Content-based Identification of Audio Material Using MPEG-7 Low Level Description

Eric Allamanche
Fraunhofer IIS-A
Am Weichselgarten 3
D - 91058 Erlangen, Germany
+49 9131 776 322
alm@iis.fhg.de

Jürgen Herre
Fraunhofer IIS-A
Am Weichselgarten 3
D - 91058 Erlangen, Germany
+49 9131 776 353
hrr@iis.fhg.de

Oliver Hellmuth
Fraunhofer IIS-A
Am Weichselgarten 3
D - 91058 Erlangen, Germany
+49 9131 776 372
hel@iis.fhg.de

Bernhard Fröba
Fraunhofer IIS-A
Am Weichselgarten 3
D - 91058 Erlangen, Germany
+49 9131 776 535
bdf@iis.fhg.de

Thorsten Kastner
Fraunhofer IIS-A
Am Weichselgarten 3
D - 91058 Erlangen, Germany
+49 9131 776 348
ksr@iis.fhg.de

Markus Cremer
Fraunhofer IIS-A / AEMT
Am Ehrenberg 8
D - 98693 Ilmenau, Germany
+49 3677 69 4344
cre@emt.iis.fhg.de

ABSTRACT

Along with investigating similarity metrics between audio material, the topic of robust matching of pairs of audio content has gained wide interest recently. In particular, if this matching process is carried out using a compact representation of the audio content ("audio fingerprint"), it is possible to identify unknown audio material by means of matching it to a database with the fingerprints of registered works. This paper presents a system for reliable, fast and robust identification of audio material which can be run on the resources provided by today's standard computing platforms. The system is based on a general pattern recognition paradigm and exploits low level signal features standardized within the MPEG-7 framework, thus enabling interoperability on a world-wide scale.

Compared to similar systems, particular attention is given to issues of robustness with respect to common signal distortions, i.e. recognition performance for processed/modified audio signals. The system's current performance figures are benchmarked for a range of real-world signal distortions, including low bitrate coding and transmission over an acoustic channel. A number of interesting applications are discussed.

1. INTRODUCTION

Stimulated by the ever-growing availability of musical material to the user via new media and ways of distribution (e.g. the Internet, efficient audio compression schemes) an increasing need to identify and classify audio data has emerged. Given the enormous amount of available audio material it has become more and more difficult for the consumer to locate music that fits his or her personal tastes.

Descriptive information about audio data which is delivered together with the actual content would be one way to facilitate this search immensely. This so-called metadata ("data about data") could e.g. describe the performing artist, composer or title of the song and album, producer, date of release, etc.. Examples of de-facto and formal standards for metadata are the widely used ID3 tags attached to MP3 bitstreams [1] and the forthcoming MPEG-7 standard [2].

Another way of retrieving these information resides in the characteristics of the medium on which the audio data is comprised. This kind of services are provided by e.g. *Gracenote,* formerly *CDDB*, [3] where the *Table Of Content* (TOC) of an audio CD is compared against a vast database. Obviously, this kind of mechanism fails when the CD is a self made compilation, or when commercially not available.

A lot of different approaches have addressed the automatic analysis of audio content, be it speech/music classification [4, 5, 6], retrieval of similar sounds ("sounds like" data base search) [7, 8, 9], or music genre classification [10].

The main topic of this paper, however, is to present a system which performs an automated identification of audio signals rather than assigning them to predefined categories. The essential property of the introduced system lies in the fact that it does not rely on the availability of metadata information that is attached to the audio signal itself. It will, however, identify all incoming audio signals by means of a database of works that are known to the system. This functionality can be considered the algorithmic equivalent of human recognition of a song from the memory of the recognizing person.

This observation yields the key criteria for the performance requirements of an audio identification system. It should be able to identify the song as long as a human being is able to do so. To come as close as possible to this aim, the system should be robust against alteration commonly applied to musical material, like filtering, dynamic range processing, audio coding, and so on. Additionally, arbitrary excerpts of the music signal should be sufficient for the recognition.

