AUDIO IDENTIFICATION USING SINUSOIDAL MODELING AND APPLICATION TO JINGLE DETECTION

Michaël Betser **Patrice Collen** **Jean-Bernard Rault**
France Télécom R&D
4 rue du clos courtel, 35510 Cesson-Sévigné, France
e-mail: firstname.lastname@orange-ftgroup.com

ABSTRACT

This article presents a new descriptor dedicated to Audio Identification (audioID), based on sinusoidal modeling. The core idea is an appropriate selection of the sinusoidal components of the signal to be detected. This new descriptor is robust against usual distortions found in audioID tasks. It has several advantages compared to classical subband-based descriptors including an increased robustness to additive noise, especially non-random noise such as additional speech, and a robust detection of short audio events. This descriptor is compared to a classical subband-based feature for a jingle detection task on broadcast radio. It is shown that the new introduced descriptor greatly improves the performance in terms of recall/precision.

1. INTRODUCTION

This last decade has seen a rapid increase in available multimedia content. The question of rapid and easy access to these data has become a strategic research subject, especially in the audio domain. Audio identification concerns numerous applications which can be gathered in three categories: research of information about a document (identification via cell-phone, CD tracks identification etc.), document detection for structuring purposes (jingle detection...) and document detection for broadcast control and copyright purposes. All media involving audio contents are concerned.

Although two families of audio identification system exist, namely audio fingerprinting and audio watermarking, the former is more popular, being more robust and non-intrusive. Fingerprinting systems involve the construction of a fingerprint for each document which should uniquely characterize the document. The fingerprint should also be robust to any alteration the document might suffer. Extensive lists of possible alterations can be found in many studies, and actually depend on the application [1] [2].

Non-random additive noises such as speech are a very frequent problem in the case of broadcast radio analysis. This problem is not usually addressed by researchers because most audio identification systems are aimed at musical pieces identification and musical database management. This type of modification can seriously alter most of the classically used fingerprints.

This article will present a new type of fingerprint, based on a sinusoidal parameter extraction, which can handle this kind of noise as well as the other usual distortions.

One other major issue, concerning audio identification constraints is computational efficiency. Most systems use hashing procedures, and precomputed look-up tables to speed up the process [2]. The purpose of this article is not computational performance comparison, but rather to demonstrate the robustness of this sinusoidal fingerprint, bearing in mind that hashing procedures can also be adapted to this parameter.

In section 2, the paper begins with a brief presentation of the fingerprinting systems, and a review of the different types of fingerprints used for audio identification. The proposed sinusoidal fingerprint extraction and comparison scheme will be presented in section 3 and 4. Application to jingle detection and experimental comparisons end the main part of the document in section 5 before a conclusion note.

2. FINGERPRINT SYSTEM

All fingerprint-based methods present the same classical analysis scheme [1], which is presented in **Figure 1**.

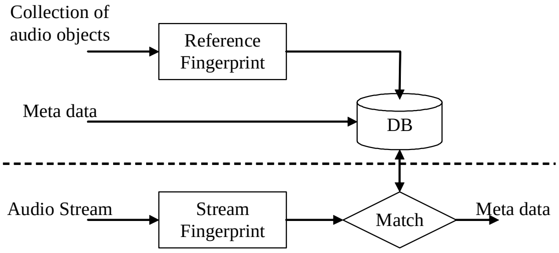

Figure 1: Fingerprint-based audio identification

A fingerprint-based identification system is composed of three distinctive parts: a fingerprint extraction module, a storage module ('DB' for database), a fingerprint comparison module ('match').

We have made a distinction between the reference fingerprint and the stream fingerprint computation modules. We will see in section 3 that computing a different fingerprint for the reference audio objects and the audio streams makes sense and is useful for sinusoidal fingerprinting in order to take superimposed sounds into account.

This article focuses on the fingerprint extraction module, which is described in detail in the next section, but also on the comparison module which directly depends on the nature of the fingerprint. Usually the comparison procedure is kept as simple as possible for computational purposes.

