A Public Audio Identification Evaluation Framework for Broadcast Monitoring

Mathieu Ramona [†] Sébastien Fenet [‡] Raphaël Blouet [§]
Hervé Bredin [¶] Thomas Fillon [‡] Geoffroy Peeters [†]

[†] IRCAM
1, place Igor-Stravinsky
75004 Paris
France

[‡] Institut Télécom
Télécom ParisTech
CNRS-LTCI
37-39, rue Dareau
75014 Paris
France

[§] Yacast
4, rue Paul Valery
75016 Paris
France

[¶] LIMSI-CNRS
B.P. 133
91403 Orsay Cedex
France

(*Received 00 Month 200x; In final form 00 Month 200x*)

This paper presents the first public framework for the evaluation of audio fingerprinting techniques. Although the domain of audio identification is very active, both in the industry and the academic world, there is nowadays no common basis to compare the proposed techniques. This is because corpuses and evaluation protocols differ between the authors. The framework we present here corresponds to a use-case in which audio excerpts have to be detected in a radio broadcast stream. This scenario indeed naturally provides a large variety of audio distortions that makes this task a real challenge for fingerprinting systems. Scoring metrics are discussed, with regard to this particular scenario. We then describe a whole evaluation framework including an audio corpus, along with the related groundtruth annotation, and a toolkit for the computation of the score metrics. An example of application of this framework is finally detailed. This took place during the evaluation campaign of the Quaero project. This evaluation framework is publicly available for download and constitutes a simple, yet thorough, platform that can be used by the community in the field of audio identification, to encourage reproducible results.

1 Introduction

Audio identification is a special case of audio event detection that covers the detection and the identification of an audio excerpt (a music track, an advertisement, a jingle ...) in an audio recording (either a short excerpt or a broadcast stream). Two techniques are used in that purpose : *audio watermarking*, that relies on the embedding within the audio signal of meta-information, robust to common alterations, and *audio fingerprinting* (sometimes called *audio hashing*), where audio occurrences are detected through the recognition of code signatures extracted from short snippets of the signal. These signatures are designed to make a compact representation of the audio content, linked to some perceptually relevant cues, that remains robust to typical distortions observed on audio signals, such as dynamic compression, various encodings, equalization, time scale changes... Since audio watermarking implies an initial processing of the signal source to embed the watermark, it cannot be applied to unknown signals. This paper hence focuses on audio fingerprint techniques for audio identification.

The audio identification technology underlies several key applications, including broadcast monitoring, internet content identification, copyrights control or interactive behavioral targeted advertising. This explains a great effort on contributions in the community during the last decade, despite the relative novelty of the domain, mostly from industrial actors, such as Philips (Haitsma and Kalker [2002]), Shazam (Wang [2003]), Google (Weinstein and Moreno [2007], Mohri et al. [2008]), Fraunhofer (Allamanche et al. [2001], Herre et al. [2001]), or Microsoft (Burges et al. [2003], Burges et al. [2002]). Many propositions also emerge from the academic research area: Ircam owns a technology based on a double-nested Fourier transform (Rodet et al. [2003]), NTT Basic Research Lab have proposed the *active search* algorithm (Smith

This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

et al. [1998]), and the Pompeu Fabra University owns a technology based on the so-called *audioDNA* model (Neuschmied et al. [2001], Cano et al. [2002]).

However, it remains impossible nowadays to compare the different systems described in the literature, since no common framework or corpus has been proposed, apart from the TRECVid evaluation on video copy detection (Smeaton et al. [2006]). Indeed, most of the evaluations are applied on private corpuses, of which volume and nature varies greatly between the articles. Also, the evaluations protocols, as well as the scoring metrics, are generally based on different use-cases and reflect different underlying priorities for the authors, which induce very different insights and conclusions on the algorithms.

Moreover, since the key point of audio fingerprinting is the detection of audio events under common distortions, evaluations are often based on the application of controlled synthetic audio distortions applied to audio items, that do not necessarily reflect the constraints of a real-world use-case.

We thus propose in this article the first public evaluation framework focused on audio identification, based on a scenario involving the detection of audio excerpts in broadcast radio streams. The framework consists of a public corpus and an evaluation toolkit named PyAFE. This corpus is not based on artificial audio degradations but on the real-world degradations induced by the radio broadcast production, which implies a wide variety of combined distortions. This corpus hence makes a challenging and realistic task for audio identification methods. Relevant metrics, related to the use-case, are also provided, along with a discussion about their respective characteristics.

These contributions define a consistent evaluation framework that is made publicly available to the community, in order to encourage benchmarking in the field of audio identification, instead of private evaluations. We will then thoroughly describe the process and the results of the Quaero project[1] first evaluation campaign on audio identification, held in September 2010, that is based on this evaluation framework.

This paper is structured as follows. The various evaluation schemes in the literature will be briefly examined in Section 2, in order to give an outline of the common protocols, as well as the synthetic audio degradations generally used to assess the robustness of the fingerprint codes. A new evaluation framework will be proposed in Section 3, that includes a corpus presented in Section 3.1 and an evaluation toolkit presented in Section 3.3. The latter includes the implementation of the scoring metrics discussed in Section 3.2. Then, Section 4 will detail an example of application of this framework on the evaluation campaign of the Quaero project, along with the results of the participants. Some comments and short-term perspectives on the framework will finally be given in Section 5.

