A GENERAL FRAMEWORK FOR DICTIONARY BASED AUDIO FINGERPRINTING

Manuel Moussallam and Laurent Daudet

Institut Langevin, Univ. Paris 7 Diderot - ESPCI ParisTech - CNRS UMR 7587
first.last@espci.fr, 1 rue Jussieu 75005 Paris

ABSTRACT

Fingerprint-based Audio recognition system must address concurrent objectives. Indeed, fingerprints must be both robust to distortions and discriminative while their dimension must remain to allow fast comparison. This paper proposes to restate these objectives as a penalized sparse representation problem. On top of this dictionary-based approach, we propose a structured sparsity model in the form of a probabilistic distribution for the sparse support. A practical suboptimal greedy algorithm is then presented and evaluated on robustness and recognition tasks. We show that some existing methods can be seen as particular cases of this algorithm and that the general framework allows to reach other points of a Pareto-like continuum.

***Index Terms*—Sparse Representation, Audio Fingerprinting**

I. INTRODUCTION

I-A. Standard audio fingerprinting approaches

Audio objects recognition systems aim at the automatic retrieval of a signal y among a collection of known sound objects $\{y^{(i)}\}$. In practice, such collection can be very large and sounds are complicated objects to compare. For this retrieval to be effective, the search must be performed efficiently, for instance by comparing low-dimensional *proxies* of the objects, or fingerprints.

An audio fingerprint is a collection of signal-characteristic features that is somehow robust to distortions and can be efficiently compared to others. There are two main families of audio fingerprinting systems, the first one adopts a bag of features approach. A low-dimensional vector of features (eg. Chroma, MFCC, etc..) is used as the fingerprint. It has for instance been proposed by Haitsma [1] with binarized Chroma. A review of such methods can be found in [2] with more recent avatars being based on wavelet transforms [3] or finer frequency models [4], [5].

The second family of methods is similar in spirit to some feature extraction methods developed in image processing. It has first emerged with the work of Wang [6] and builds on the idea of selecting a subset of *keypoints* in a Time-Frequency (TF) representation, pairing them to form *landmarks* and using each of these *landmarks* as an index in a structured database (*e.g.* a hash-table or any fast indexing system). This approach is at the basis of the well known Shazam service [7], but also led to the works of Cotton and Ellis[8] and Fenet *et al* [9] among others. While in his seminal work [6], Wang selected *keypoints* as local maxima in a simple spectrogram, Cotton and Ellis [8] use a greedy algorithm on a multiscale Gabor dictionary and Fenet *et al* [9] propose a logarithmic transform instead of windowed Fourier.

All of these methods share a common formalism, that is conveniently exposed using a dictionary-based point of view. Given a dictionary Φ, one seeks a combination of k elements of Φ (labeled *atoms*) that can be efficiently used as *keypoints* in a fingerprinting system. State of the art methods, mainly propose different dictionaries (*e.g.* Gabor [6], [7], Union of Gabor[8], MDCT [10], Logarithmic [9]..) and selection algorithms (Local Peak Picking [6], [7], [9], Matching Pursuit [8], [10]).

I-B. A unifying framework

Quite naturally, one would hope to design a unifying framework for all these methods. Mostly what distinguish them is the stress that is imposed either on the robustness of the landmarks or their discriminative power. A robust landmark is one that remains unaltered by distortions such as additive noise, compression, time or pitch shifting etc. The discriminative power is harder to quantize, but will directly be linked to recognition performances. A landmark is discriminative when it is highly characteristic of an object fingerprint, that is, it is unlikely to appear in the fingerprint of an object that is fairly different.

Unfortunately, robustness and discriminative power seem to be concurrent objectives. Indeed, such discriminant information will be found in the high frequencies of audio signals, but these frequencies are the most easily altered by distortions. Additionally, for the search to be efficient, the number of keypoints and landmarks must be kept as low as possible.

