AUDIO IDENTIFICATION BASED ON SPECTRAL MODELING OF BARK-BANDS ENERGY AND SYNCHRONIZATION THROUGH ONSET DETECTION

Mathieu Ramona, Geoffroy Peeters

Ircam - CNRS (Sound Analysis/Synthesis Team)
1, pl. Igor Stravinsky - 75004 Paris - France
mathieu.ramona, geoffroy.peeters@ircam.fr

ABSTRACT

In this paper, we present for the first time the fingerprint IRCAM system for audio identification in streams. The baseline system relies on a double-nested Short Time Fourier Transform. The first STFT computes the energies of a filter-bank, that are then modelled over 2 s, using a second STFT. We then present recent improvements of our system: first the inclusion of perceptual scales for amplitude and frequency (Bark bands), then the synchronization of stream and database frames using an onset detection system. The performance of these improvements is tested on a large set of real audio streams. We compare our results with the results of re-implementations of the two state-of-the-art systems of Philips and Shazam.

1. INTRODUCTION

Audio identification aims at detecting occurrences of known audio samples (or items) in an unknown audio signal or stream, generally through the use of a fingerprint code. The latter is designed to make a compact numerical representation of audio samples that is highly discriminative between different items, while remaining robust to typical distortions expected on an audio recording. Its applications are numerous, though the most straightforward is the detection of musical tracks on broadcast radio or TV streams, for copyright supervision.

The field of audio fingerprinting is covered by many industrial actors, among which Philips [1] proposes a very compact representation (32 bits) of sub-band energy differences, combined with an exact match search in a hash table. The Shazam system [2] is based on numerous compact key signatures representing peak pairs in the spectrogram ; the accumulation of many keys for a given item during search determines its detection. AudibleMagic [3] (based on the Muscle Fish technology) relies on a common pattern classification framework to classify the codes among the tracks in the database. The *AudioID* technology developed by Fraunhofer [4], is also built on a classical pattern classification framework, using a standard Nearest Neighbor rule on MPEG-7 descriptors, coded through Vector Quantization.

While some authors propose detection schemes based on the entire signal of a sole analyzed audio track [3], most address the problem of the audio identification in a live stream. This implies both a short delay on the system output, and a highly optimized code computation, in order to cope with real-time constraints. The signal frame used for the creation of each code must thus be kept as short as possible, while maintaining it discriminative enough to avoid false alarms. Philips and Shazam systems both rely on very short term horizons for each fingerprint (a few milliseconds), compensating the lack of robustness of the short-term code by an analysis of the succession of multiple codes over time. We propose here a different scheme for audio identification, based on an original and more robust fingerprint, computed through the spectral modeling of Bark-band energies, and extracted from larger frames (a few seconds) at a much lower rate. The latter characteristic drastically reduces the volume of the search database. However, we will see that reducing the sampling rate of the fingerprint codes necessarily implies considering time shifts between the frames in the stream and in the tracks used for learning. We address this problem by proposing a new method to synchronize the codes efficiently between the analyzed signal and the database items.

We provide here, for the first time, a detailed description of the IRCAM audio fingerprint framework (originally developed in 1998), recent improvements on it, as well as a comparative evaluation with our implementations of two major contributions of the literature (Philips and Shazam), that is unprecedented, as far as we know.

The article is structured as follows : the base-line IRCAM system will be presented in section 2, which also includes a brief presentation of the search strategy in section 2.2. The perceptual scales in the code computation will then be introduced in section 3 and the frame synchronization issue will be addressed in section 4. Section 5 will then assess the importance of synchronization through a study on the code robustness in subsection 5.1. We then provide evaluation of both the onset detection proposition, in subsection 5.3, and the overall system, compared to state-of-the-art techniques, in subsection 5.4. A brief conclusion in section 6 will sum up our contributions on the problem and perspectives for future work.

2. BASE-LINE SYSTEM

Our base-line audio identification system is the one proposed in [5] and [6][1]. The system is composed of a coding and a search part.