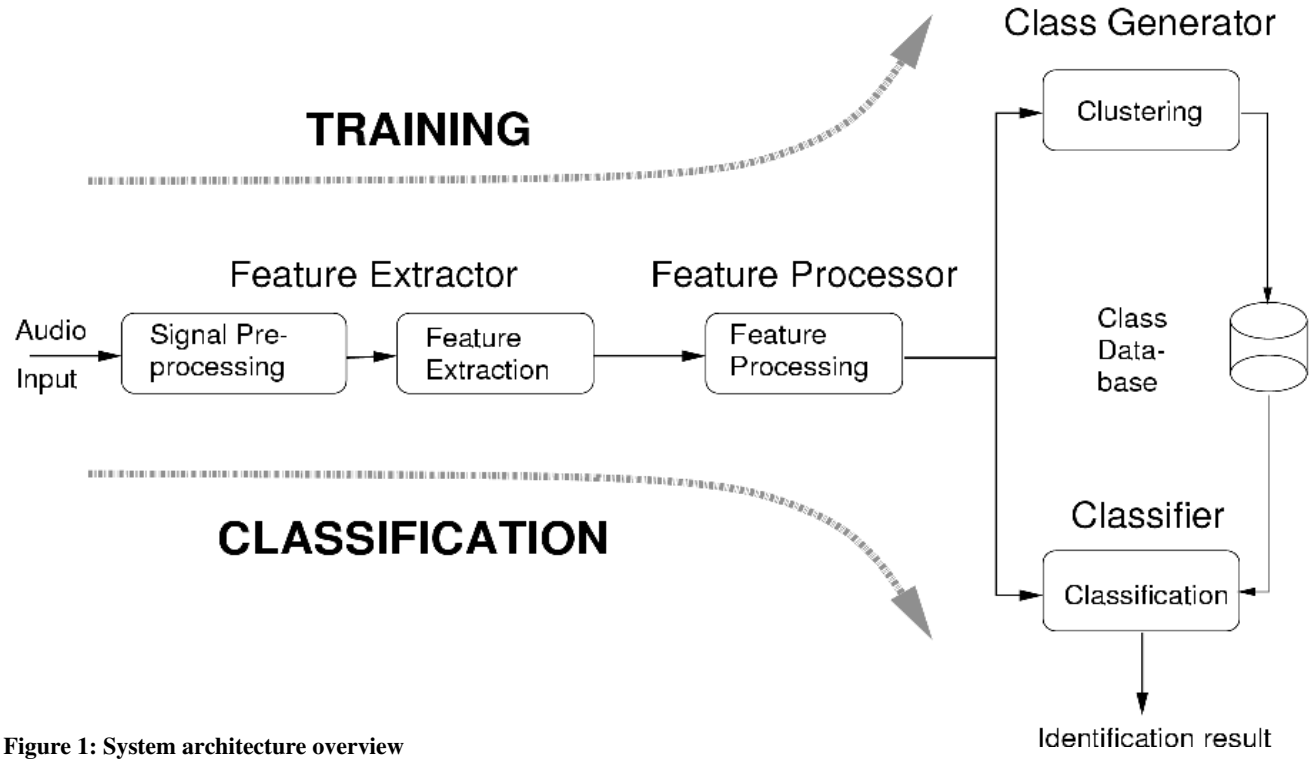

Figure 1: System architecture overview

The segment size needed for recognition by an ideal system should not be longer than a few seconds, with other words as long as it would take a human listener to identify a piece of music correctly.

On top of that the system should be able to operate with large databases of registered works while reliably discriminate between the items, and computational complexity should stay within acceptable limits ("System Scalability").

While the task of music recognition may appear easy to human listeners, lately introduced technologies definitely fall short of reaching these high goals, e.g. in terms of robustness of recognition [11] or computational complexity [12].

