Content-based Identification of Audio Titles on the Internet

Helmut Neuschmied
Institute of
Information Systems
Joanneum Research
Steyrergasse 17
A-8010 Graz, Austria
+43 316 876 1159
helmut.neuschmied@joanneum.ac.at

Harald Mayer
Institute of
Information Systems
Joanneum Research
Steyergasse 17
A-8010 Graz, Austria
+43 316 876 1136
harald.mayer@joanneum.ac.at

Eloi Batlle
Institut Universitari
de L'Audiovisual
Universitat Pompeu Fabra
Passeig de Circumvalació, 8
08003 Barcelona, Spain
+34 93 542 2200
Eloi.Batlle@iua.upf.es

ABSTRACT
The increasing usage of the Internet for the distribution of audio content and the increasing number of audio broadcasting stations require new supporting tools and methods for the observation of occurrence frequencies and of possible copyright-infringements. This paper describes a general approach for the identification of audio titles and its application on Internet observation. The concept of an AudioDNA is developed, allowing a highly compressed representation of a sequence of acoustic events. Audio titles are identified by using a sequence matching method which determines similarities between observed and reference AudioDNA stored in a database. This method is implemented in an highly scaleable architecture allowing to identify several audio titles in parallel. The first results with this system are very promising, only highly distorted audio titles are not identified correctly.

Keywords
Audio recognition, content based identification, copyright ownership protection, sequence matching

1 INTRODUCTION

Automatic identification of audio titles becomes a more and more important application due to the highly increasing amount of audio distribution channels, including radio stations, Internet radio, file download and exchange facilities. For content owners it is important to know where their music is played and consequently if they receive proper royalties.

In this paper we describe a highly-scalable system for the automatic identification of audio titles, which is based on the concept of AudioDNA, a highly compressed representation of audio signals.

Most approaches for automatic music identification implement watermarking methods, i.e. embedding inaudible digital data containing a unique identifier within the audio signal [8]. This approach is also used by the Secure Digital Music Initiative (SDMI), an organisation founded by the music industry [10]. But in practise it is difficult to develop a watermark which is really inaudible to the listeners, especially if it has to be robust against psycho acoustical compression techniques like MP3 encoding. And watermarking is no choice for already released material. In our system design we do not rely on any additional information within the observed signal. Of course such recognition methods may be included as an additional component to our system.

There are several approaches on content-based identification and search of audio material. E. Wold et al. [2] from Muscle Fish describe a system for finding similar sounds to a given example. This system extracts time-varying properties from sampled sound files and for each property the mean, variance and autocorrelation over the entire file is recorded. At the time of their publication the system was used for comparison of noises, like scratches, bells and laughing but was not used for whole song identification.

Nowadays there are several companies claiming to provide music identification technology and services, like tuneprint [11], eTantrum [12], cantametrix [13] and audible magic [14]. The last one is based on the technology of Muscle Fish and already offers a Windows application named Clango [15] for end users to identify music currently playing on their PC. However, due to the commercial orientation there are not too much details available about the used methods and technologies.

In the following sections we describe the manifold requirements to a recognition system, the fundamentals of our approach and its application.

2 RECOGNITION REQUIREMENTS

For a human listener it seems quite easy to recognize a known audio title. However, trying to have a computer identifying audio title is a very challenging task. Though the comparison of two fully identical audio signals is not too complicated, there are several practical aspects which have to be taken in account when building a identification system.

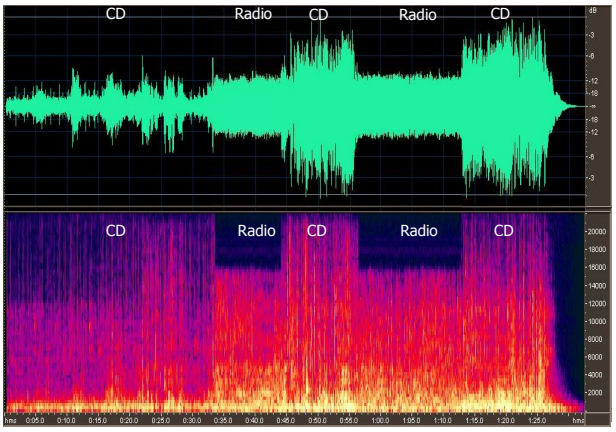

Figure 1: Audio signal (upper image) and frequency distribution (lover image) of an audio title captured alternatively from a CD and from a radio station.