The purpose of this work is not to propose yet another audio fingerprint system, but to generalize existing ones within a common framework that is theoretically motivated. With this in mind, we first propose a formulation of the fingerprint design problem as a multi-objective optimization of a dictionary-based processing system. Then we introduce a proxy for the discriminative power using information theoretic tools. Using a structured sparsity model for the keypoints (*e.g.* atoms of the dictionary) one can model the probability of selecting a keypoint and even the probability of their combinations which allows to use an entropy measure to characterize the quantity of information carried by a single landmark. We then propose a general greedy algorithm to build (suboptimal) solutions and show that some particular parameter sets correspond to the state of the art algorithms described above.

The rest of this paper is organized as follows: Section 2 exposes our proposal to write the fingerprint design problem as a multi-objective sparse representation one. In Section 3, we present a structured sparsity model using Boltzmann machines and propose a penalized greedy algorithm to build hybrid fingerprints. The behavior of this algorithm and its relation to state of the art approaches is demonstrated in Section 4 on robustness and recognition experiments.

II. DICTIONARY BASED AUDIO FINGERPRINTS

Let $\boldsymbol{y} \in \mathbb{E}^N$ be a N - dimensional discrete signal ($\boldsymbol{E} = \mathbb{R}$ or $\mathbb{C}$) and $\boldsymbol{\Phi} = \{\phi_i\}_{i=1..M}$ a dictionary of M *atoms* $\boldsymbol{\phi}_i$ of same dimension than $\boldsymbol{y}$, one speaks of a *representation* $\hat{y}$ of $\boldsymbol{y}$ in Φ as a linear combinations of the atoms, *i.e.* $\hat{y} = \sum_{i=1}^{M} \alpha_i \phi_i$ where the weights coefficients stacked in an $M-$dimensional vector $\boldsymbol{\alpha}$ now carry the information. The nature and quantity of information conveyed by each (or a combination of) α_i depend on how the dictionary is designed and what *a priori* knowledge on the signal is available.

In an audio fingerprint context, it is interesting to further decompose $\boldsymbol{\alpha}$ as the element-wise product $\boldsymbol{\alpha} = \boldsymbol{x} \odot \boldsymbol{s}$ where x is real or complex valued and $\boldsymbol{s}$ is called the support and restricted to binary values: $s_i = 1$ if atom i is selected as a keypoint and

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

zero otherwise. In the following, the terms *keypoints* and *atoms* are equivalent.

II-A. Formalizing fingerprint properties as constraints

In this formalism, limiting the number of keypoints can be straightforwardly transcribed as a sparsity constraint on s. The robustness property is harder to characterize since different types of distortions may occur. For the sake of clarity, let us consider only the case of additive white Gaussian noise. The best way to resist such distortion is to select atoms minimizing a reconstruction error. More generally, most types of robustness can be enforced by constraints of *descriptiveness* of the keypoints.

Expressing the discriminative power, however, is more challenging. This can be done by using information theoretic metrics in general and entropy in particular. Audio signals often carry more energy in their low than high-frequencies. Corresponding keypoints thus have a higher probability of being selected. Intuitively, they provide a less discriminant information on a signal than the least probable ones. If one is able to fully evaluate the probability distribution of the support then one would want to constrain its *entropy* to be the highest possible.

The problem of finding k keypoints that have maximum descriptive and discriminative potentials can thus be stated as:

$$\mathcal{P}_{\lambda,k} : \min_{s} \|y - \sum_{i=1}^{M} x_i.s_i.\phi_i\|_2 - \lambda H_\Phi(s) \text{ s.t. } \sum_{i=1}^{M} s_i = k \quad (1)$$

where $H_\Phi(s)$ is the entropy of the vector s given the dictionary and λ a penalty weight.