The system presented in this paper has been designed to meet many of the requirements mentioned above. The system's complexity is low enough to allow operation on today's personal computers and other cost-effective computing platforms and the described algorithm is based on well-known feature extraction/pattern recognition concepts [13]. It includes extraction of a set of robust features with a psychoacoustic background. The extraction process itself is based on so called *Low Level Descriptors* that will be part of the upcoming MPEG-7 standard.

In the following chapters an overview of the presented system is provided first. The architecture of the system as well as the basic underlying concepts are explained. Subsequently, the system requirements for robust recognition are discussed by identifying a suite of typical alterations of the original audio material. The influence of the audio feature selection on the recognition performance is addressed thereafter based on test results using different sets of test data. In the following two chapters potential applications of the proposed system are identified and the compliance to the upcoming MPEG-7 standard is accounted for. Finally, a conclusion section will present promising future improvements and directions for further enhancement of the overall system performance.

2. SYSTEM OVERVIEW

The audio identification system presented here follows a general pattern recognition paradigm as described in [13]. From the block diagram shown in Figure 1, two distinct operating modes can be identified, namely the *training* mode and the *classification* (recognition) mode. During training a condensed "fingerprint" of each item from the training sample is created which is used in the recognition phase to identify the item under test. In a preprocessing step a signal preprocessor converts the audio input signal into a fixed target format with predefined settings. In the present configuration, the signal is converted to a mono signal using common downmix techniques and then, if necessary, resampled to a sampling frequency of 44.1 kHz.

2.1 Feature Extraction

Feature extraction is a central processing step which has a high influence on the overall system performance. The chosen feature set should be robust under a wide class of distortions (see Section 3.2) and the computational burden should be low enough to allow for real-time calculation. In the present configuration the audio time signal is segmented by a windowing function and each window is mapped to a spectral representation by means of a DFT (Discrete Fourier Transform). A set of psychoacoustic features is extracted from the spectrum of each analysis window to form a feature vector. This vector is regarded as an elementary feature at a discrete time instant t and undergoes further processing.

The elementary features are then normalized to have component-wise unit variance. Note that no removal of the mean is necessary

prior to normalization, as suggested in [14], since only the difference between the feature vectors to be classified and the reference vectors from the "fingerprint" are considered. Through this normalization step, a balanced feature vector is generated which can be filtered optionally.

Normalized features from subsequent time steps are then grouped together to form a composite feature vector of higher dimension. In addition, the feature statistics of the single vectors are estimated.

2.2 Vector Quantization for Pattern Recognition

The identification system uses a linear classifier which is based on a compact representation of the training vectors, the above mentioned fingerprint. The classification is performed using a standard *NN* (Nearest Neighbor) rule. To obtain a compact class representation a *VQ* (Vector Quantization) algorithm is applied for training. This method approximates the training data for each class with a so-called vector codebook by minimizing a *RMSE* (Root Mean Square Error) criterion. The codebook consists of a certain number of code vectors depending on the maximum permitted RMSE. An upper limit of the number of code vectors may be specified. The VQ clustering algorithm is an iterative algorithm which approximates a set of vectors by a much lower number of representative code vectors, forming a codebook. Such a codebook is needed for each class (audio item). In Figure 2 an example of the representation of a set of 2-D feature vectors by 6 code vectors is shown.

The code vectors are obtained using a simple *k-means* clustering rule. The code vectors computed during training phase are stored in a database together with other associated descriptive information of the music items, such as title and composer of the item.

In Figure 3 the approximation error of a feature vector set is shown, depending on the number of code vectors used for the codebook. The training set can be approximated ideally if the number of code vectors reaches the number of training vectors. For distorted versions of the training vectors, on the other hand, the approximation error does not converge toward zero.