2 Audio identification evaluation

2.1 *Audio identification in the context of event detection*

Audio event detection covers a wide range of scenarios of audio analysis. Depending on the type of events that must be recognized, their duration, their acoustical variability, varying techniques are employed.

Frame-based classification methods are generally used when each event is defined by an acoustic source, such as gun shots (Clavel et al. [2005]), applause or cheers (Cai et al. [2003]). More general acoustic classes, like speech or music, can also be considered as audio events, and involve a very large literature (Lin et al. [2005], Ramona and Richard [2009], ...). This kind of problem implies the use of statistical methods to cover the whole range of variability of the acoustic sources, and focuses on the possible confusion beween the classes.

On the other hand, events may not denote sources, but audio signals themselves, as in the detection of jingles (Pinquier and André-Obrecht [2004]), advertisements or musical tracks. This scenario dramatically

[1] Please consult `http://quaero.org` for more information on the project.

reduces the scope of variations of the event occurrences, which is restricted to typical audio degradations that affect the signal while keeping it perceptually recognizable.

This case covers what is denoted by *audio identification* in the scope of this paper. It is characterized by the fact that a single example is available for each event in the training process, and involves specific techniques, typically audio fingerprinting and audio watermarking. Another typical aspect of this scenario is the very large number of events generally involved. For instance a music track identification task can scale up to several million different classes. The confusion between classes is thus critical, since the slightest overlap might lead to a very large number of false alarms. The audio identification scenario thus focuses on the compromise between discrimination and robustness to audio degradations.

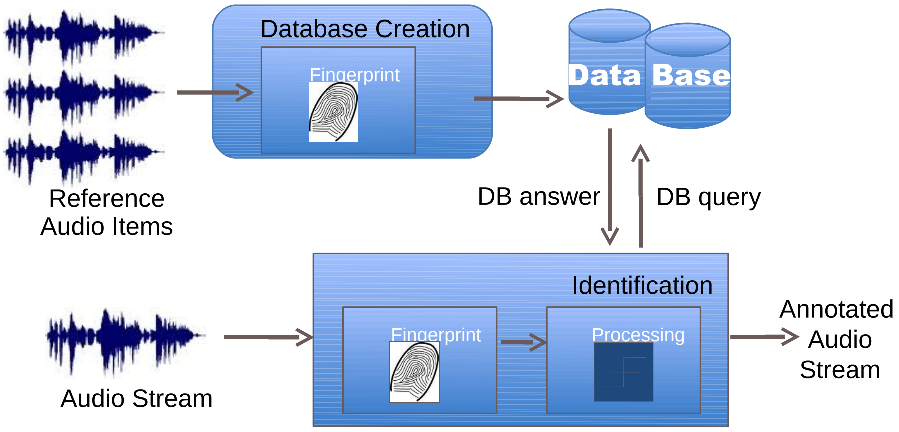


Figure 1. Illustration of the audio fingerprinting workflow.

Figure 1 illustrates the typical workflow of an audio fingerprinting system. As all machine learning systems, audio fingerprinting requires two modes. The first one is the *database creation*, where the set of target audio items is processed by the system for the extraction of their fingerprints. All fingerprints are stored in a database and allow to link a given content to tag or metadata. The other mode allows *audio identification*, based on the fingerprints computed from the audio stream.

While the framework presented here is originally dedicated to the evaluation of works on audio fingerprinting, it can indeed be used for any audio identification task, with no restriction on the technique employed.

The following sections detail the typical audio degradations found in the literature, as well as the evaluation protocols for audio identification systems.

2.2 *Considering usual audio degradations*

A very diverse panel of audio degradations can be found in the literature, designed to reproduce most of the audio effects that can be applied to an audio signal, affecting its quality, without changing its semantic content (i.e. what is perceptually received by the listener). Most of them are inspired by the studies on robustness of Haitsma and Kalker [2002] and Allamanche et al. [2001].

The main issue of this discussion is the distinction between three kinds of degradations:

- *Numerical degradations*, by far the most convenient to apply, since they can be simulated numerically.
- *Acoustic degradations*, that involve somehow a conversion to acoustic waves. Their simulation requires more equipments (microphones, loudspeakers...), but remains possible.
- *"Real-world" degradations*, are a much more complex blend that combines numerous degradations and require a whole sound production chain, e.g. broadcast radio production and transmission.

We detail here most of the degradations found in the literature, that fall in the first two categories:

Audio encodings (numerical):

- *MP3 or WMA encoding/decoding,* from very low (8 kbps) to usual bitrates (128 kbps),
- *Real Media encoding/decoding,*
- *GSM encoding/decoding,* at full rate, with a controlled carrier to interference (C/I) ratio,
- *Resampling,* down to half the sample rate and up again.

Filtering (numerical):

- *All-pass filtering,* using an IIR filter,
- *Equalization,* with a 10 to 30-band equalizer,
- *Band-pass filtering,*
- *Telephone band-pass,* between 135 and 3700 Hz,
- *Echo filter,* simulating old time radio.