II-B. Probabilistic modeling of the sparse support

By definition, $H_\Phi(s)$ only exists when the probabilistic distribution of s is available. Experimentally, we are able to characterize this distribution quite efficiently. Let Φ be a time-frequency dictionary, and let us observe the solutions to $\mathcal{P}_{0,k}$, that is the sparse reconstruction problem without entropic constraint. Figure 1 displays the empirical distribution of the first 100 keypoints selected with an algorithm from the Matching Pursuit (MP [11]) family as in [8], [10]. The dictionary is a union of 7 MDCT scales replicated such as to form a highly over-complete shift-invariant dictionary of roughly 65 millions atoms. Atoms are uniformly selected in time while a strong bias on their frequency localization can be observed.

At en even deeper level, one can empirically observe the covariance matrix of the support. Figure 2 shows empirical co-occurrences of keypoints relative to their frequency and time index respectively. Both matrices have strong coefficients near the diagonals, it reveals the neighborhood correlations between keypoints close to each other in the time-frequency plane. The frequency matrix also exhibit strong subdiagonals that reflect the harmonic correlations.

This basically tells us that landmarks built on neighboring and harmonically related keypoints are less informative (*i.e.* discriminative) than others.

II-C. Relation to existing work

Problem (1) is, in general, NP-hard to solve. Many works in the literature can be understood as suboptimal methods to tackle this problem. For instance, the method at the basis of Shazam [7] and its avatars [9], never explicitly express the entropic constraint. However, their strategy is to enforce the selection of keypoints that are spread all over the time-frequency plane. Basically, this amounts to forbidding the construction of landmarks from neighboring keypoints which are the less discriminative. Overall, this local peak-picking strategy can be understood as an entropy-oriented one.

On the other hand, systems such as the one proposed in [8] put all the emphasis on the robustness to distortions. Their landmarks

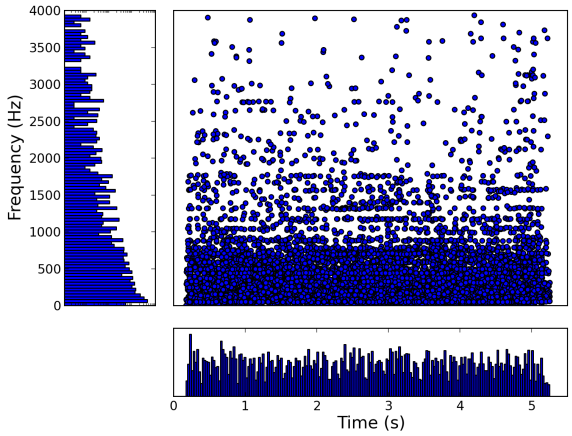

Fig. 1. Empirical Time-Frequency positions of the first 100 selected atoms (blue dots) and their marginal distributions observed on 600 audio segments of 5 seconds each, taken from the GTZAN[12] dataset. Signals are down-sampled to 8KHz. The marginal on frequency is presented in log scale.

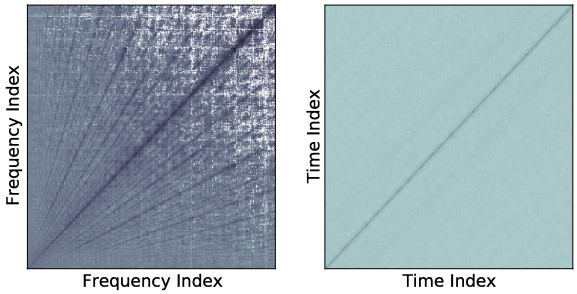

Fig. 2. Empirical Co-occurrence of Time Frequency atoms observed on the same 600 segments.. The empirical bias has been subtracted. Darker regions indicate higher co-occurrences. The strong diagonal components indicates neighborhood relationships both in time and frequency. Harmonic correlations can be observed in the frequency matrix.

are tailored for retrieval of highly distorted objects, but it might be at the expense of their discriminative power.

Finally, let us draw a parallel with some fingerprinting techniques such as the Distortion Discriminant Analysis exposed by Burges *et al* [13], to our knowledge, this would also fit in this framework but with a dictionary learning paradigm.