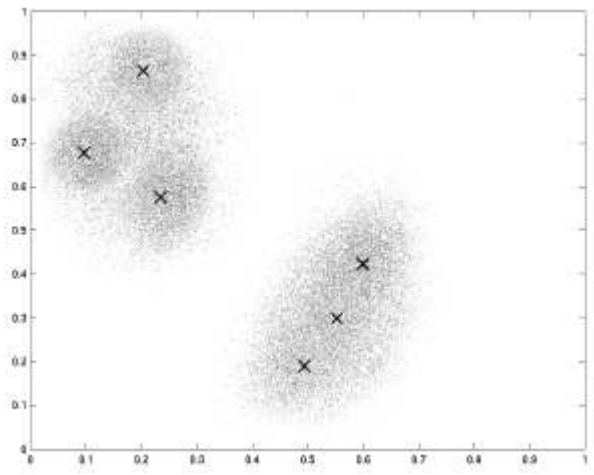

Figure 2. Example of 2-D feature set and it's approximation using 6 code vectors.

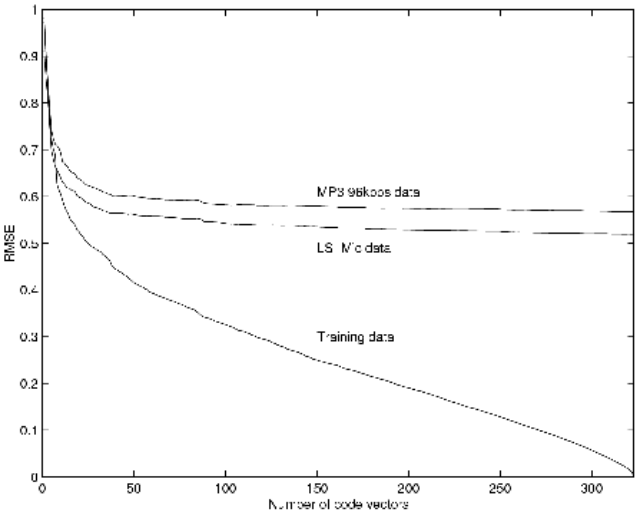

Figure 3. RMS error as a function of the number of code vectors.

2.3 Classification

The music identification task here is an *N-class* classification problem. For each of the music items in the database one class, i.e. the associated codebook, is generated. To identify an unknown music item which is included in the reference database, a sequence of feature vectors is generated from the unknown item and these features are compared to the codebooks stored in the database.

In more detail, during the identification process each vector is subsequently approximated by all stored codebooks using some standard distance metric. For each of the classes the approximation errors are accumulated and, as a result, the music item is assigned to the class which yields the smallest accumulated approximation error.

In a more recent version of the system, the statistics of the features is used for the classification task instead of the features themselves. The extracted features are collected over a certain period of time and short-time statistics are calculated. Furthermore, the temporal dependencies between the features are taken into account. This results in both higher recognition performance and lower processing time.

3. SYSTEM REQUIREMENTS

3.1 Robustness Requirements

For a human listener, just a few seconds, even in noisy environments, may be sufficient to identify a song. In order to design a prototype system which approximates this behavior, special attention has to be paid to the alterations an audio signal can be subjected to and to measure the impact of these degradations on the recognition performance. It is therefore of great importance for an audio identification system to handle "real world" audio signals and distortions. Some of these types of

distortions are discussed subsequently, forming the basis of the development process of a robust identification system.

A basic type of signal "degradation" which exists in real world are time shifted signals. If a feature turns out to be very sensitive towards this kind of signal modification, it is likely that this feature will also yield a poor recognition performance when faced with "real world" signals.

Another essential aspect is the sensitivity of the identification system against level changes. This is particularly important when the level of the input signal is unknown, or even worse, may slowly vary over time. Such situations arise when, for example, a song is recorded via a microphone. When considering this kind of distortion, the selected features should be invariant to scaling. This is, for instance, the case for energy envelopes and loudness. However, appropriate post processing of such features can avoid this dependency. A simple example could be the calculation of the difference of two consecutive feature vectors (these are the so-called *delta features*). Other transforms may be applicable as well to overcome this deficiency.