Noise addition, of controlled SNR (numerical):

- *White noise addition,* using a uniform or gaussian white noise
- *Real-world noise addition,* adding a capture of noisy environment (e.g. a crowded pub)
- *Speech addition.*

Dynamics changes (numerical):

- *Amplitude dynamic compression,*
- *Multiband companding,* specifically defined in the TRECVid evaluation,
- *Volume change,* affecting the global volume with a constant of slightly varying factor.

Temporal changes (numerical):

- *Time scale modification :* up to +4% and -4% without affecting the pitch,
- *Linear speed change:* up to +4% and -4%, with both tempo and pitch affected,
- *Time shift:* the signal is slightly shifted in order to affect the alignment of the temporal frames (This will be discussed thoroughly in the next section).

Acoustic conversions (acoustic):

- *D/A A/D conversion:* using a commercial analog tape recorder,
- *Re-recording:* through a loudspeaker/microphone chain.Possibly in a noisy environment.

While these reproducible degradations are shared by almost all the experiments in the literature (Belletini and Mazzini [2010], Liu et al. [2009], Jang et al. [2009]), very few examples of real-world degradations are found (Betser et al. [2007], Cano et al. [2002], Pinquier and André-Obrecht [2004]), all based on radio broadcast recordings. In fact, such recordings include most of the artificial degradations detailed here: the signal is generally numerically encoded, and all the filtering processes, such as all-pass and band-pass filtering and equalization, are very common effects in radio broadcast production. This also stands for time scale, pitch shift and linear speed modifications, that are used especially on musical tracks. Finally, real-world noise addition is also observed, since most radio shows hosts speak on the instrumental introduction of the songs.

Synthetic distortions are strictly controlled and studied independently, whereas real-world radio broadcast signals provide a varied set of complex combinations between all these distortions. Finally, the audio streaming constraint induces the loss of alignment between the original audio excerpts and the observed audio frames, which will be discussed later on.

These remarks motivate our proposition of an evaluation framework based on a radio broadcast corpus.

2.3 *Existing evaluation protocols*

As stated in the previous section, a large majority of the past contributions rely on synthetic audio degradations. This statement constraints the evaluation protocol and the corpus. Indeed, in most cases, the corpus consists in a collection of unrelated audio (mostly music) tracks. The queries are subsets of the original tracks, on which various degradations are applied. The dominant evaluation protocol thus consists in detecting in the queries the original tracks learnt from the corpus.

However, it remains focused on the false rejects (also called false negatives or deletion errors). A slightly different protocol involves a collection of matching and non-matching pairs of audio excerpts. The matching pairs include original clean tracks and their degraded versions, and the non-matching ones are arbitrary. Through the number of matching and non-matching pairs, the balance between false rejects and false alarms (false positives or insertion errors) can be controlled.

Another interesting approach deals with the distances between fingerprint codes themselves. It cannot evaluate directly the performance of an algorithm, but the comparison of the distributions of matching and non-matching fingerprint pairs gives very useful insights on the discriminativity of a fingerprint code. This was initiated by Haitsma and Kalker [2002], who measure the Bir Error Rate (BER, i.e. the Hamming distance) on the binary Philips fingerprint.

As stated in the previous section, the last protocol encountered in the literature is based on real broadcast recordings that include occurrences of the corpus audio items. The drawback is a reduced control over the number of occurrences, but the "real-world" combinations of degradation and the presence of long sections with no occurrence enforce the realism of the evaluation.

Another major argument in favor of broadcast-oriented evaluation is the arbitrariness of the occurrences position in the audio streams, which imply random de-synchronization between the original item signal and the occurrence signal. Indeed, as stated in a previous publication (Ramona and Peeters [2011]), a slight time-shift between the original audio and the degraded audio induces distortions on the fingerprint that are more important than most degradations. Many evaluations skip this issue since they apply degradations directly on the original audio sample and thus preserve the temporal alignment .A real-world corpus implicitly induces random time-shifts in the occurrences.

Evaluation protocols in the literature also cover a wide range of score metrics. Generally used with the corpus-subset-queries protocol, the accuracy rate (the number of queries correctly identified) is by far the most common criterion. However, it does not cover false alarms (false positives). A more thorough approach is to use the false negative / false positive (FN/FP) pair as a measure of performance. The TRECVid evaluation plan for Copy Detection (Smeaton et al. [2006]) also defines a refined metric that specifically balances false negatives and positives.

When the method involves a threshold or a tunable parameter, a Receiver Operating Characteristic (ROC) curve is used to illustrate the evolution of the FP/FN measures with regard to the parameter. The Precision and Recall metrics (derived from the FP/FN measures) are also used in a similar way. Finally, since audio identification is often based on a nearest neighbor scheme, instead of computing the accuracy on the first result, a tolerance can be set to accept detections that are within the N best ranked. This measure is denoted by "Top-N".

Table 1 summarizes the different evaluation protocols presented here, along with the score metrics, the size of the corpuses and the number of queries. The last line describes the TRECVid Copy Detection evaluation task, mentioned earlier. It is the only other evaluation campaign related to our subject, but it is mostly dedicated to video indexing, and does not propose a specific task for audio identification.

It can be noted that, even though audio identification is typically considered as a large-scale problem, most of the evaluations are limited to a few thousands, or a even a few hundred tracks in the corpuses or as queries. We hope the framework provided here, and described in the next Section, offers a larger scale than most evaluations mentioned in the table.