One of these aspects is the robustness of the recognition method with regard to variations to the audio signal. Depending on the audio source (e.g. CD, Radio, Internet audio sources, etc.) the actual audio signal may be very different from a master source.

Figure 1 shows an example of the frequency distribution of an audio signal, which has been captured both from a CD and from a radio station. The main reasons for such audio signal differences are:

- **Audio storage:** Signal-to-noise ratio, digitization parameters and frequency range of the storage media, influence the degeneration of the audio signal. Especially for audio sources from the Internet differences are caused by the various compression formats and compression rates which are used for saving bandwidth.
- **Transmission effects:** The broadcasting techniques (e.g. frequency modulation (FM), amplitude modulation (AM)) specify amongst other things the bandwidth and the signal-to-noise ratio of the observed audio signal. The Internet compression technologies used within streaming media protocols could also assigned to this category.
- **Manipulative effects:** Radio stations use complex sound processing to apply their own sound characteristics. For example they often enlarge the stereo base, which enhances the impression of stereo listening, or they play the songs faster.

A recognition method has to be robust against the above mentioned distortions and has to identify titles even with small fractions of a title (e.g. some seconds of music).

In addition to these transmission and manipulative effects, there are several additional issues which make identification a complicated task. Broadcasting stations often play only parts of a title and mix it with the voice of a moderator.

And there are some cases where it depends on the application if a title should be identified as equal to one master recording, e.g. when a title stems from the same artist but from a different recording (live versions), when a title is performed by a different artist (cover versions) or when parts of title are reused within DJ mixes.

A useful recognition system has also to fulfil certain requirements on processing speed. It has to be capable to identify one title against a database with several ten thousand of master recordings considerably faster than real time.

3 OUR APPROACH

The aim of our approach is to build up a system which fulfils the above mentioned requirements consisting only of standard PC equipment, i.e. no special hardware developments.

The idea is to extract characteristic information not only from the whole piece of music but also from smaller parts of the acoustic signal. Therefore we see the audio signal as a sequence of acoustic events. Such a sequence identifies a music title. In analogy to the biological terminology we name the acoustic events as audio genes. A piece of music is composed of a sequence of audio genes which is called the AudioDNA.

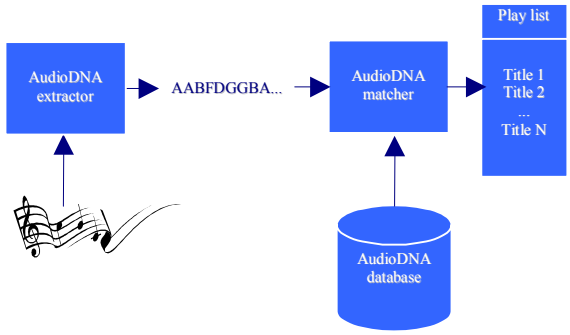

Figure 2: Principle recognition method.

In speech recognition audio genes can be described in terms of phonemes. In music modelling notes can be used. But

notes are often played simultaneously (accords, polyphonic music) and music samples contain additional voices or other sounds. Hence our main objective is to engineer audio genes which can be used for all kind of audio signals (see also AudioDNA Extraction).

Figure 2 shows the principle scheme to recognise audio titles by using an AudioDNA. The AudioDNA is extracted from the observed audio signal. The AudioDNA extractor produces continuously a sequence of audio genes coded denoted here as letters. This sequence is matched against a database of AudioDNA's of already known titles. If in the database a similar AudioDNA subsequence with a minimum length is found then an audio title has been detected. The name of the identified title together with the time interval entered in the play list.