The following list enumerates a selection of signal distortions which were used during the development process of the identifications system to form a test suite of typical "reference distortions", each representing a different aspect of robustness.

- Time shift: Tests the system's robustness against arbitrary time shifts of the input signal. This can be performed very easily by accessing the original audio signal randomly. Care should be taken that the entry points do not correspond to a block boundary used during training.
- Cropping: It is desirable that an audio identification system may be able to identify a small excerpt from a musical item with sufficient accuracy. In this way, identification of an entire song would be possible when only parts (such as the introduction or chorus) are used for recognition. As a consequence, the duration of a song to be entered in the base class database cannot be used as a feature.
- Volume change: By scaling the input signal by a constant or slightly time varying factor, the signal amplitude (volume) may be varied within a reasonable range. In order to counter level dependency, all features/post processing chosen for the identification system were designed to be level independent. Thus, no separate test results will be listed for this type of robustness test.
- Perceptual audio coding: An ever-increasing amount of audio is available in various compressed audio formats (e.g. MP3). It is therefore important for an identification system to maintain high recognition performance when faced with this kind of signals. The bitrate should be selected within a reasonable range, so that the degradation of subjective audio quality is not excessive. A bitrate of 96kbps for an MPEG-1/2 Layer-3 coded stereo signal is considered to be appropriate for general testing.
- Equalization: Linear distortion may e.g. result from applying equalization which is widely used to adapt the frequency characteristics to the users personal taste. For robustness testing of the audio identification system, octave band equalization has been used with adjacent band attenuations set to -6dB and +6dB in an alternating fashion.
- Bandlimiting: Bandlimited signals occur when the signal was represented at a low sample rate or, simply, if a low pass filter has been applied. This can be regarded as a special case of equalization.
- Dynamic range compression: Dynamic range compression is frequently used in broadcast stations. In order to identify audio signals from these stations, robustness against this *time-variant* type of processing must be considered.
- Noise addition: *White noise* or *pink noise* with a reasonable SNR (like e.g. 20-25 dB) was added to the item with a constant level in order to simulate effects such as analog background noise.
- Loudspeaker-microphone transmission (Ls-Mic): This kind of distortion appears when a musical item is played back over a loudspeaker and the emitted sound is recorded via a microphone. The resulting analog signal is then digitized by means of an A/D converter and presented to the input of the system. Such a setup provides a realistic combination of both severe linear and non-linear distortions and has been found to be one of the most challenging types of distortions with respect to automatic audio recognition. A system exhibiting robustness with respect to such a scenario is perfectly suitable for a wide range of applications. The test setup used in the presented work consists of a pair of small multimedia PC speakers and a standard PC microphone, which is directed towards the speakers at a distance of around 10cm.

While this list is by far not exhaustive, it should be sufficient for a general assessment of a system's robustness qualities. In particular, the robustness of each feature with respect to these distortion types can be quantified effectively by such a test suite and then taken into account for the final feature selection process.

3.2 Computational Requirements

When investigating the necessary computational resources of all the software components involved in the identification process, it becomes apparent that that there exists a clear asymmetry between the feature extractor and the classifier in terms of processing power and memory space (both RAM and disk space). More precisely, the extraction process ("fingerprint generation") can be performed several times faster than real-time, since it only consists of a signal analysis followed by a feature calculation. This processing step is independent from the used classification scheme and from the database size, and thus only requires a small amount of CPU processing power and RAM storage.

In contrast, the required resources for the classification task are directly related to the underlying matching algorithm, the size of the database (i.e. the number of trained reference items) and the size and type of the fingerprint information.

While there is a trade-off between the degree of tolerable distortions, the fingerprint size and the computational complexity of the matching algorithm, it was the goal of the work described in this paper to find efficient configurations which would allow for both reliable recognition of real-world audio signals and real-time operation on today's standard PC computing platforms.

4. RECOGNITION PERFORMANCE

This section discusses the recognition performance achieved by the prototype system depending on the choice of features. More specifically, the performance of the system is investigated when faced with distorted audio signals like the ones listed in the previous section. Figures are provided for different configurations of the system, including three features and different sizes of the test database.