2.1. Classical feature extraction

The generation of the fingerprint is usually based on a short-time frequency analysis, using a windowed Fast Fourier transform (FFT). The fingerprint of an audio document is therefore composed of a collection of sub-fingerprints regularly spaced in time. In most systems, the FFT spectrum is divided into sub-bands, and a spectral characteristic is extracted from each sub-bands to form the sub-fingerprint. The spectral characteristic extracted can be the spectral magnitude [3] , the energy difference along the time and frequency axes [2], the spectral flatness measure [4], the modulation spectrum of the energy flux [5], the binary state of activation [6], the Mel cepstrum coefficients [7].

The method retained as reference for comparison is the one proposed by Haitsma et al. [2]. This method is one of the most utilized for comparison purposes since it exhibits very good performances on musical identification tasks.

2.2. Limitations of classical fingerprints

Perceptually, recognizing a sound object essentially relies upon the information carried by the object's predominant sinusoidal components. Even if there is no complete psycho-acoustic study on the subject, experiments seem to confirm this fact [6]. The main flaw of the descriptors described in the previous section is that they only partially take into account this fact. Another problem concerns the low energy portions of the audio signal which are often kept to compute the fingerprints and which make them more fragile.

The proposed solution is to analyze the reference signal, in order to extract the strongest sinusoidal components, which will be less prone to noise perturbation. Actually severe distortions on these components will result into such important changes in the perception of the sound object that the object can hardly be considered identical anymore.

3. SINUSOIDAL FINGERPRINT

The general procedure of the sinusoidal fingerprint creation is presented on **Figure 2**. It has four steps, the pre-selection and the compression steps being optional.

3.1. Preselection

The pre-selection step corresponds to any pre-processing done before the FFT, like band filtering. Here it is a low-pass filtering of the signal with a frequency cut at 4 kHz. It will render the signature less prone to band pass limitation, and concentrate the processing on the most informative part of the signal.

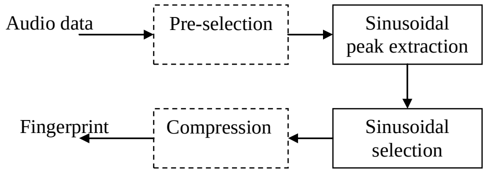

Figure 2: Fingerprint-based audio identification

3.2. Sinusoidal peak extraction

Sinusoidal modeling is based on the decomposition of audio signals into a sum of sinusoidal components plus a noise residual part. The sinusoidal components are modeled by a sinusoid with a set of parameters including amplitude, phase and frequency. Sinusoidal parameters extraction consists in estimating these parameters for each sinusoid present in the signal. For audio identification, the phase is not a discriminant parameter and therefore will not be retained.

Numerous approaches have been proposed, many of which being based on Fourier analysis. The Fourier-based estimation procedure has proven almost optimal, given that the sinusoids are well resolved by the Fourier transform and respect the underlying sinusoidal model [9]. These methods are also computationally effective, many of them being only slightly more complex compared to a FFT. The method retained for frequency estimation is one of the so-called Discrete Fourier Interpolator using phase, described in [10].

3.3. Sinusoidal peak selection

All the extracted sinusoidal peaks are not of equal interests: many of them have low amplitudes, which make them more vulnerable to noise perturbation, whilst others do not verify the sinusoidal model, which causes imprecision in the method of estimation. In fact, only the most predominant and stable peaks, relatively to the estimation method, are required to identify an audio document. Consequently, a selection procedure is needed both for computational purposes, as fewer peaks will require less processing, and robustness purposes.

The selection in itself can be different for the reference fingerprint and stream fingerprint. This has two advantages. First, the number of sinusoidal component in the reference may vary from frame to frame, and if a noise with strong sinusoidal peaks is added, the peaks to be detected might not be the strongest ones anymore. It is therefore interesting to keep more peaks in the stream fingerprint than in the reference. Secondly, contrary to the reference fingerprint, the stream fingerprint is usually computed on-the-fly. A less complex signature creation will therefore save precious time. A concrete example of peak selection will now be detailed.

The reference peak selection consists of three steps. First, all of the low energy peaks are removed using an adaptive threshold. Only the sinusoids whose amplitude is superior to a fraction of the power of the signal are kept. Secondly, to avoid the time-shifting problems, and to suppress the unstable peaks, the frequency of peaks has to lie within a tolerance of Tf Hz when the frame is shifted by H/2 and –H/2 samples, where H is the step size between two frames. Finally, the M most energetic peaks are kept. In order to have a reliable fingerprint, M should be superior to a hundred.