III. PROPOSED FRAMEWORK

III-A. Structured Sparsity model

Empirical evidence suggest the sparsity pattern of the support vector in time-frequency dictionaries is highly structured. We propose to use Boltzmann machines as a model for the distribution of s:

$$p(s) \propto \exp(b^T s + s^T W s) \quad (2)$$

This distribution has first been proposed in [14]. It models the interaction in a graph of connected nodes (keypoints in our case) using two parameters: a bias b and a connectivity matrix W. This model recently appeared in dictionary based processing setups. Dremeau *et al* [15] show that it generalizes many structured sparsity

models. Under this model, we can evaluate the probability of a state using the difference of energy for atom i:

$$\Delta E_i = \sum_j w_{ij} + b_i \tag{3}$$

Fixing the states of all other variables, the probability of atom i being turned on (*i.e.* keypoint i being selected) writes:

$$p(s_i = 1 | \{s_{j \neq i}\}) = \frac{1}{1 + \exp(-\Delta E_i)} \tag{4}$$

III-B. Reducing model complexity

The expressiveness of the Boltzmann machine is essentially captured by the W matrix which is of size $M \times M$ where M is the number of atoms in the dictionary. Clearly, for real scale data, the resulting model complexity will become prohibitive. Fortunately, the considered dictionaries are further structured. Assume each atom ϕ_i can be indexed by a unique triplet $(f_i, t_i, l_i) \in \mathcal{F} \times \mathcal{T} \times \mathcal{L}$ of its frequency and time centroids and length. A way to drastically reduce the complexity is to assume separability of the time and frequency centroid variables. Such hypothesis seems reasonable because a keypoint frequency localization is essentially linked to other keypoints frequencies and lengths, independently of their time position. Symmetrically, time localizations may be considered apart from the frequency localization.

In practice, this implies cutting many vertices in the Boltzmann machine graph, or equivalently putting many elements of W to zero. We have seen empirically that keypoints are uniformly located in time, we can thus drop this dependency:

$$b_i = b(f_i, t_i, l_i) = b(f_i, l_i) \tag{5}$$

Similarly, each element w_{ij} of the W matrix can be expressed as a product:

$$\begin{aligned} w_{ij} &= w\left[(f_i, t_i, l_i)(f_j, t_j, l_j)\right] \\ &= w^F\left[(f_i, l_i)(f_j, l_j)\right] w^T\left[(t_i, l_i)(t_j, l_j)\right] \\ &= w_{ij}^F w_{ij}^T \end{aligned}$$

where w_{ij}^F and w_{ij}^T are taken in two factoring matrices W_F and W_T. We have seen empirical estimators of such bias and W matrices in Figures 1 and 2.

III-C. Algorithm

Addressing problem (1) is a complicated issue. Indeed, even with $\lambda = 0$, it requires either a relaxation of the sparsity constraint or the use of suboptimal greedy algorithms such as MP. Given that the hard sparsity constraint is strict in this case, we have chosen to modify an MP algorithm by simply changing the atom selection rule.

Such algorithm makes a series of local decisions (*i.e.* keypoint selection), based only on the knowledge of the previous choices (*i.e.* which keypoints have already been selected). The residual signal r^n at iteration n is usually updated by subtracting from the original signal its projection on the subspace spanned by the selected atoms. At iteration n the decision boils down to solving:

$$\arg\max_{\phi_i \in \Phi} |\langle r^n, \phi_i \rangle| (1 + \lambda_H H(\phi_i | s_{n-1})) \tag{6}$$

where $H(\phi_i | s_{n-1})$ is the entropy of choosing atom ϕ_i knowing the support s_{n-1} and writes:

$$\begin{aligned} H(\phi_i | s_{n-1}) &= -p(\phi_i | s_{n-1}) \log\left[p(\phi_i | s_{n-1})\right] \\ &= \frac{\log\left[1 + \sum_{j \in \Gamma_{n-1}} w_{ij} + b_i\right]}{1 + \sum_{j \in \Gamma_{n-1}} w_{ij} + b_i} \end{aligned} \tag{7}$$

with Γ_{n-1} being the indices of the non zero elements of s_{n-1}, *i.e.* the keypoints selected so far. An advantage of this algorithm is that it can be quickly implemented using existing MP libraries such as PyMP[1]. Additionally, existing algorithms can be seen as particular cases.