4.1 Features

A decisive factor in the performance of the identification system is the selection of features. An extensive review of potentially interesting features led to the selection of the following candidate features which have been used for further experimentation.

- An important part in the perception of sound is represented by the so-called Loudness. Loudness belongs to the category of intensity sensations [15]. It seems intuitive that this basic aspect of an audio signal could serve as a robust feature for audio identification. Simple computational models of loudness are known, including both the calculation of the signal's total loudness and partial loudness in different frequency bands. This provides plenty of flexibility for defining a loudness-based feature set. For the following investigations a multi-band loudness feature was used.
- Besides the loudness sensation, another important characteristics of the audio signal relates to the distinction between more tone-like and more noise-like signal quality. The so-called SFM (Spectral Flatness Measure) [16] is a function which is related to the tonality aspect of the audio signal and can therefore be used as a discriminating criterion between different audio signals. Similar to loudness, the SFM can be used to describe the signal in different frequency bands. Such a multi-band version of the SFM features was used for the following evaluations.
- Similar to SFM, a so-called SCF ("Spectral Crest Factor") feature was investigated which is related to the tonality aspect of the audio signal as well. Instead of calculating the mean value for the numerator the maximum is computed, i.e. the ratio between the maximum spectral power within a frequency band and its mean power is determined. In the same way as for SFM, a multi-band version is used.

The next sections present classification results based on different setups. Each setup consists of a data base holding an increasing number of music items.

For each setup, a few tables are provided which reflect the recognition performance of the identification system. The performance is characterized by a pair of numbers, where the first stands for the percentage of items correctly identified (top 1), while the second expresses the percentage for which the item was placed within the first ten best matches (top 10).

4.2 1,000 Items Setup

An experimental setup of 1,000 musical items was chosen first, each item stored in the compressed MPEG-1/2 Layer 3 format (at a data rate of 192 kbit/s for a stereo signal). The items were chosen from the combined genre rock/pop, to make a distinction between the items more demanding than if material with a wider diversity of characteristics would have been used. To achieve a fast classification of the test items the processed length was set to 20 seconds while training was limited to 30 seconds, i.e. the data had to be recognized based on an excerpt of the sequence only. The feature extractor uses a block size of 1,024 samples. Both the Loudness and the SFM feature were using 4 frequency bands. After feature extraction, temporal grouping and subsequent transformation techniques were applied prior further processing. The generation of the base classes was conducted as described above (VQ clustering algorithm). The setup described here allowed a classification time of 1 second per item (measured on a Pentium III 500 MHz class PC). A selection of the recognition performance for this setup of the system is reported in Table 1.

Table 1. Recognition performance of Loudness and SFM features (1,000 item setup, top 1/ top 10)

Feature	Loudness	SFM
No distortion	100.0% / 100.0%	100.0% / 100.0%
Cropping 15s	51.0% / 75.5%	92.3% / 99.6%
Equalization	99.6% / 100.0%	14.1% / 29.8%
Dynamic Range Compression	89.5% / 94.9%	99.0% / 99.3%
MPEG-1/2 Layer 3 @ 96 kbit/s	19.0% / 33.3%	90.0% / 98.6
Loudspeaker / Microphone Chain	38.3% / 61.7%	27.2% / 59.7%

As can be seen from these figures, the Loudness feature provides a rather low recognition performance for the case of cropping effects (further restriction to 15s length) or MPEG-1/2 Layer-3 robustness. In contrast to this, SFM shows very good performance concerning these robustness tests. Both features do not perform very well in this configuration for the loudspeaker/microphone chain experiment.