This method relies on a database which contains information about titles which should be identified. Therefore the AudioDNA of reference (master) audio titles has to be acquired and saved in the database together with the according meta information (title name, artist, etc.).

4 AudioDNA EXTRACTION

The AudioDNA extraction procedure is mainly based on statistical pattern matching algorithms [4] and the procedure is based on speech recognition system.

First of all, the audio signal is split into frames with fixed length (this length can vary from application to application but, once it is fixed, it remains the same during the training and recognition stages). After this splitting, a Hamming window is applied to the audio frames (to avoid border effects) and some parameters are extracted (mainly timbre related ones). To increase the performance of the system, these parameters are decorrelated using a projection matrix [6]. The goal of the decorrelation is to allow the use of a diagonal covariance matrix but also to reduce the number of parameters in the system (we keep only the vectors with higher eigenvalues).

After the parameter extraction process, these feature vectors are used to compute a probability for each AudioGene using Hidden Markov Models (HMM) [5].

At the final step of the AudioDNA extraction, the Viterbi algorithm [7] gives the most probable sequence of AudioGenes.

5 AudioDNA MATCHING

Usually the observed audio signal is manipulated by broadcasting effects or by lossy compression techniques(see Recognition requirements). For this reason the AudioDNA of the observed signal differs from the AudioDNA which is saved in the database. To find an audio title in the database not only exact matches but also similar AudioDNA sequences have to be found.

For the calculation of the similarity of sequences we have to deal with inexact or approximate matching methods. This is a characteristic problem in text searching applications and in the field of computational biology.

In the applied matching algorithm there are two main processing steps. The same principle method is also used within FASTA [3], one of the most effective practical database search methods for biological gene sequences:

1. First the search space in the database is reduced by exact matching of short subsequences.
2. On the positions of the exact matches the similarity of a longer sequence is examined by an approximate matching algorithm.

In difference to text strings or biological gene sequences the genes of our AudioDNA have additionally a time information. Each audio gene represented by a letter correspond to a specific time period of the acoustic signal (see Figure 3)

Exact Matching of Subsequences

From the audio gene stream which is produced by the AudioDNA extractor (see Figure 2) short subsequences are extracted continuously. It will be checked if a specific number of these subsequences can be found in one of the database AudioDNA's in the same order. For this database search the time values of audio genes are not used.

The exact matches can be efficiently found by creating an index of all subsequences which occur in the database. For that the database AudioDNA's are split into overlapping substrings of a specific length. From this overlapping substrings an index is created.

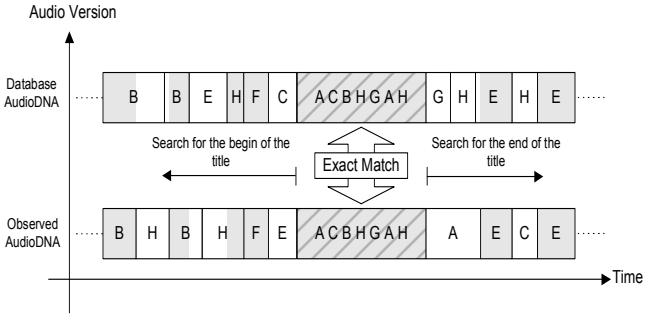

Figure 3: The exact match result is used to detect equal genes which have a time overlap (indicated by the grey areas).

The exact match results are validated using the time period values of the audio genes. A correlation value is calculated from all time values of the observed and of the database genes, which occur in the subsequences. If the correlation value is lower then a predefined threshold value the result of the exact matching will be rejected.

The Approximate Matching Algorithm

To detect similarities of longer sequences at the position of the exact matches an approximate matching algorithm has

to be used. To start from the exact match position the audio genes are consecutively compared in each direction (see Figure 3). If in a defined time period the sequence similarity falls short of a predefined threshold the start or the end of the matching sequence and therefore of the audio title is detected.