IV. EXPERIMENTS

In the framework described above, many parameters need to be chosen. The dictionary Φ, the sparsity k of the representation (*i.e.* the number of keypoints), λ_H, the bias b and the W matrix. We will adopt the notation W03 and C10 by reference to [6] and [8] respectively. Note however that all results presented here are obtained with our own implementation of these methods. W03 corresponds to a local peak picking strategy with a monoscale Gabor dictionary. C10 is equivalent to our algorithm with λ_H being set to 0. We investigate hybrid strategies with a simple synthetic frequency bias b and a neighbor penalizing matrix W.

Experiments are run in a framework that is similar in nature to the one presented in [6], [9], [8]. Landmarks are binarized and stored as index keys in a Hashtable implemented using the open source Berkeley DataBase C Library[2]. Each key is a combination $(f_1, f_2, \Delta t)$ where f_1 and f_2 are the frequency centroids of the two keypoints and Δ_t the difference between their time centroids. Each key corresponds to a value that is a combination of the file index and the time of occurrence of the landmark in the file. In this work we are interested in comparing the keypoints and landmarks selection procedures. For fair comparison, the hashing and key formatting parameters are not optimized to any of the methods but fixed to common values.

In the following, a set of parameters will be identified by the triplet (λ_H, b, W). Whenever the bias (respectively W) is set to zero we will write $\lambda_H\,(0, W)$ (respectively $\lambda_H\,(b, 0)$). In this work we use a synthetic bias that is a simple decreasing exponential of the keypoints frequencies. W is decomposed in W_T and W_F that are zeros everywhere except near the diagonals. This particular setting corresponds to penalizing the selection of keypoints in the neighborhood of previously selected ones. For now we do not penalize harmonic relationships. Non zero coefficients in W_T and W_F are adapted to the desired sparsity level and corresponds to the same Time-Frequency widths as the ones used in W03 for local peak picking.

IV-A. Keypoints and Landmarks entropy

We expect keypoints selected with the entropic penalization to have a distribution somehow more "uniform" that those selected on purely energetic considerations. Indeed Figure 4 shows empirical distributions measured on the decomposition of 600 random 5 second length segments taken from the GTZAN[12] dataset. Keypoints and Landmarks selected with the W03 method are almost uniformly distributed, while the ones built by C10 exhibit a strong bias towards low frequencies. The hybrid approach allows to reach a new compromise.

An illustration of this behavior is provided in Figure 3. For 4 different settings, 100 landmarks are built and figured with black segments. Recall that C10 is similar to [8]. The algorithm selected atoms on a purely energetic basis in a union of 6 scales Gabor shift-invariant dictionary. A first hybrid approach (labeled $\lambda_H = 5\,(0, W)$) uses the same dictionary with $\lambda_H = 1$ the bias is set to zero and W as described above. The penalization led to the selection of a slightly different set of keypoints. A second hybrid approach (labeled $\lambda_H = 10\,(b, W)$) uses both W and the bias b to penalize the selection. This time the algorithm has selected a very different set of keypoints. Finally, the last case