Articles	Corpus	Queries	Protocol	Metric
Allamanche et al. [2001](Fraunhofer)	15 k	NA	Subsets	Acc + Top 10
Cano et al. [2002]	50 k	12h (104)	Subsets+Broad	FP/FN
Haitsma and Kalker [2002] (Philips)	4	4	Subsets+BER	Nb hits
Wang [2003] (Shazam)	10 k	250	Subsets	Acc + FP
Pinquier and André-Obrecht [2004]	200	10h (132)	Broad	Acc
Seo et al. [2006]	8 k	NA	Subsets	ROC (Pre/Rec)
Covell and Baluja [2007] (Google)	10 k	1000	Subsets	Acc + ROC (FP/FN)
Betser et al. [2007]	30	18h (243)	Broad	Recall
Kim and Yoo [2007]	NA	7 M	Pairs	ROC (FP/FN)
Mohri et al. [2008]	15 k	3,600	Subsets	Acc
Liu et al. [2009]	13 k	2400	Pairs	Acc + ROC (FP/FN)
Jie et al. [2009]	500	NA	Subsets	Top 1,5,10,20,50
Jang et al. [2009]	100	44 k	Pairs	Acc + ROC (FP/FN)
Belletini and Mazzini [2010]	100 k	NA	Subsets	Acc
Li et al. [2010]	1,822	100	Subsets+BER	Top 1,5,10 + FP/FN
Smeaton et al. [2006] (TRECVid)	400h	12000	Subsets	Balanced FP/FN

Table 1. Comparative list of the evaluation protocols in the literature, specifying corpus and queries sizes, protocols and score metrics.

3 Proposed evaluation framework

3.1 *Broadcast-oriented corpus*

The evaluation corpus described in this paper comes from a basic media monitoring use-case. Given a set of target musical tracks, it consists in determining if and when an audio item has been broadcasted, i.e. establishing the time of broadcasting and the identity of the target track occurrences.

The evaluation corpus has been drawn in the framework of the sub-task *Audio Identification/ Fingerprint* of the Quaero project. It provides a corpus of target audio items, for the database building, and several continuous radio broadcast streams, for the identification. The annotation process was done semi-automatically and entirely checked by human operators. For the evaluation within the Quaero project, the corpus is characterized by around 8,000 audio items of one minute. These audio excerpts correspond to audio segments previously broadcasted and manually annotated and extracted. There is one excerpt per audio item, available in `RAW` format, little-endian, with 16 bits per sample, at a sampling rate of 11025 Hz.

The test audio streams consists in full days of radio stream, from different stations, captured and saved in succeeding five-minutes chunks, encoded in `AAC`, with a bit-rate of 64 kb/s and a sampling rate of 11025 Hz. The item signatures may not be completely broadcasted in the streams. The beginning of a signature is not known.

For legal issues, we are not allowed to distribute the whole stream associated to a media. We hence have built an *artificial* stream made of a concatenation of short audio excerpts (from 3 s to 45 s) coming from different media. The stream is made up of around 8,000 audio items provided in 4 files of 4 hours. As the proposed corpus includes real radio broadcasted items from a set of 15 media, it is likely to cover all the challenges of audio identification for radio monitoring. For each target audio item, 30 seconds of audio data are provided to compute the fingerprint signature.

The provided corpus is partly synthetic, since it relies on a concatenation of excerpts. Nevertheless, it is important to note that the excerpts themselves come from real-world signals, and thus provide realistic degradations, as stated in section 2.2, considered as the main issue for a serious study on robustness.

The ground truth is determined by a set of XML files, one for each media file. Of course the XML annotation only specifies the items that are present in the corpus delivered. The annotation, as stated earlier, comes from an audio identification engine manually checked, with a precision of about 1 s. Each file lists occurrences of each target audio item, with the following XML structure:

```
<MusicTrack>
    <id>123456</id>
    <idMedia>548</idMedia>
    <title>Some kind of wonderful</title>
    <artist>Grand Funk Rail road</artist>
    <album>Caught in the act</album>
    <genre>Pop/Rock International</genre>
    <startDate>2010-07-05 00:03:43</startDate>
    <endDate>2010-07-05 00:03:55</endDate>
</MusicTrack>
```

where:

- `<id>` identifies the audio item on air between `<startDate>` and `<endDate>`,
- `<idMedia>` identifies the source media, in order to extract identification scores for specific media,
- `<title>`, `<artist>`, `<album>` and `<genre>` are various metadata on the target item. This makes it possible, for instance, to extract identification scores by genre,
- `<startDate>` and `<endDate>` respectively indicate the start and end time of the elements denoted by the ID. However, since the item signatures provided for the database build are subsets of the original songs (as stated earlier), these time limits are actually larger than the actual occurrence of the item signature alone. This issue will be thoroughly discussed in the next section.

Audio streams and signature files are freely available for academic research use at:
`http://pyafe.niderb.fr`.

3.2 *Scoring*

The present use-case implies the following constraints :

- Occurrences of known audio items are to be detected in an audio stream.
- The audio items are only known through short snippets called *audio signatures*.
- The position of these signatures within the original tracks is unknown.

Please note that if audio tracks in the stream are not occurrences of any item in the corpus, they are not considered as "occurrences". They are part of the "no-item" areas described later on, like any unknown signal, since they cannot be recognized.