The calculation of the similarity of sequences is a standard application of dynamic programming [1]. The time information of the audio genes makes it possible to use a simpler and faster approximate matching algorithm which is described below.

By the exact match result we get a time basis for the alignment of the observed AudioDNA and the AudioDNA, which is saved in the database. By this alignment it is possible to determine the time periods Δt_{equal} where equal audio genes occur (see Figure 3). The similarity S of the AudioDNA sequences in a specific time period Δt_{obs} is given by the equation:

$$S(\Delta t_{obs}) = \frac{\sum_{i=1}^{n} \Delta t_{equal}(i)}{\Delta t_{obs}}$$

where $\Delta t_{equal}(1)$ is the first and $\Delta t_{equal}(n)$ is the last Δt_{equal} in the time period Δt_{obs}. The similarity S can be calculated in $O(N)$ time where N is the length of the compared sequence.

6 FIRST RESULTS

Up to now nearly 2000 songs are saved in our AudioDNA reference database. To test the recognition of audio titles we continuously captured 12 hours broadcasted by radio stations. 75 songs have been played during these 12 hours.

The evaluation of this test material yielded a very promising result: 72 songs were detected. Two songs were not recognised because they were not in the reference database. Another song, which was not detected, was very noisy (by speaker and non-linear limiter).

The software was also tested with MP3 compressed files. Even files with lower compression rates were correctly identified.

The extraction of the AudioDNA can be done faster than real-time, which allows the simultaneous observation of several audio streams at the same PC. The matching of the AudioDNA data of 12 hour test data against the reference database required only two minutes.

7 APPLICATION

The Internet is more and more used for the distribution of audio content. Therefore technologies for monitoring the usage of audio material on the Internet become increasingly important. For this reason the proposed recognition technique is used to built up a system which systematically searches the Internet for audio titles and identifies them.

Audio content on the Internet appears in several variants:

- Audio files in different formats which can be downloaded from Web or FTP servers.
- Audio content transmitted in a streaming audio format (Internet radio).
- Sharing of audio files among Internet users with the help of tools like Napster and Gnutella.

The audio content of streaming audio servers can be checked by continuously receiving and processing the audio stream, much like conventional radio. In this case an operator tunes into the selected radio station.

Finding audio files on the Internet can not be done manually if a huge number of such files should be processed. In this case a Web crawler is used to systematically visit Web and FTP server.

The technologies for Web crawlers are well known [9], the challenging part in our development is the scalability of the system and the ability to find titles which are hidden to automatic systems by obscuring them. For example audio files may be stored in software archives (ZIP files) and these files have extensions like image files. Or FTP servers are often secured by a username and password which can be retrieved by opening a text file named "HOW TO DOWNLOAD FROM THIS SERVER.txt".

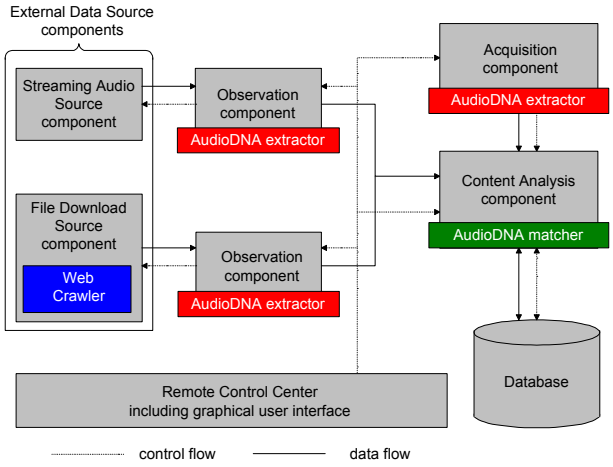

Figure 4: System architecture.

