AMADEUS: A SCALABLE HMM-BASED AUDIO INFORMATION RETRIEVAL SYSTEM

Eloi Batlle, Jaume Masip, Enric Guaus

Audiovisual Institute
Universitat Pompeu Fabra
{eloi, jmasip, eguaus}@iua.upf.es

ABSTRACT

The new transmission and storage technologies now available have put together a vast amount of digital audio. All this audio is ready and easy to transfer but it might be useless with a clear knowledge of its content as metadata attached to it. This knowledge can be manually added but this is not feasible for millions of on-line files. In this paper we present a method to automatically derive acoustic information about audio files and a technology to classify and retrieve audio examples.

1. INTRODUCTION

Music Information Retrieval (MIR) technologies allow the identification and management of audio content without using any external metadata or embedded watermarking. In a general Music Information Retrieval scheme, a compact digest derived from the acoustic properties of the audio is processed against a database for a specific purpose.

MIR is a very wide field in the Information Technologies area. Many MIR algorithms have been proposed [1]. All of them can be classified into many subgroups, but at two different conceptual levels. In the first conceptual level, we can find MIR systems which use Decision Trees, Neural Networks, Artificial Intelligence, Hidden Markov Models and so on, while at the second conceptual level we can find MIR systems which use spectral features, time-domain descriptors, multiresolution techniques and so on.

AMADEUS is a framework specially designed to deal with Hidden Markov Model (HMM) related problems. Applications designed by using AMADEUS should ideally be compact, for complexity issues, robust against distortions and should retain as much acoustic relevant information as possible to manage possibly near-infinite data.

As an example of an application using Amadeus, we will present a fingerprinting system which goes beyond the template matching paradigm to a statistical pattern matching paradigm. The goals are to incorporate robustness by statistically modeling the audio evolution while reducing the fingerprint size by considering local and global redundancies in a corpus of audio content. The reduction of the fingerprint data is important to eficiently use the bandwidth in internet applications.

2. SYSTEM'S OVERVIEW

2.1. The identification process

To identify an unknown piece of audio, we use the property of a Hidden Markov Model from what an HMM can be seen as a double stochastic process. Therefore, HMM could be used to generate observations and we can calculate the probability that some observations are generated by a given HMM.

Figure 1 represents a sequence of HMM that models a song. Each HMM is a part of the temporal structure belonging to music. Since music can be seen as a sequence of events sorted in time, this music can be modeled with a sequence of HMMs. The evolution in time of the song is represented with the jumps from one state to the next one.

Let's suppose we have an unknown fragment of audio, where O are all the vectors of parameters (i.e. Mel-cepstrum, rhythm features, harmonic structure description, etc.). If a known song in our database is modeled using an HMM λ, the probability that the generation of this song was the same than the generation of the unknown fragment is [2]:

$$P(O \mid \lambda) = \sum_{q_1,q_2,\ldots,q_T} \pi_{q_1} b_{q_1}(o_1)\, a_{q_1 q_2} b_{q_2}(o_2) \cdots a_{q_{T-1} q_T} b_{q_T}(o_T) \quad (1)$$

With this equation in mind, the identification process can be seen in the following way. We have a database of HMMs sequences that model our repository. When we are given an unknown audio fragment from which we have derived some observations (melody parameters, mel-cepstrum, etc.) Eq. 1 answers the question whether the HMMs of a song λ in the repository would be able to generate that unknown music. This approach has several advantages over a more classical matching approach. The first advantage is that it is more robust to noise because two songs can be modeled in a very different way because one is noisy and the other one is clean, and however the identification process will give the

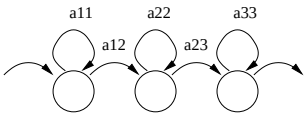

Fig. 1. Song representation with an HMM sequence

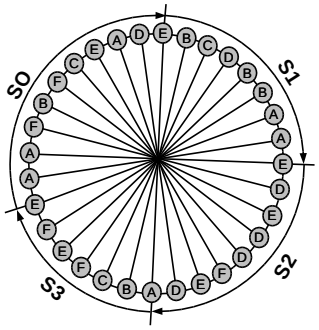

Fig. 2. Song Model

right result. Another advantage is that the same process can be used to retrieve audio similarity as explained in [3].

The identification algorithm matches an input streaming audio against all the fingerprints to determine whenever a song section has been detected. The Viterbi algorithm is used again with the purpose of exploiting the observation capabilities of the HMM models contained in the fingerprint sequences. Nevertheless, this time the model is not a complete graph but the HMM ring shown in Fig. 2. In this structure, each HMM only has two links, one to itself and one toward its immediate neighbor. The identification algorithm scales linearly with the number of songs in the database because no backtracking is required for single path models.

2.2. Feature extraction

There are a lot of features we can extract from music that can help our MIR system. Here are some of them, available in AMADEUS.