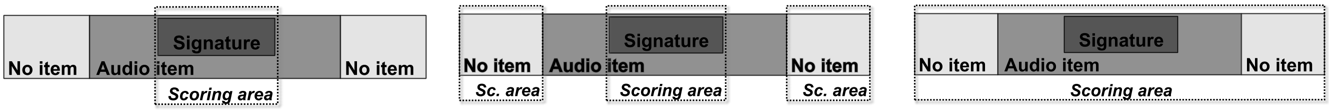
3.2.1 ***Scope of evaluation.*** Since only a part of the audio items (the signature) is taken into account in the training process, one could consider the sole signature area being the track itself, as in Figure 2(a), where an occurrence of a target item is shown in medium gray, between two areas with no item in light gray ; the signature snippet is shown in dark gray. In this context, detections timestamps can be evaluated precisely.

The exclusion of the "no-item" areas greatly reduces the scope of evaluation. The evaluation should rather imply the whole signal. But since music tracks generally imply repetitive structures (chorus, verses...), the areas of the song outside the signatures are likely to present high correlations with the signatures themselves[1]. It would therefore be preferable to exclude the latter from the evaluation scope, as shown in Figure 2(b), since the notions of missed detection and false alarms are ambiguous outside the signatures.

However, as stated earlier, the position of the signatures in the tracks is unknown ; it is thus impossible to define the scope of evaluation according to it. Therefore, the case described in Figure 2(c), implying the whole signal, is the only one applicable here. Although this configuration is theoretically less reliable than

[1] This correlation cannot be quantified, since it highly depends on the fingerprint code design, but it suffices to say that it is used as a basic assumption for automatic musical structure retrieval (see Peeters et al. [2002], for an example based on fingerprint techniques)

the previous schemes, this issue is answered by not considering the temporal position of the detections in the items, as discussed in the next paragraph on score metrics.



(a) signature only (b) signature + "no-item" (c) all the signal

Figure 2. Possible evaluation scopes.

3.2.2 ***Score metrics.*** The metric for the evaluation of audio identification is based on a punctual event detection scheme, which means that only instants of decision are taken into account, instead of segments. The possible segmentation of the signal into frames is exclusively related to the audio identification process, and is not relevant to the following score metrics.

A previous section exposed the main score metrics found in the literature. Most evaluations are based on the accuracy (e.g. number of correctly detected occurrences over the number of occurrences), which is here equivalent to the Precision measure, and inversely proportional to the false reject rate (or deletion error rate). The Recall measure is similarly related to the false alarm rate (or insertion error rate). Note that in the context of audio identification, no distinction is made by the community between substitutions (e.g. mistaking an item for another one) and insertions. ROC curves are also commonly used, but not adapted to an evaluation framework designed to compare different algorithms. Indeed, in a proper benchmark, each system is evaluated as a standalone application that does not require any parameter tuning.

This framework is thus restricted to false reject and false alarm measures. The TRECVID evaluation plan (Smeaton et al. [2006]) defines several balances between the two measures, but the present use-case has no preference for a specific error. The counting of the false alarms is discussed in this section.

Let us denote a collection of N audio occurrences of items. Each occurrence n is between time boundaries t_n^{sta} and t_n^{end} in the signal, and is related to an item of index i_n in the database. The evaluation considers a set of D punctual detections. Each detection d is related to an item index j_d and instant τ_d. As stated before, the signature is not precisely located, so that a correct detection only involves the observation of at least one detection of the correct item within the scope of the occurrence (the medium gray "Audio item" scope in Figure 2). The number of correct detections (Accuracy) is thus defined as:

$$S_{\mathrm{OK}} = \mathrm{Card}\{n \in [1 \ldots N],\ \exists\, d,\ j_d = i_n \ \text{ and } \ t_n^{sta} \leq \tau_d \leq t_n^{end}\}. \tag{1}$$

The number of false rejects is straightforward and is implicit in the Accuracy definition ($S_{\mathrm{FR}} = N - S_{\mathrm{OK}}$). The global score is defined as the following rate:

$$R = 1/N \cdot \left(S_{\mathrm{OK}} - [S_{\mathrm{FA}} + S_{\mathrm{FA}}^{\mathrm{out}}]\right), \tag{2}$$

where S_{FA} and $S_{\mathrm{FA}}^{\mathrm{out}}$ respectively denote the false alarm rate within and outside the occurrences.

The expression of R depends on the definition of the false alarm rates, which depends on the tolerance accepted. The most straightforward definition counts each false alarm as one error, as shown in Figure 3(1):

$$\begin{cases} S_{\mathrm{FA},1} = \mathrm{Card}\{d \in [1 \ldots D],\ \exists\, n,\ t_n^{sta} \leq \tau_d \leq t_n^{end} \ \text{ and } \ j_d \neq i_n\}, \\ S_{\mathrm{FA},1}^{\mathrm{out}} = \mathrm{Card}\{d \in [1 \ldots D],\ \nexists\, n,\ t_n^{sta} \leq \tau_d \leq t_n^{end}\}. \end{cases} \tag{3}$$

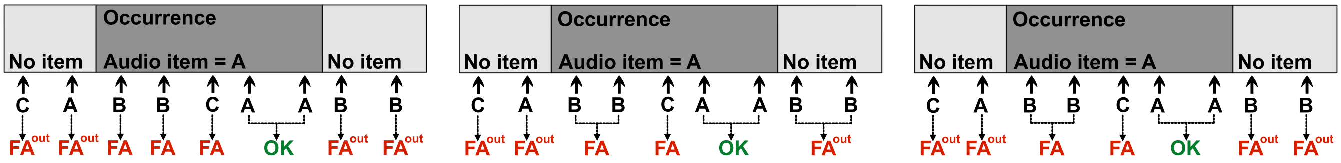
However, this measure is strongly biased since false alarms are upper bounded by D (i.e. can be arbitrarity numerous), while correct detections are bounded by the number of occurrences N. The non-homogeneity between the two measures can lead to negatives scores, especially with a high number of detections.