System Design

Figure 4 shows the software architecture of the system. The system is built up from components which are controlled by the so called Remote Control Center. The components are implemented as standalone applications and use CORBA (Common Object Request Broker Architecture) as communication platform. Hence the components can be run on a network of distributed computers. Several instances of components can be launched depending on the number of

observed audio sources. This architecture is the basis for a highly scaleable framework.

The Acquisition component is responsible to produce the reference data (AudioDNA and meta data) which are saved in the database. Usually a CD drive or an interface to an audio database of content owners is used as reference audio source.

The External Data Source components receive and convert the audio data from the different observed audio sources to a standardized format. The preprocessed audio data are transmitted to an Observation component where the AudioDNA is calculated (see AudioDNA Extraction).

The Content Analysis component detects similarities between the AudioDNA from the Observation components and the reference AudioDNA's (see AudioDNA Matching). The search results are saved as playlists in the database.

8 CONCLUSION

We presented a new approach for the automatic content based identification of audio titles. The audio signal is transformed into a symbolic representation which we call AudioDNA in the context of this project. The AudioDNA extraction is robust against variations and distortions of the audio signal and enables the identification of a title even if only a part of the song is observed.

The research involved around the AudioDNA opened several future lines of development. Among these we can find the inclusion of other musical-based parameters (like rhythm and melodic trajectories) into the pattern matching algorithm as well as improvements into the HMM structure in order to better fit the musical needs.

The proposed technique is used to built up a system for monitoring copyright infringements on the Internet. But it is also planned to observe radio broadcast stations and the automatically generated play-lists can be used to determine market statistics.

ACKNOWLEDGEMENTS

The main work described in this paper was done within a European project, which is partially funded by the European Commission (IST-1999-12585) and is also supported by the Austrian Federal Ministry of Science and Research. More information on the project can be found at http://raa.joanneum.ac.at.

The authors would like to thank all other partners of the Consortium for their cooperation, especially Walter Plaschzug and Peter Uray from HS-art digital service for their contribution to the overall system design.

REFERENCES

1. Guisfield, D. Algorithms on Strings, Trees, and Sequences. *Computer Science and Computional Biology (Cambrigde, 1999)*, Cambridge University Press, ISBN 0-521-58519-8.
2. Wold, E., Blum, T., Keislar, D., and Wheaton, J. Content-Based Classification, Search, and Retrieval of Audio. IEEE Multimedia, Vol. 3., No.3, 1996, 27-36.
3. Perason, W. R. and Lipman, D. J. Improved tools for biological sequence comparision. Proc. Natl. Academy Science, Vol. 85, 1988, 2444-2448.
4. Duda, R. O and Hart, P. E. Pattern Classification and Scene Analysis. John Wiley & Sons, New York, 1973.
5. Rabiner, L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE. Vol. 77, n. 2, 1989, 257-286.
6. Batlle, E., Nadeu, C. And Fonollosa, J.A.R. Feature Decorrelation Methods in Speech Recognition. A Comparative Study, International Conference on Spoken Language Processing, Vol. 3, Sydney, 1998, 951-954.
7. Viterbi, A.J. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. IEEE Transactions on Information Theory, Vol. 13, n. 2, 1967, 260-269.
8. Laurence B., Ahmed H. T., and Khaled N. H. Digital Watermarks for Audio Signals, IEEE Int. Conf. on Multimedia Computing and Systems June 17-23. Hiroshima, Japan, 1996, 473-480.
9. Miller R., Bharat K., SPHINX: A Framework for Creating Personal, Site-Specific Web Crawlers, In Proceedings of WWW7, Brisbane Australia, April 1998.
10. SDMI Identifies Audio Watermark Technology, SDMI (Aug. 9, 1999) <https://www.sdmi.org/dscgi/ds.py/Get/File611/sdmiaug9.htm>.
11. http://www.tuneprint.com
12. http://www.etrantrum.com
13. http://www.cantametrix.com
14. http://www.audiblemagic.com
15. http://www.clango.com