2.2.1. Mel-Cepstrum

The *Mel-cepstrum* is a measure of the spectrum shape or the sound "color". The term *cepstrum* was introduced by Bogert et Al. in [4], and it has come to be the accepted terminology for the inverse Fourier Transform of the logarithm of the spectrum of a signal:

$$Cepstrum = IFFT\left[Log\left(FFT\left[x(t)\right]\right)\right] \qquad (2)$$

2.2.2. 4Hz Modulation

As defined in [5], the 4Hz Modulation Energy Peak is a characteristic feature of speech signals due to a near 4Hz syllabic rate. It is calculated by decomposing the original waveform into 20 or 40 mel-frequency bands.

2.2.3. Zero Crossing Rate

As defined in [6], the Zero Crossing Rate (ZCR) of the time domain waveform provides a measure of the weighted average of the spectral energy distribution, similar to a spectral center of mass or Spectral Centroid of the input signal (see Sec.2.2.4). It also can be interpreted as the noisyness of the input signal. From a mathematical point of view, it can be calculated with:

$$ZCR = \frac{1}{2}\sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])| \qquad (3)$$

where the *sign* function is 1 for positive arguments and 0 for negative arguments, and x[n] is the time domain signal for frame t.

2.2.4. Spectral Centroid

As defined in [5], the Spectral Centroid is the *balancing Point* of the spectral power distribution. It can be interpreted as a measure of the average frequency, weighted by amplitude, of a spectrum, that is, a measure related with the brightness of the signal. From a mathematical point of view, the Spectral Centroid can be calculated with:

$$SC = \frac{\sum f_i a_i}{\sum a_i} \qquad (4)$$

where f_i is the frequency value of each bin of the FFT and a_i is its amplitude.

2.2.5. Spectral Flatness

The Spectral Flatness can be defined, according to Ozgur Izmirli in [7], as the ratio of the geometric mean to the arithmetic mean of the power spectral density components in each critical band. Some times, the Spectral Flatness Measure is expressed in decibels:

$$SFM_{dB} = 10log_{10}\frac{G_m}{A_m} \qquad (5)$$

2.2.6. Voice2White

The Voice2White parameter is a measure of the energy inside the typical speech band ($300Hz..4KHz$) respect the energy of the whole audible margin (in case of $sr = 4410Hz$) or global band (in case of $sr < 44100Hz$). From a mathematical point of view:

$$v2w = 10log_{10}\frac{\sum_{f_i=300}^{4500} B_{f_i}}{\sum_i B_i} \qquad (6)$$

2.2.7. Rhythmical features

As defined in [8], some rhythmical features can be extracted from the input signal by calculating the periodogram of the derivative of the multi-band filtered input signal. Some applications and features are explained in the paper.

2.3. Training of the system

The biggest problem that arises with a music recognition scheme based on HMMs is how to find the more suitable HMM set that will lead to a good recognition even in bad noisy environments. The training of a speech recognition system has still some issues but it is a well studied problem. In speech, the target for each HMM is a phoneme (or other phonetic related characteristic), but in music there are not such "phonemes". In [3] we present a way to define some properties for the units that can suit the music identification problem as well as music similarity.

To automatically derive some good units to represent the music, we follow an iterative approach based on the EM algorithm [9].

The algorithm is composed by several steps:

Number of different HMM The first decision has to be taken and it is the number of different HMM that will be used. In other identification tasks, this decision is easy and usually an HMM for each phoneme is used. In the music identification problem, since there are no "phonemes", the number of HMM has to be carefully chosen. The higher the number of HMMs used, the higher the accuracy of the models for each song, but this would mean also more complexity for both the training and identification. If the numbers of HMM is very low, the accuracy of the representation will be very poor and the system will need a longer fragment of unknown audio to identify it. In our case the number of HMM is set to 1024.

Initialization Originally the bootstrap models we used were pure random and all the means and variances were chose at random and the transition probabilities were set to 0.5 for both stay and jump. Unfortunately, this methods lead often to a local maximum that is not good enough for identification because the discrimination capacity for each HMM was very poor. A second method of k-means to create the bootstrap parameters shows a very good performance.

Realignment With the current parameters, the system calculates a new sequence of HMM in order to increase the observation probability. This is done with the Viterbi algorithm [10].

Update With the alignment calculated in the previous step, we use the Baum-Welch algorithm to update the means, variances and transition probabilities.

Loop Steps 3 and 5 are repeated until the global probability of generation is not growing from one iteration to the next one.

3. AUDIO FINGERPRINT SIZE SCALABILITY

Since there is a growing number of available songs due to the digitalization of big repositories, identification and similarity systems should have a good behavior in scalability. In this section we show how we can control the system requirements thanks to the stochastic properties of the fingerprint based on HMMs.

Because of the algorithms behind the audio identification system explained above, it is possible to use some techniques borrowed from information theory that can improve dramatically the performance of the system and the resources need.

Figure 1 shows a small part of a song when it is represented with an HMM sequence. The variables a_{ij} are the probabilities to jump from state i to state j which is the same than saying than represent the probability to change from one acoustical event in a song to the next one. The sequence of events is a Markov Chain and the speed of change is controlled by the transition probability matrix A [11].