The scheme presented in Figure 3(2) corrects this balance by grouping as a single error the false alarms of a same wrong item in a given area. However, false alarms of distinct items are still counted separately, otherwise the balance would be strongly biased in favor of correct detections. In order to unify the evaluation scheme, the areas between the occurrences ("No item" areas in Figure 3) are considered as occurrences where only false alarms are counted (no item can be detected). Hence, the evaluation does not consider individual detections, but only the items detected within areas. False alarms are expressed as follows:

$$\begin{cases} S_{\text{FA},2} = \text{Card}\{n \in [1 \dots N],\ \exists\, d,\ t_n^{sta} \leq \tau_d \leq t_n^{end} \text{ and } j_d \neq i_n\}, \\ S_{\text{FA},2}^{\text{out}} = \text{Card}\{n \in [1 \dots N],\ \nexists\, d,\ t_n^{sta} \leq \tau_d \leq t_n^{end}\}. \end{cases} \tag{4}$$

The last proposition, illustrated in Figure 3(1.5), is a compromise between the previous two. On a musical radio station, the "no-item" areas are rather short, and only contain transitions between musical tracks. Other stations, though, can produce mostly talk shows, lasting several hours. The grouping of false alarms of the same item separated by long durations is therefore not relevant. The metric 1.5 counts the false alarm by items within the occurrences, and by detections outside them :

$$\begin{cases} S_{\text{FA},1.5} = S_{\text{FA},2} \\ S_{\text{FA},1.5}^{\text{out}} = S_{\text{FA},1}^{\text{out}} \end{cases} \tag{5}$$



(1) Score 1: each FA = 1 error in or outside the occurrences.

(2) Score 2: FAs of same ID in a same area = 1 error.

(1.5) Score 1.5: FAs not grouped in "no-item" areas.

Figure 3. Comparison of the false alarms counting methodologies.

The three metrics we have detailed are jointly provided and used in the present evaluation framework. None is favored *a priori* over the others. The global scores stand as follows:

$$R_1 = 1/N \cdot \left(S_{\text{OK}} - [S_{\text{FA},1} + S_{\text{FA},1}^{\text{out}}]\right) \tag{6}$$

$$R_2 = 1/N \cdot \left(S_{\text{OK}} - [S_{\text{FA},2} + S_{\text{FA},2}^{\text{out}}]\right) \tag{7}$$

$$R_{1.5} = 1/N \cdot \left(S_{\text{OK}} - [S_{\text{FA},2} + S_{\text{FA},1}^{\text{out}}]\right) \tag{8}$$

3.3 *The* PYAFE *evaluation toolkit*

As simple as it may seem at first, the actual implementation of these scoring metrics can be complex. Researchers should not have to implement their own version. This would increase the risk of getting several (yet differing) implementations of the same scoring metric, therefore compromising the fair comparison paradigm that we aim at. Moreover, as state-of-the-art audio fingerprinting systems are getting really close to perfection, it becomes crucial that the performance of two systems can be compared accurately. That is why we introduce the **PYAFE**[1] toolkit:

PYthon **A**udio **F**ingerprinting **E**valuation

It was developed in the framework of the *Evaluation* work-package of the Quaero project and is made freely available as open-source software downloadable from:

`http://pyafe.niderb.fr`

[1] **PYAFE** is pronounced like the last name of Edith Piaff, the famous french singer.

It was designed as a modular piece of software, in order to be easily extended in the future:

- two modules provide the necessary functions to parse the groundtruth and detection XML files (whose formats are described on the **PyAFE** website),
- the core module includes the implementation of score metrics R_1, R_2 and $R_{1.5}$ described in Section 3.2. The number of correct detections S_{OK}, false rejects S_{FR} and false alarms S_{FA} and S_{FA}^{out} are provided.

Included in the **PyAFE** toolkit, an all-in-one command line evaluation tool is also available. It provides an easy to use, straightforward way of getting evaluation results. It gets as simple as typing:

```
$ python full_eval.py --groundtruth=[path_to_groundtruth_directory]
                      --submission=[path_to_detection_directory]
```

This tool actually browses all subdirectories of the *groundtruth* base directory looking for annotation files. For each of them, the corresponding detection file in the *detection* base directory is evaluated. Depending on the requested level of verbosity, it can output one single score for the whole set of files, one score per file, or even the detailed list of errors made for every single file. Another useful option allows to perform the evaluation using only a subset of audio targets, by providing the list of their identifiers.

The full documentation can be found on the **PyAFE** website, along with sample groundtruth and detection files.