2.1. Fingerprint code computation

The global idea of our code is to represent directly the evolution of the signal characteristics over time and not only the characteristics around a specific time (as most coding schemes do). The goal is to allow a fast search, which is possible since our code directly represents the evolution of the signal. We can therefore only use a single search, if the frame is long enough (several seconds). On the

[1] It should be noted that the code used for audio identification has also been used in [7] in the framework of audio structure by similarity (estimating repetitions inside a single track).

opposite, algorithms based on local (short-term) codes need to perform several searches corresponding to the succession of codes. The evolution of the characteristics of the signal is represented using a spectral representation of the energy content of the signal in several different spectral bands. Although this method is known today as ”modulation spectrum” [8] [9] [10], our initial proposal of this formulation dates back to 1998 [5]. We summarize the computation of our code in the following.

We consider the sampled audio signal $x(n)$ of sampling rate f_s. The signal is first normalized according to $\hat{x}(n) = x(n)10^{\frac{96}{20}}$. The signal is then analyzed using a Short-Time Fourier Transform with a Blackman window $w(n)$ of $l = 100$ ms duration and $h = 25$ ms hop size. The STFT at time m and frequency k is expressed as

$$X(k,m) = \sum_{n=0}^{N-1} \hat{x}\left(m+n\right)\, w(n)\, \exp\left(-j2\pi\frac{k}{N}n\right), \qquad (1)$$

where m is the start of the so-called short-term analysis window, of length N samples. The amplitudes $|X(k,m)|$ are considered as a set of K low-pass signals over time m. Considering the properties of the analysis window, the output signal $|X(k,m)|$ has a sampling rate of 40 Hz.

Second STFT A second STFT is then performed over the times m (times of the short-frames) for a specific frequency κ:

$$Y(k,\kappa,p) = \sum_{m=0}^{M-1} \left|X\left(k,p+m\right)\right| w(m) \exp\left(j2\pi\frac{\kappa}{M}m\right), \qquad (2)$$

where p is the start of the long-term analysis window of length M short-term frames. $w(m)$ is a rectangular window of $L = 2$ s duration and $H = 0.5$ s hop size. The final fingerprint code at time p is then created by grouping the frequencies k and κ (merging adjacent frequencies) until reaching a 36 dimensional concatenated vector.

2.2. Search strategy and post-processing

The search strategy is straightforward and consists in selecting among the code database the k nearest neighbors to the code analyzed. The result is a pair of matrices containing the audio item indexes and the time position in the item. The columns denote the sorted neighbors and the rows the timestamps p. Figure 1 shows an example of these matrices, where the each color denotes an item.

The post-processing correlates the results of adjacent frames, and prunes the accidental detections of erroneous items. As shown in the figure, it consists in finding pairs of elements in the matrix for which the time differences correspond in the analysis stream and the audio item detected (as in the green example). This procedure discards a huge amount of elements, and when applied to sequences of several correlated examples, only keeps the correct item index.

3. PERCEPTUALLY-SCALED CODE

Grouping of frequency in Bark-bands: Compared to the base-line code as indicated in Eq. 1, a first modification consists in grouping the frequencies k of $|X(k,m)|$ in $b \in [1, B = 24]$ Bark-bands [11]. Conversion of the frequencies in Hz to the Bark scale is given by

$$\mathrm{Bark}(f_k) = 13\arctan\left(\frac{f_k}{1315.8}\right) + 3.5\arctan\left(\frac{f_k}{7518}\right). \qquad (3)$$

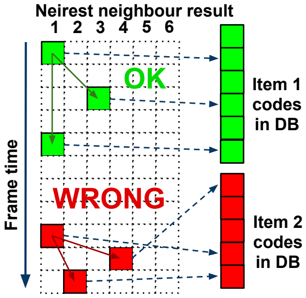

Fig. 1. Illustration of the post-processing scheme based on the matching of time differences between the stream and the item.

Grouping is performed by summing the energy of $|X(k,m)|^2$ for $f_k \in b$. The resulting signal for band b is denoted by $Z(b,m)$.