4.3 15,000 Items Setup

This setup represents one significant step on the way to a "real world" scenario. A set of 15,000 items was chosen as a database for the classification system, representing a clearly more demanding task. Again the chosen test items belong mostly to the rock/pop genre. To cope with the two main points of interest (namely speed and discrimination) while handling this amount of data, some improvements were made compared to the previous setup. To realize an even faster classification speed with a larger number of items, the statistical analysis of the features was exploited and used for classification instead of the raw features themselves. Furthermore, the number of frequency bands was increased from 4 to 16 bands in order to achieve a more precise description of the audio signal.

A further difference compared to the previous setup is the fact that the features were implemented in accordance with the time/frequency resolution as specified for the extraction of Low Level Descriptors (LLDs) by the MPEG-7 audio standard [2] (i.e. same window/DFT and shift length).

Tables 2 and 3 show the recognition performance achieved for this experimental setup, now investigating the behavior of the promising features which are related to the signal's spectral

flatness properties (and thus "tone-likeness"). Table 2 reports the classification results of a standard Vector Quantization approach, whereas Table 3 shows the results for a more advanced matching algorithm including aspects of temporal relationship between subsequent feature vectors. As can be seen from the figures, both features (SFM and SCF) perform extremely well even under severe distortion conditions, such as the loudspeaker/microphone chain. It can be observed that the SFM feature performs very good while using a standard VQ classifier. This is further increased to recognition rates above 97% with the more sophisticated matching algorithm. In both cases, SCF shows an even better recognition performance. Being at some kind of "saturation level" further tests with an increased amount of items and additional robustness requirements are mandatory for a better discrimination of the two features. Classification time is 7.5 seconds for standard and 2.5 seconds for advanced matching (per item).

Table 2. Recognition performance of *SFM* and *SCF* features using *standard* matching (15,000 item setup)

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	96.1% / 97.2%	99.4% / 99.6%
MPEG-1/2 Layer 3 @ 96 kbit/s & Cropping	92.2% / 94.3%	98.8% / 99.3%

Table 3. Recognition performance of *SFM* and *SCF* features using *advanced* matching (15,000 item setup)

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	99.6% / 99.8%	100.0% / 100.0%
MPEG-1/2 Layer 3 @ 96 kbit/s & Cropping	97.9% / 99.9%	99.7% / 100.0%
Loudspeaker / Microphone Chain & Cropping	98.0% / 99.0%	98.8% / 99.5%

5. APPLICATIONS

The identification of audio content based on matching to a database of known works has many attractive applications, some of which are presented in the following:

- **Audio Fingerprinting:** Matching of audio signals as described in this paper is closely related to the much-discussed topic of "Audio Fingerprinting". A compact representation of the signal features for matching (e.g. the VQ codebooks) resembles the condensed "essence" of the audio item and is thus usable as a fingerprint of the corresponding item.
- **Identification of music and linking to metadata:** Automated identification of audio signals is a universal mechanism for finding associated descriptive data (metadata) for a given piece of audio content. This is especially useful when the format the content has been delivered in is irrelevant for the identification process and when furthermore this format does not support the transport of associated metadata or reference thereto. Under these premises recognition of the song will also serve to provide links to the corresponding metadata. Since the metadata is not necessarily embedded in the audio content, access to a remote database could carry updated information on the artist, concerts, new releases and so on.
- **Broadcast monitoring:** A system for automatic audio recognition can identify and protocol transmitted audio program material on broadcasting stations. With a system like the one introduced in this paper this can be achieved without the need for special processing of the transmitted audio material, as would otherwise be required when using branding methods like watermarking. Applications that require monitoring of radio programs would include verification of scheduled transmission of advertisement spots, securing the composer's royalties for broadcast material or statistical analysis of program material (charts analysis).
- **Music Sales:** Automatic audio identification can also be used to retrieve ordering and pricing information of the identified material and additionally offer similar material. The recording of sound/music and storage of the signature on small handheld devices (such as Personal Digital Assistants) will enable the customer to find the recorded music item in the music store or by connecting to the Internet.

6. MPEG-7 AND ROBUST IDENTIFICATION OF AUDIO

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 goal can be achieved by using pre-computed descriptive data ("metadata") which is associated with the content. One example of a number of upcoming metadata standards for audiovisual data is the MPEG-7 [2] 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) [17].