4 Application of the framework for the Quaero project

The Quaero project includes a sub-project focused on *Audio Identification and Fingerprinting*. Starting in 2010, an audio fingerprinting evaluation campaign is organised every year during summer, throughout the existence of the Quaero project. The pilot evaluation took place in September 2010, for which a dedicated corpus was collected.

Annotated radio broadcast corpus. It consists in the recording of 5 weeks of the french radio station RTL, captured and saved on disk in 5 minutes chunks. Therefore, the total duration reaches 840 hours. Similarly to what was described in Section 3.1, the whole dataset was annotated by manually checking the output of an audio identification engine (with precision around 1 second). The whole corpus is divided into two parts: four weeks make the training set and the remaining week constitutes the test set. A set of $7,309$ one-minute-long target audio signatures was gathered to build the database.

Pilot evaluation. A few months before the submission deadline, participants were provided with the training set, the corresponding annotations and the target signatures. They were also provided with the **PyAFE** toolkit, knowing that this very tool would be used by the evaluation coordinator for the actual evaluation. The test set – obviously free of any annotation – was then distributed to participants and they submitted back the output (XML files) of their audio fingerprinting systems. Three participants submitted at least one run for the Quaero evaluation campaign 2010:

- **Participant 1** provided a system based on double-nested FFT, combined to a kNN search among the database codes, and a post-processing that correlates succeeding detection timestamps.
- **Participant 2** has developed a fingerprinting system based on the Shazam algorithm (Wang [2003]).
- **Participant 3** audio fingerprinting system implements some slights modification over the system described in Haitsma and Kalker [2002]. It is based on quantizing differences of energy measures from overlapped short-term power spectra.

After a necessary adjudication phase during which the test corpus annotations were corrected when necessary, the final results were publicised within the Quaero consortium. Table 2 reports the results of the various submitted runs, as provided by the **PyAFE** toolkit.

System	S_{ok}/N	$S_{FA,1}$ (S^{out})	R_1	$S_{FA,1.5}$ (S^{out})	$R_{1.5}$	$S_{FA,2}$ (S^{out})	R_2
Participant 1	445 / 459	0 (2)	96.5%	0 (2)	96.5%	0 (2)	96.5%
Participant 2	381 / 459	0 (0)	83.0%	0 (0)	83.0%	0 (0)	83.0%
Participant 3	442 / 459	0 (2)	95.9%	0 (2)	95.9%	0 (2)	95.9%

Table 2. Pilot Quaero evaluation results. The false alarm scores are indicated inside and outside the occurences (respectively before and between the parentheses).

5 Conclusion & Future work

Audio fingerprint is one of the main industrial challenges of the last years related to the diffusion of music. While many systems has been proposed to perform this task, no comparison between existing technologies has been performed because of the lack of unified evaluation frameworks. In this paper we described a proposal for the evaluation of audio fingerprint algorithms in the case of broadcasted music. This framework contains the definition of score metrics, their public implementation and a public test set corresponding to the use-case of broadcast monitoring of music. Because this test-set contains radio streams, it naturally allows representing several degradation types artificially created in previous evaluations. The whole framework is accessible at the following URL: `http://pyafe.niderb.fr`. As an example, we presented the use of this framework and the results obtained during the first Quaero audio fingerprint evaluation.

The current framework focuses on the punctual detection of music tracks ("when has this music track been broadcasted ?") in a corpus, given a short signature of each track. Further scenarios will include the detection of the exact boundaries of music track diffusion ("when did this broadcast radio or TV started playing the song and ended it") or boundaries within the music tracks themselves ("which part of the track has been played ?"). Further works will concentrate on extending the framework to the detection of advertisement and jingles in audio streams, as well as blind recurrent patterns detection in audio streams (therefore without previous knowledge of signatures).

In a further step, it could be worth defining distortion measures between the reference signature and the broadcasted audio. This could lead to an objective distortion measure for each corpus.

References

Eric Allamanche, Jürgen Herre, Oliver Hellmuth, Bernhard Fröba, Throsten Kastner, and Markus Cremer. Content-based identification of audio material using MPEG-7 low level description. In *Proc. International Symposium on Music Information Retrieval (ISMIR '01)*, Bloomington, Indiana, USA, 2001.

Carlo Belletini and Gianluca Mazzini. A framework for robust audio fingerprinting. *Journal of Communications*, 5:409–424, May 2010.

Michaël Betser, Patrice Collen, and Jean-Bernard Rault. Audio identification using sinusoidal modeling and application to jingle detection. In *Proc. International Symposium on Music Information Retrieval (ISMIR '07)*, Vienna, Austria, September 23-27 2007.

Christopher J.C. Burges, John C. Platt, and Soumya Jana. Extracting noise-robust features from audio data. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP '02)*, volume 1, pages 1021–1024, Orlando, Florida, USA, May 13-17 2002.

Christopher J.C. Burges, John C. Platt, and Soumya Jana. Distortion discriminant analysis for audio fingerprinting. *IEEE Transactions on Speech and Audio Processing*, 11:165–174, May 2003.

Rui Cai, Lie Lu, Hong-Jiang Zhang, and Lian-Hong Cai. Highlight sound effects detection in audio stream. In *Proc. IEEE International Conference on Multimedia and Expo (ICME '03)*, volume 3, pages 37–40, Baltimore, Maryland, USA, July 6-9 2003.