Sone-scale: We denote by $s(n)$ a pure audio signal and by $t(n)$ the impulse response corresponding to a transmission channel (equalization performed in radio or TV, loudspeaker or microphone characteristics, ...). Our recorded signal can then be expressed as $x(n) = s(n) * t(n)$, or $X(k) = S(k)T(k)$ in the frequency domain. The influence of the transmission channel can be avoided by expressing $X(k)$ in log-scale: $\log(X(k)) = \log(S(k)) + \log(T(k))$. Since $T(k)$ is considered constant over time, it will be located into the DC component of our second STFT. We therefore apply the log-scale to $Z(b,m)$ before computing the second STFT. In practice, a sone-scale is used rather than a log-scale:

$$Z_{\mathrm{sone}}(b,m) = \begin{cases} 1 & \text{if} \quad Z(b,m) < 1 \\ 2^{\frac{Z(b,m)-40}{10}} & \text{if} \quad 1 \le Z(b,m) \le 40 \\ \left(\frac{Z(b,m)}{40}\right)^{2.642} & \text{if} \quad Z(b,m) > 40 \end{cases} \qquad (4)$$

4. CODE SYNCHRONIZATION

A study on the code robustness in the subsection 5.1 below will show that, among a collection of common audio degradations, the sole shifting of the frame positions between the stream signal and the audio excerpts is a major cause of distortions on the fingerprint codes. Figure 2 shows an example of comparison between an original frame (in solid green) and a slightly shifted frame (in dotted blue).

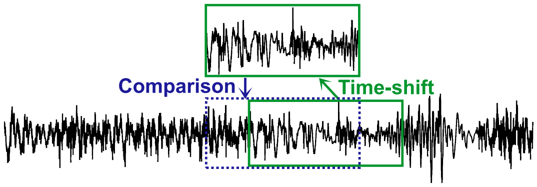

Fig. 2. Illustration of the time-shift audio degradation, observed when comparing frames with slightly different offsets.

In order to reduce the effect of time-shifts, we propose a basic scheme of synchronization through the detection of reliable timestamps. These timestamps are used to define the starting position of the frames (i.e. the p variable in Eq. 2) in the signal before the fingerprint computation. This implies that for each timestamp detected, the STFT frames are synchronized with it. The rest of the code computation in unchanged. In order to cope with the possible signal distortions, we rely on the onset positions, assumed to be robust under additional noise.

The onset detection phase follows here the algorithm proposed in [12] and [13], which has shown very good results in the latest MIREX contests. This algorithm aims at detecting transient peaks through the evolution of the center of gravity of the time domain energy (t_{cg}) in a short sliding window. The estimation of t_{cg} is based on the phase derivative, which proves to be more efficient, computationally speaking, than the direct estimation from the signal.

$$t_{cg} = \frac{\int_W -\frac{\partial \phi(\omega,t)}{\partial \omega} A(\omega,t)^2 d\omega}{\int_W A(\omega,t)^2 d\omega}, \tag{5}$$

where W stands for the sliding window, $A(\omega,t)$ and $\phi(\omega,t)$ for the spectral energy and the phase of the frequency ω at instant t. It is stated that a transient is likely to occur when t_{cg} decreases under a given threshold C_e, i.e. when the center of gravity reaches the center of the window, from its right part. Please refer to the original article [13] for more precision on the onset detection algorithm.

The threshold C_e is empirically tuned during the training phase, to control the expected mean rate of onsets detected. However, the use of a threshold does not guarantee a regular rate of onset detections, nor does it guarantee the detection of any onset for a given track. This issue will be addressed in the future.

5. EVALUATION

5.1. Study on the code robustness

In order to achieve the proper identification of audio items in a signal, the fingerprint code must be robust against common distortions on the signal, while remaining discriminant between the different items' codes. To assess the reliability of our code, we reproduce here the experimental protocol proposed by Haitsma and Kalker [1].

This experiment consists in applying a series of controlled distortions on clean audio samples, and computing the distance between the fingerprint codes extracted from the clean and the distorted samples. While the authors originally focus on 4 audio tracks under copyright, we extend here this study to a collection of 500 tracks[2] extracted from the public dataset Magnatagatune [14].

The set of distortions used here is a subset of [1] :

- **MP3 encoding/decoding** at a low bitrate (8 Kbps),
- **GSM encoding/decoding** at full rate,
- **Amplitude compression**, with ratios 8.94:1 if $|A| \geq -28.6$ dB; 1.73:1 if -46.4 dB $< |A| < -28.6$ dB; 1:1.61 if $|A| \leq -46.4$ dB,
- **Equalization** with a 10-band equalizer,
- **Noise addition** using uniform white noise,
- **Time shifting** with a delay varying between 0.02 s and 0.5 s,
- **Broadcast**, simulating a distorted real broadcast emission by cascading equalization, dynamic compression and MP3 encoding.

The mean distances (along with the standard deviation, std) between distorted and original codes are indicated in Table 1. In order to give an upper bound, we indicate in the last row the distance with codes computed from random audio samples outside the dataset.