[1]https://github.com/mmoussallam/PyMP

[2]http://www.oracle.com/us/products/database/berkeley-db/overview/index.html

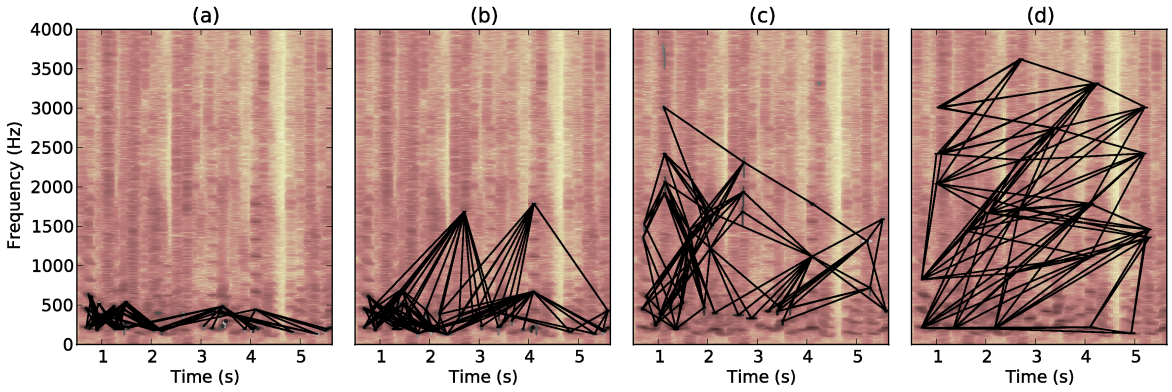

Fig. 3. Time-frequency *landmarks* built by the algorithm with varying parameters on a 5s audio excerpt of female speech. Each case has built 100 landmarks. (a): C10 ($\lambda_H = 0$) (b): $\lambda_H = 1\,(0, W)$ (c): $\lambda_H = 10\,(b, W)$(d): W03

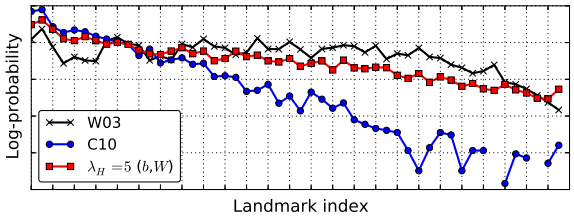

Fig. 4. Empirical distribution of landmarks. Flatter distribution has higher entropy and corresponds to cases where each landmark is more discriminant. Landmarks are indexed by increasing frequencies of the first keypoint.

(W03) use a monoscale Gabor dictionary with 25% overlap and a selection procedure equivalent to [6]. This is similar to what could be obtained with no bias, the same W and $\lambda_H = +\infty$.

IV-B. Robustness and Recognition performances

The primary objective of a fingerprint system is its recognition performance. The main parameter affecting the quality of the results is the sparsity, or equivalently the number of landmarks per seconds on which to base the decision. For each query, the system returns a best candidate file in the database and an estimated time of occurrence. The score is simply the ratio of the number of correctly retrieved segments over the number of queries. To assess for the robustness of the fingerprints, we measure the Proportion of Identical Landmarks (PIL) that remain unaffected by an additive white Gaussian noise.

Using different parameters a learning an a testing phase are run. During the learning phase, a database of fingerprints is built out of the 1000 files of the GTZAN dataset[12]. Each file is sliced in 5 seconds-long segments with a 50% overlap. On each segment the number of allowed keypoints is fixed to the sparsity level k. During the test phase, 2500 randomly chosen 5 seconds-long segments are used for queries. The testing segments thus have a very low probability of being aligned with the learning ones. The compromise between the two concurrent objectives is illustrated on Figure 5. C10 is the most robust method but performs poorly on the recognition task at low levels of sparsity. On the opposite, W03 reaches very good recognition scores at low sparsities, but is also the most affected by the additive noise. Between them, the hybrid approaches allows to reach different compromises.

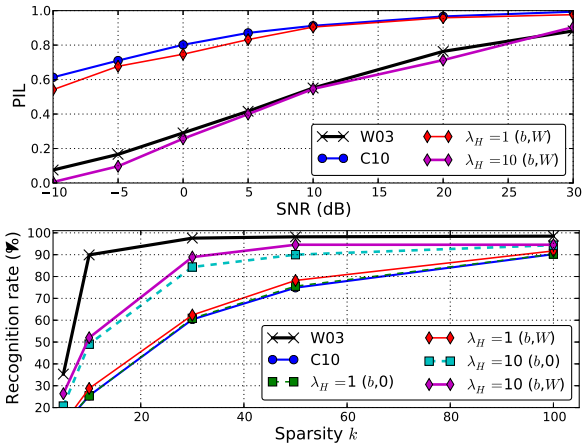

Fig. 5. Top: Robustness results with synthetic W and varying λ_H averaged over 5 trials of random Gaussian noise on each of 600 random segments of 5 seconds taken in the GTZAN dataset. Bottom: Recognition performances for isolated 5sec audio excerpts from the complete dataset with various settings, function of the number of keypoints k.

