111th AES-Convention*, New York, 2001. Convention Paper 5463.

[16] ISO/IEC JTC1/SC29/WG11 (MPEG). Multimedia content description interface - part 4: Audio. International Standard 15938-4, ISO/IEC, 2001.

[17] ISO/IEC JTC1/SC29/WG11 (MPEG). Overview of the MPEG-7 standard. http://mpeg.telecomitalialab.com, December 2001.

[18] Adam Lindsay and Jürgen Herre. MPEG-7 and MPEG-7 Audio: An Overview. *AES*, 49(7/8):589–594, July/August 2001.

[19] ISO/IEC JTC1/SC29/WG11 (MPEG). MPEG-7 Applications Document. http://mpeg.telecomitalialab.com/, January 2001.

[20] European Telecommunications Standards Institute. 650, route des Lucioles, 06921 Sophia-Antipolis Cedex, FRANCE. http://www.etsi.org.