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.

In the area of audio signal description, MPEG-7 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 [2]. 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.

7. CONCLUSIONS AND OUTLOOK

This paper discussed methods for achieving automatic content-based identification of audio material by means of robust matching to a set of known reference items. Particular attention was paid to aspects of robustness with respect to common types of signal alterations, including both linear and non-linear distortions, audio compression and cropping to a reasonably-sized excerpt. The ability to handle these types of distortions is vital to the usability of systems for content-based processing in many real-world application scenarios.

Relying on a general feature extraction/pattern recognition paradigm, a prototype system for automatic identification of audio material was described in its architecture and background. Clearly, the selection of appropriate robust features can be considered crucial for achieving a good recognition performance under a wide range of possible distortions.

Recognizing the importance of the application, the upcoming MPEG-7 audio standard defines a descriptor designed to provide the functionality of robust matching of pairs of audio signals which relates to the "un-flatness" of the signal's power spectrum and thus the tone-like quality of the signal in a number of frequency bands.

Using this (and related) features, the recognition performance of the identification system was assessed in a number of experiments. The system configuration used showed excellent matching performance for a test set comprising 15,000 songs. A correct identification rate of better than 98% was achieved even for severe distortion types, including an acoustic transmission over a loudspeaker/microphone chain. The system runs about 80 times real-time performance on a Pentium III 500MHz class PC.

Clearly, there is still a long way to go until such an automatic system will be able to match the recognition performance of a human listener. Nonetheless, the current level of performance already opens the door for a number of very interesting applications, including finding associated metadata for a given piece of audio content, broadcast monitoring and music sales.

8. REFERENCES

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

[2] ISO-IEC/JTC1 SC29 WG11 Moving Pictures Expert Group. Information technology – multimedia content description interface – part 4: Audio. Comittee Draft 15938-4, ISO/IEC, 2000.

[3] Gracenote homepage: http://www.gracenote.com

[4] E. Scheirer, and M. Slaney. *Construction and evaluation of a robust multifeature speech music discriminator*. In ICASSP, 1997.

[5] R. M. Aarts, and R. T. Dekkers. *A real-time speech-music discriminator*. J. Audio Eng. Soc., 47(9), 1999.

[6] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal. *Speech music discrimination for multimedia applications*. In ICASSP, vol. IV, pages 2445-2448, 2000.

[7] Cantametrix homepage: http://www.cantametrix.com

[8] Musclefish homepage: http://www.musclefish.com

[9] E. Wold, T. Blum, D. Keislar, and J. Wheaton. *Content-based classification, search, and retrieval of audio*. In IEEE Multimedia, vol. 3, pages 27-36, 1996.

[10] D. Pye. *Content-based methods for the management of digital music*. In ICASSP, vol. IV, pages 2437-2440, 2000.

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

[12] C. Papaodysseus, G. Roussopoulos, D. Fragoulis, T. Panagopoulos, and C. Alexiou. *A new approach to the*

automatic recognition of musical recordings. J. Audio Eng. Soc., 49(1/2), 2001.

[13] A. K. Jain, R. P. W. Duin, and J. Mao. *Statistical Pattern Recognition: A Review*. IEEE Transaction in Pattern Analysis and Machine Intelligence, 2(1), 2000.

[14] D. Kil, and F. Shin. *Pattern Recognition and Prediction with Applications to Signal Characterization*. American Institute of Physics, 1996.

[15] E. Zwicker, and H. Fastl. *Psychoacoustics*. Springer, Berlin, 2nd edition, 1999.

[16] N. Jayant, and P. Noll. *Digital Coding of Waveforms*. Prentice-Hall, Englewood Cliffs, NJ, 1984.

[17] ISO/IEC JTC1/SC29/WG11 (MPEG): "Introduction to MPEG-7", available from http://www.cselt.it/mpeg.