V. DISCUSSION

The proposed framework is flexible and there are many parameters one can modify. The role played by each of this parameters need to be further investigated. Let us stress that the expressive power of the model is quite high and one could use it to introduce more prior knowledge on the data. Taking harmonic correlations into account would be a natural next step. More generally, specific relationships could be learn on a variety of sound classes, such as speech, instrumental or environmental.

In this work, we avoided the issue of inferring the Boltzmann machine parameters by using empirical estimators. It is arguably not satisfying, but allowed us to conduct these proof of concept experiments. Moreover, the suboptimal strategy proposed here to address Problem (1) has also been chosen for its simplicity and serve as comparison basis. Future work will investigate smarter optimization scheme, such as Bayesian versions of MP (*e.g.* as in [15] with Boltzmann machines) or convex relaxations methods.

VI. REFERENCES

[1] J. Haitsma, T. Kalker, and J. Oostveen, "Robust Audio Hashing for Content Identification," in *International Workshop on Content-Based Multimedia Indexing*, 2001.

[2] P. Cano, E. Battle, T. Kalker, and J. Haitsma, "A Review of Algorithms for Audio Fingerprinting," in *IEEE Workshop on Multimedia Signal Processing*, pp. 169–173, 2002.

[3] S. Baluja, "Audio Fingerprinting: Combining Computer Vision and Data Stream Processing," in *IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 213–216, 2007.

[4] M. Betser, P. Collen, and J.-B. Rault, "Audio Identification using Sinusoidal Modeling and Application to Jingle Detection," in *International Society for Music Information Retrieval Conference*, (Vienna, Austria), Sept. 2007.

[5] E. Dupraz and G. Richard, "Robust Frequency-Based Audio Fingerprinting," in *IEEE International Conference on Acoustics Speech and Signal Processing*, (Dallas,USA), pp. 2091–2094, Mar. 2010.

[6] A. Wang, "An Industrial-strength Audio Search Algorithm," in *International Society for Music Information Retrieval Conference*, pp. 7–13, 2003.

[7] A. Wang, "The Shazam Music Recognition Service," *Communications of the ACM*, vol. 49, Aug. 2006.

[8] C. V. Cotton and D. P. W. Ellis, "Audio fingerprinting to identify multiple videos of an event," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2386–2389, 2010.

[9] S. Fenet, G. Richard, and Y. Grenier, "A Scalable Audio Fingerprint Method with Robustness to Pitch-Shifting," in *International Society for Music Information Retrieval Conference*, pp. 121–126, 2011.

[10] S. Fenet, M. Moussallam, Y. Grenier, L. Daudet, and G. Richard, "A Framework for Fingerprint-Based detection of Repeating Objects in Multimedia Streams," in *European Signal Processing Conference*, pp. 1464–1468, 2012.

[11] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.

[12] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[13] C. Burges, J. Platt, and S. Jana, "Distortion discriminant analysis for audio fingerprinting," *IEEE Transactions on Audio, Speech and Language Processing*, vol. XX, pp. 1–10, 2003.

[14] G. Hinton and T. Sejnowski, "Learning and relearning in Boltzmann machines," *MIT Press, Cambridge, Mass*, 1986.

[15] A. Dremeau, C. Herzet, and L. Daudet, "Boltzmann Machine and Mean-Field Approximation for Structured Sparse Decompositions," *IEEE Transactions on Signal Processing*, vol. 60, pp. 3425–3438, July 2012.