From an information theory point of view, we can reduce the size of the fingerprint while keeping a good statistical representation of it by increasing the probability a_{ii} (probability to stay at the same state) and reducing at the same time a_{ij} (for $i \neq j$). This modification will lead to an increase of the state duration that by default is an exponential density function. Since the order of the acoustic events cannot be modified, there exist only a_{ii} and a_{ii+1}. Thus, the adjustable parameter p to control the size of the fingerprint will be used in the new transition probabilities as

$$a_{ii}^{'} = a_{ii} + p \qquad a_{ii+1}^{'} = a_{ii+1} - p$$

constricted to $a_{ii}^{'} + a_{ii+1}^{'} = 1$ and $0 \leq a_{ii}^{'}, a_{ii+1}^{'} \leq 1$.

All this shows that the transition matrix is a good control to set the amount of CPU and memory requirements to use the fingerprinting system. With a high p, the probability to stay at the same state will be higher and, therefore, the size of the fingerprint will be smaller.

Of course there is a drawback for this fingerprint reduction. There is a trade-off between robustness and fingerprint size and the value of p should be carefully chosen with the working conditions of the identification system in mind.

Another well known technique to reduce the fingerprint size is to decimate the feature space by a factor of N.

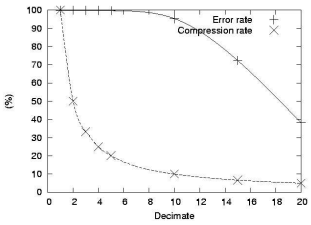

Fig. 3. Transition prob.

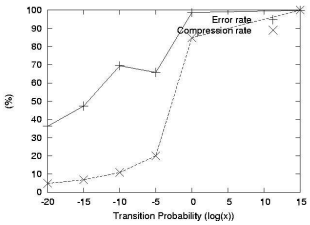

Fig. 4. Decimate

4. EXPERIMENTS AND RESULTS

The input for the experiments was a continuous single channel 16bit 22050KHz audio stream generated with the addition of samples of 10 seconds extracted randomly from a 1570 song database. Each experiment took less than 1 hours to complete on a computer cluster with 32 Pentium-III CPUs at 1GHz interconnected through MPI.

One experiment has been selected as reference to study the different system variables. This reference experiment used 1024 HMM with non compressed fingerprints (transition probability modification=0, decimate=1). The results show the identification error rates for different probabilities and decimates. The compression rates over the fingerprint are shown in the same graphic.

4.1. Transition probability and fingerprint depths

The reference experiment has been confronted against the transition probability system variable. The reference result for this experiment was the same than before. Fig. 3 plots the error rate versus the modification in the transition probability from one state to the next one. For convenience the parameter p is shown proportional to a logarithmic scale, where 0 means no modification of the probability.

4.2. Decimate and fingerprint depths

The reference experiment for the changes in the decimate value was the maximum reached in experiments from section 4.1. The graph show what was already expected and a high reduction of the framerate involve a reduction in the accuracy.

5. CONCLUSIONS

The fingerprinting approach designed with the AMADEUS technology shows very promising results. The proposed architecture allows a very flexible system able to adapt to different MIR environments, including similarity issues of MIR. In the other hand, for compressed data transmission in Internet or for noisy environments, the transition probabilities adjustments allow a better model for each song and therefore a more accurate data managing. From this point of view, AMADEUS seems to be a very powerful tool for MIR applications where the amount of data tends to infinite but the physical support is limited.

6. REFERENCES

[1] Jonathan Foote, "An overview of audio information retrieval," *Multimedia Systems*, vol. 7, no. 1, pp. 2–10, 1999.

[2] L. R. Rabiner, "A Tutorial on HMM and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[3] E. Batlle, E. Guaus, and J. Masip, "Open position: Multilingual orchestra conductor. lifetime opportunity," in *Proceedings of 26th ACM/SIGIR International Symposium on Information Retrieval*, Toronto, Canada, 2003.

[4] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Symposium on Time Series Analysis*, 1963, pp. 209–243.

[5] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," in *Proc. ICASSP*, 1997, pp. 1331–1334.

[6] J. Saunders, "Real-Time Discrimination of Broadcast Speech/Music," in *Proc. ICASSP*, 1996, pp. 993–996.

[7] Ozgur Izmirli, "Using Spectral Flatness Based Feature for Audio Segmentation and Retrieval," Tech. Rep., Center for Arts and Technology, Department of Mathematics and Computer Science, Connectucut College, 1999.

[8] E. Guaus and E. Batlle, "Visualization of metre and other rhythm features," in *Proceedings of IEEE Symposium on Signal Processing and Information Technology*, 2003.

[9] A. P. Dempster and et altri, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[10] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Identification," *IEEE Trans. Info. Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[11] P. Bremaud, *Markov Chains. Gibbs fields, Monte Carlo Simulation and Queues*, Springer, 1999.