Pedro Cano, Eloi Batlle, Harald Mayer, and Helmut Neuschmied. Robust sound modeling for song detection in broadcast audio. In *Proc. 112th AES Convention*, pages 1–7, Munich, Germany, May 10-13 2002.

Chloé Clavel, T. Ehrette, and Gaël Richard. Events detection for an audio-based surveillance system. In *Proc. IEEE International Conference on Multimedia and Expo (ICME '05)*, pages 1306–1309, Amsterdam, The Netherlands, July 6-8 2005.

Michele Covell and Shumeet Baluja. Known-audio detection using waveprint: Spectrogram fingerprinting by wavelet hashing. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, pages 237–240, Honolulu, Hawaii, USA, April 15-20 2007.

Jaap Haitsma and Ton Kalker. A highly robust audio fingerprinting system. In *Proc. International Symposium on Music Information Retrieval (ISMIR '02)*, Paris, France, October 13-17 2002.

Jürgen Herre, Eric Allamanche, and Oliver Hellmuth. Robust matching of audio signals using spectral flatness features. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '01)*, pages 127–130, New Paltz, New York, USA, October 21-24 2001.

Dalwon Jang, Chang D. Yoo, Sunil Lee, Sungwoong Kim, and Ton Kalker. Pairwise boosted audio fingerprint. *IEEE Transactions on Information Forensics and Security*, 4:995–1004, December 2009.

Tang Jie, Liu Gang, and Guo Jun. Improved algorithms of music information retrieval based on audio fingerprint. In *Proc. 3rd International Symposium on Intelligent Information Technology Application Workshops (IITAW '09)*, November 21-22 2009.

Sungwoong Kim and Chang D. Yoo. Boosted binary audio fingerprint based on spectral subband moments. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, volume 1, pages 241–244, Honolulu, Hawaii, USA, April 15-20 2007.

Wei Li, Yaduo Liu, and Xiangyang Xue. Robust audio identification for MP3 popular music. In *Proc. 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 627–634, Geneva, Switzerland, July 19-23 2010.

Chien-Chang Lin, Shi-Huang Chen, Trieu-Kien Truong, and Yukon Chang. Audio classification and categorization based on wavelets and support vector machine. *IEEE Transactions on Speech and Audio Processing*, 13:644–651, September 2005.

Hui Lin, Zhijian Ou, and Xi Xiao. Generalized time-series active search with kullback-leibler distance for audio fingerprinting. *IEEE Signal Processing Letters*, 13:465–468, August 2006.

Yu Liu, Hwan Sik Yun, and Nam Soo Kim. Audio fingerprinting based on multiple hashing in DCT domain. *IEEE Signal Processing Letters*, 16:525–528, June 2009.

Mehryar Mohri, Pedro Moreno, and Eugene Weinstein. Efficient and robust music identification with weighted finite-state transducers. *IEEE Transactions on Audio, Speech and Language Processing*, 18: 197–207, January 2008.

Helmut Neuschmied, Harald Mayer, and Eloi Batlle. Identification of audio titles on the internet. In *Proc. International Conference on Web Delivering of Music (Wedelmusic '01)*, Florence, Italy, November 23-24 2001.

Geoffroy Peeters, Amaury La Burthe, and Xavier Rodet. Toward automatic music audio summary generation from signal analysis. In *Proc. International Symposium on Music Information Retrieval (ISMIR '02)*, Paris, France, October 13-17 2002.

Julien Pinquier and Régine André-Obrecht. Jingle detection and identification in audio documents. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, volume 4, pages 329–332, May 17-21 2004.

Mathieu Ramona and Geoffroy Peeters. Audio identification based on spectral modeling of bark-bands energy and synchronisation through onset detection. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP '11)*, pages 477–480, Prague, Czech Republic, May 22-27 2011.

Mathieu Ramona and Gaël Richard. Comparison of different strategies for a SVM-based audio segmentation. In *Proc. 17th European Signal Processing Conference (EUSIPCO '09)*, Glasgow, Scotland, August 24-28 2009.

Xavier Rodet, Laurent Worms, and Geoffroy Peeters. Brevet FT R&D/03376: Procédé de caractérisation d'un signal sonore - Patent 20050163325 Method for characterizing a sound signal, July 2003.

Jin S. Seo, Minho Jin, Sunil Lee, Dalwon Jang, Seunjae Lee, and Chang D. Yoo. Audio fingerprinting based on normalized spectral subband moments. *IEEE Signal Processing Letters*, 13:209–212, April 2006.

Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *Proc. 8th ACM International Workshop on Multimedia Information Retrieval (MIR '06)*, pages 321–330, Santa Barbara, California, USA, October 26-27 2006.

Gavin Smith, Hiroshi Murase, and Kunio Kashino. Quick audio retrieval using active search. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP '98)*, volume 6, pages

3777–3780, Seattle, Washington, USA, May 12-15 1998.

Avery Li-Chun Wang. An industrial-strength audio search algorithm. In *Proc. International Symposium on Music Information Retrieval (ISMIR '03)*, Washington, D.C., USA, October 26-30 2003.

Eugene Weinstein and Pedro Moreno. Music identification with weighted finite-state transducers. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, volume 2, pages 689–692, Honolulu, Hawaii, USA, April 15-20 2007.