While this experiment indicates that the code is quite robust against basic distortions, such as additional white noise or equalization, the audio encoding schemes (GSM and MP3) induce more noise, as does our simulation of a broadcast emission. However, time shifting is clearly an important source of distortion, even with a slight shift delay of 0.1 s (i.e. only 5% of the window size used to compute the code). Our measures actually show that the evolution of the mean distance between the codes is roughly proportional to the shift delay when the latter is between 0 and the hop size (H=0.5 s).

Degradation	mean	std
Noise addition	0.28	1.47
Compression	0.99	0.85
Equalization	1.23	1.63
GSM	1.53	1.42
MP3 8 Kbps	2.52	2.21
Broadcast	3.67	3.64
Shift 0.05 s	0.72	0.88
Shift 0.1 s	1.85	1.52
Shift 0.25 s	4.28	2.58
Shift 0.5 s	7.38	3.82
Random	20.66	9.02

Table 1. Mean and standard deviation of the code distances under different audio degradations.

In the case of a live analysis of broadcast audio streams, this latter observation is a major bottleneck to the performance. Indeed, since the item occurrences are obviously not synchronized with the training items, the periodic pattern of frames with a hop size of H induces an expected mean shift of $H/4$ between the frame of an occurrence and the closest frame in the database.

5.2. Corpus for the evaluation

The evaluations are performed on a part of the train set of the Quaero project. This set consists of a collection of 10 whole days (i.e. 240 hours) of broadcast radio stream encoded in WMA at a very low bitrate (about 10 kbps), and is provided by the media monitoring company Yacast, as a partner of the Quaero project. The reference items are excerpts of 30 s of the same quality. The corpus contains around 2000 occurrences in the streams, searched among a database of 1000 training items.

The corpus has been carefully re-annotated, by locating very precisely the start and end times of each item signal in the occurrences in the streams (with a precision of about 0.01 s). This allows us to provide a reliable evaluation of the correspondence between the onsets found in the items and in the streams.

5.3. Assessing the onset selection scheme

For each onset detection in the stream, we determine the minimum delay between its timestamp (minus the start time of the occurrence in the stream) and the one of the onset detected on the corresponding audio item. Figure 3 shows the distribution of these minimum delays over the whole set of onsets detected in the stream. The empirical distribution estimated from the system with regular frames is shown in blue. In dotted black is represented the corresponding expected distribution, that is uniform between 0 and half the hop size ($\frac{H}{2} = 0.25$ s), assuming that the occurrences are randomly drawn in the streams. The gain of using the onset detection scheme is quite clear from observing the distribution in red. Indeed, a large proportion of the stream frames are clearly closer to the nearest track frame than with regular frames. Since the delay between two adjacent frames is not upper bounded anymore, a counterpart is observed in the presence of non-zero probabilities for delays above $\frac{H}{2}$. How-

[2] The exact list of audio items used for this experiment can be found at `http://www.mathieuramona.com/Main/Mag500`.

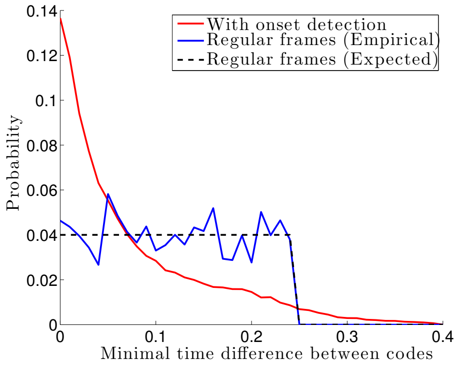

Fig. 3. Distributions of the time delay between each stream code and the nearest code in the corresponding item.

ever Tab. 2, that indicates the empirical mean minimal delay on all the frames, shows that the onset detection remains advantageous, when compared to a regular frame basis.

Onset detection	0.104
Regular frames	0.124

Table 2. Empirical mean of the minimum delays between corresponding codes in the database and the stream.