The stream peak selection consists in keeping the *Q* most energetic peaks per frame. *Q* should be greater than the maximum number of peaks kept within a frame, during the reference signature extraction step.

3.4. Compression

The compacting module consists in keeping only the frequency of each peak, coded on 16 bits. The corresponding precision, for a frequency interval of [0,4000], is 0.1 Hz. The amplitude is not kept, being more prone to noise perturbation. For a given frame, the final signature is a vector containing the frequencies of all the selected peaks.

4. FINGERPRINT COMPARISON AND DECISION

Figure 3 represents the general scheme of sinusoidal fingerprint comparison. The stream fingerprint is a block of T frames, referred to as Bs. This block is compared to all the possible blocks Bj,t of length T in all the reference object, where j is the index of the reference object and t the frame index in the reference object j.

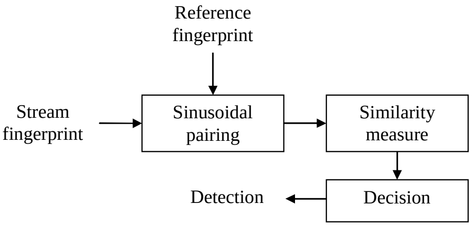

Figure 3: Fingerprint-based audio identification

The comparison of B_s and $B_{j,t}$ is done frame by frame. A frequency of one frame in $B_{j,t}$ is considered as paired (detected) with a frequency of the corresponding frame in B_s if they are equal, within a tolerance T_f. This tolerance is the same as the one used in the reference fingerprint creation, in section 3.

The chosen similarity measure is the number of reference frequencies correctly detected in the audio stream normalized by M_j, the number of peaks per second in the reference *j*. Dividing by M_j favors the parts of the reference which have a number of frequencies per frame superior to the mean, and are therefore more reliable. If all the frequencies are detected, and if the length of the reference audio object is equal to *T*, then the similarity is equal to 1. In the general case, when the similarity is close to one or above, the detection is considered as very reliable. The robustness on additive non-random noise is ensured by the fact that only correctly detected sinusoidal peaks will increase the similarity measure. Additional sinusoidal peaks in the stream fingerprint will have no impact.

The similarity measure can be considered as a function of the time t for each reference j. If a maximum is detected in this function of time, a decision is taken using two thresholds as in [3].

5. APPLICATION TO JINGLE DETECTION

5.1. Corpus

The system is applied to a jingle detection task for broadcast radio. The task consists in detecting 30 extracts of jingles from the French news radio France Info. The jingles to be detected have a length varying from 3 seconds to 10 seconds. Their length is not determining for the performance evaluation because the detection is realized for a fixed block size of 1 second.

The train corpus used for reference fingerprint creation is composed of one example of each jingle recorded in FM. The development corpus for parameter tuning is composed of 15 extracts of one minute of France Info recorded separately, each containing one jingle.

The test corpus is composed of 18 hours of France Info radio program recorded in AM. A total of 243 jingle occurrences are present in the corpus. Among them 33 corresponds to shorter versions of the jingles. In radio programs, these short jingles are used to announce new topics for example. They are usually fragments of their longer counterparts. In order to test noise robustness, this corpus has been altered using two other kinds of distortions: mp3 compression at 16 kb/s (AM+MP3) and speech addition with a 0 dB SNR (AM+SP). We have also added 48 hours from two other musical French radios, RFM and Skyrock for false alarm verification.

5.2. Parameters

In **Table 1**, the parameters used for both algorithms are summarized. `Sinusoidal' refers to the algorithm described in this article and `HKO' to the classical algorithm described in [2].

	Sinusoidal	HKO
Sampling frequency	8000	8000
Frame size	512	4096
Step size	128	96
Frame per block	62	84

Table 1: Parameter comparison

The two approaches have very different FFT parameters: the former approach uses long Fourier transforms, with an important overlap, whereas the sinusoidal method uses short Fourier transform with a small overlap. For both methods the parameters have been set up to respect an entrance block size of approximately 1 second.