5.4. Comparison against state-of-the-art

This second evaluation compares the performance of the baseline system with and without the Bark and Sone-scale improvements introduced here, along with the onset detection proposed later on. A comparison with two state-of-the-art methods is also proposed, relying on a re-implementation of the Philips system [1] and of the Shazam system [2], based on the code provided by Dan Ellis [15]. The Philips evaluation includes both a version without and with the improvement involving least reliable bit tests, over 5 bits (detailed in [1]). The score metric, based on the evaluation protocol of the Quaero campaign, is simply the difference of the correct identification rate and the false alarm rate. We also consider here the missed detection rate (equal to 1 minus the correct rate). Results are shown in Tab. 3. While it is not possible to comment on our implementation of the Shazam system, since most of the parameters are not specified by the author, the results nevertheless clearly show the relevance of our fingerprint code definition, when compared to the state-of-the-art, especially the Philips system, which is thoroughly detailed in their publication. Moreover, results show that the improvements detailed here increase the performance of our original baseline system.

6. CONCLUSION AND FUTURE WORKS

We have proposed here an original fingerprint code for audio identification, based on a more robust scheme than the major state-of-the-art methods, along with two improvements. The first one is based on perceptual scales of frequency and amplitude, and the second aims at improving the synchronization between the compared codes, through onset detection. The IRCAM system yields excellent results that clearly outperform our implementation of the system from Philips and Shazam.

Future works will be dedicated to the improvement of the onset detection scheme, in order to tighten the minimal delay distribution further more, will guaranteeing a more regular temporal distribution of detected onsets. We hope to emphasize more clearly the advantage of the onset detection step by providing comparative results on a much larger test set. Another major perspective for our system is to address the problem of track occurrences with altered time-scales.

System	False	Missed	Global
	Alarm	Detection	Score
IRCAM Baseline	0.4	5.4	94.2
IRCAM Bark & Sone	0.1	4.3	95.6
IRCAM Onsets	0.0	3.8	**96.2**
Philips	0.0	12.1	87.9
Philips 5 bits	0.0	10.1	89.9
Shazam	10.3	11.3	78.4

Table 3. Comparative results on the Quaero corpus. The global score is the difference of the correctly identified and false alarm rates.

7. ACKNOWLEDGMENT

This work was partly supported by the "Quaero" Program funded by Oseo French State agency for innovation.

8. REFERENCES

[1] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *Proc. ISMIR '02*, 13-17 octobre 2002.

[2] A. Li-Chun Wang, "An industrial-strength audio search algorithm," in *Proc. ISMIR '03*, 2003.

[3] T. L. Blum, D. F. Keislar, J. A. Wheaton, and E. H. Wold, "Method and article of manufacture for content-based analysis, storage, retrieval, and segmentation of audio information," juin 1999.

[4] E. Allamanche, J. Herre, O. Hellmuth, B. Fröba, T. Kastner, and M. Cremer, "Content-based identification of audio material using MPEG-7 low level description," in *Proc. ISMIR '01*, 2001.

[5] L. Worms, "Reconnaissance d'extraits sonores dans une large base de données," Master thesis, Ircam, 1998.

[6] X. Rodet, L. Worms, and G. Peeters, "Brevet FTRandD/03376: Procédé de caractérisation d'un signal sonore - Patent 20050163325 method for characterizing a sound signal," 2003.

[7] G. Peeters, A. Laburthe, and X. Rodet, "Toward automatic music audio summary generation from signal analysis," in *Proc. of ISMIR*, Paris, France, 2002, pp. 94–100.

[8] P. Balabko, "Speech and music discrimination based on signal modulation spectrum," Tech. Rep., 1999.

[9] S. Greenberg and Brian E. D. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," in *Proc. of IEEE ICASSP*, Munich, Germany, 1997, vol. 3, pp. 1647–1650.

[10] L. Atlas and Sh. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Applied Signal Processing*, vol. 7, pp. 668–675, 2003.

[11] E. Zwicker and E. Terhardt, "Analytical expression for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Am.*, vol. 68, pp. 1523–1525, 1980.

[12] G. Peeters, *Modèles et modélisation du signal sonore adaptés a ses caractéristiques locales*, Phd thesis, Universite Paris VI, 2001.

[13] A. Röbel, "Onset detection in polyphonic signals by means of transient peak classification," in *ISMIR/MIREX*, 2006.

[14] E. Law and L. Von Ahn, "Input-agreement: A new mechanism for collecting data using human computation games," in *Proc. CHI*, 2009.

[15] D. Ellis, "Robust landmark-based audio fingerprinting," `http://labrosa.ee.columbia.edu/matlab/fingerprint/`.