The other parameters have been optimized on the development corpus. The tolerance Tf is connected to the precision of the frequency estimator used. If the sinusoidal model is respected, i.e. the amplitude and frequency is locally constant, then the maximum error on the frequency estimation will stay much lower than the Fourier precision, even for strong white noise perturbations [9]. Tf should be slightly higher than this maximum. For the parameters used in our experiments, the maximum error for a -10 dB white noise perturbation is approximately 2Hz [10], and Tf has been set to 3Hz. The HKO algorithm uses 33 frequency bands with an exponential repartition, and the maximum bit error rate has been set to 0.25, which is the same value as in [2].

5.3. Results

The algorithms are compared in terms of recall/precision. Nevertheless, as both algorithms present no false alarm, the precision has been omitted, being always equal to 100 %. Two different measures of the recall are used. The recall in terms of occurrences is the number of jingle detected divided by the number of jingle present in the corpus. The recall in terms of duration is the total block length correctly detected divided by the total length of the jingles in the corpus.

	AM	AM+MP3	AM+SP
Sinusoidal	97	95	83
HKO	89	85	67

Table 2: Occurrence recall comparison in percent

	AM	AM+MP3	AM+SP
Sinusoidal	79	68	53
HKO	60	57	34

Table 3: Duration recall comparison in percent

A significant difference between the occurrence recall and the duration recall appears. As every jingle is at least several seconds long, there is still a high probability to find at least one of the blocks corresponding to an occurrence of the jingle. To a lesser extent, another fact explaining the differences between the two measures comes from the block-based comparison scheme. The blocks on edges of the jingles might not be detected if the blocks do not contain enough jingle frames.

For an AM only perturbation, both algorithms perform well, with an advantage to the sinusoidal algorithm. The remaining errors, in the case of the occurrence measure, come from the short jingles. Some of them are too short to offer a reliable detection. Both algorithms are fairly robust to a strong mp3 compression, in terms of occurrence. As expected the sinusoidal algorithm performs particularly well on a speech additive perturbation compared to the HKO algorithm. The duration recall decreases significantly, especially in the case of the HKO algorithm.

6. CONCLUSION

In this article, a new type of fingerprint dedicated to audio identification and based on sinusoidal modeling has been presented. The advantages of this fingerprint compared to classical subband-based fingerprint are twofold. First a better modeling of the signal to be recognized, focused on the most informative part of the signal, allows to reliably recognizing segments of sounds as short as 1 second. Secondly, as the comparison itself is based only on frequency, this fingerprint presents an increased robustness to compression, to noise addition, even for strong non-random signals such as speech, and to subband filtering modifications (equalization).

In future work, a fast version of the comparison algorithm, based on the principle of hashing tables, will be investigated. Adaptations of the algorithm for time stretching deformations using dynamic programming will also be explored.

7. REFERENCES

[1] P. Cano et al., “Audio fingerprinting: concepts and applications,” *Studies in computational intelligence,* vol. 2, pp. 233–245, 2005.

[2] J. Haitsma, et al., “Robust audio hashing for content identification,” *International Workshop on Content-Based Multimedia Indexing,* Sep 2001.

[3] J. Pinquier and R. André-Obrecht, “Jingle detection and identification in audio document,” *Proc. of ICASSP,* 2004.

[4] O. Hellmuth, et al., “Advanced audio identification using mpeg-7 content description,” *AES 111th convention,* Sep 2001.

[5] J. Laroche, “Patent : Process for identifying audio content,” Number : WO88900, Nov 2001.

[6] D. Fragoulis, et al., “On the automated recognition of seriously distorted musical recordings,” *trans. on Sig. Proc.,* vol. 49, no. 9, pp. 898–908, 2001.

[7] L. Gomes, et al., “Audio watermarking and fingerprinting: for which applications?,” *J. of New Music Research,* vol. 32, no. 1, pp. 65–81, 2003.

[8] C. Burges, et al., “Distortion discriminant analysis for audio fingerprinting,” *IEEE trans. on Sp. and Audio Proc.,* vol. 11, no. 3, pp. 165–174, 2003.

[9] M. Betser, et al., “Review and discussion on STFT-based frequency estimation methods,” *AES 120th Convention,* 2006.

[10] M. Betser, et al., “Frequency estimation based on adjacent DFT bins,” *Proc. of EUSIPCO,* 2006.