A GRANULAR APPROACH FOR THE ANALYSIS OF MONOPHONIC AUDIO SIGNALS

Lorcàn M. DONAGH, Frédéric BIMBOT, Rémi GRIBONVAL

IRISA (CNRS & INRIA) / METISS
Campus Universitaire de Beaulieu 35042 Rennes - France
{lmcdonag, frederic.bimbot,remi.gribonval}@irisa.fr

ABSTRACT

This paper describes a method for analyzing audio signals with an adaptive "parametric dictionary". We use sliding frames to extract elementary signals or *grains* from the analysis signal. We search for similarities amongst the collected grains to form *classes*, which we then use to derive a signal model for each class. These signal models or *prototypes*, are used to decompose the audio signal and compute analysis parameters for each grain. As a preliminary evaluation, we tested the method with real-life, monophonic and monaural recordings and obtained encouraging results.

1. INTRODUCTION

In [1], Gabor introduces the concept of *atoms* , *i.e.* signals localized in both the time and frequency domain. He states that any musical signal could be described as a superimposition of a large number of such atoms. Techniques such as Short Time Fourier Transform and Wavelet Analysis [2] rely on this assumption. Local cosine functions, including chirps and other variants, are common atom types for audio signal analysis. Noise + transient models can be used to account for non-pitched components also found in audio signals.

Drawing upon Gabor's idea and Wavetable / Granular synthesis techniques [3], we present a scheme for representing musical signals with a dictionary of parametric "wave" (*i.e.* with no simple analytical expression) atoms in order to obtain a sparse decomposition. We give the name *grain* to such atoms.

The main idea is to find and make use of both short- and long-term redundancies in the analyzed audio signal. *Short-term redundancy* is present in pitched or stationary parts. For example, a segment associated with a single note of a pitched instrument exhibits some form of periodicity. We could extract one cycle of the waveform which we would repeat to reconstruct this part of the signal, using PSOLA [4]. *Long-term redundancy* we associate with one note often being repeated several times at different places of the song. As there is no hope to find two exactly identical grains, we must use a measure of "similarity" between grains, introduced in Sec.2, and a model of the variability to describe similar grains, which we introduce in Sec.3. We shall discuss a practical analysis algorithm in Sec.4, while Sec.5 is devoted to the presentation of our experimental results.

2. FINDING REDUNDANCY

We split the main signal into short-time frames. This provides us with a number of *grain* signals localized in time, which we can compare to one-another. This transformation is a means to search for redundancy inside the main signal.

2.1. Obtaining the grains from the signal

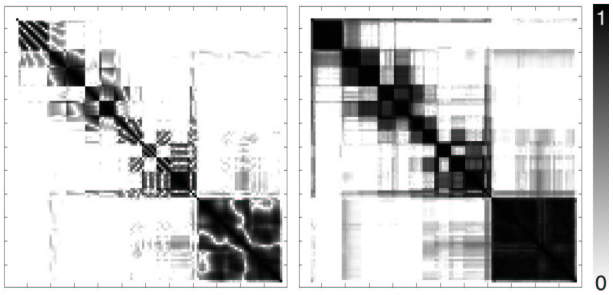
We define n_g "grain" signals g_i of duration d_g (typically approximately 20 milliseconds) by applying n_g Hanning windows W overlapping by 50%:

$$g_i[t] = W\,[t] \cdot s\,[t + \Delta_i]$$

Δ_i is the time-shift of the beginning of the i-th analysis window. Thanks to the properties of Hanning windows, we can write:

$$s[t] = \sum_{i=1}^{n_g} g_i\,[t - \Delta_i] \qquad (1)$$

Eq. 1 yields an exact, albeit trivial decomposition of s into shifts of vectors taken from a dictionary of n_g vectors. We aim to obtain a *sparser* decomposition of the signal, that is to decompose the signal with a dictionary of$K \ll n_g$ vectors. Therefore, our goal is now to find and model some form of redundancy amongst the grains. We evaluate similarity between grains by means of a similarity measure Γ, which is chosen to be invariant to a certain number of transformations of the input vectors. We then assign grains to classes, in which all grains are similar to each-other, with respect to Γ.



(a) Γ_{SP} (b) Γ_{LP}

Fig. 1. $\Gamma\left(g_i, g_j\right)$ for $n_g = 118$, $d_g = 32$ ms

2.2. A similarity measure

The similarity measure can be seen as a measure of the distance between two grains. We require Γ to be equal to one for a pair of identical grains. Below we give two examples of similarity measures and state the corresponding invariance properties.

$$\Gamma_{SP}\left(g, g'\right) = \frac{\langle g | g' \rangle}{\|g\|_2 \|g'\|_2} \quad (2)$$

$$\Gamma_{\Psi}\left(g, g'\right) = \langle |g(f)|, |g'(f)| \rangle \quad (3)$$

Γ_{SP} (Eq.2) is the scalar product of the normalized grains, it is invariant with respect to the amplitude of the grains. Γ_{Ψ} (Eq.3) is the scalar product of the normalized amplitude spectra of the grains. Γ_{Ψ} is invariant with respect to the phase spectrum of the grains. Fig.1(a) shows an example computed on all the pairs $\left(g_i, g_j\right)_{i,j \in \{1..n_g\}}$, of grains extracted from a clarinette signal[1][5]. Note the symmetric block structure of the matrix $\Gamma_{SP}\left(g_i, g_j\right)$. Dark squares mostly correspond to stationary parts of the signal (*i.e.* notes of the melody). But there are some oscillations of the similarity inside those blocks, a problem which we describe below.

2.3. Overcoming frame-related problems

One disadvantage of frame-based methods is that the bounds of the sonic events we want to identify may not be synchronized with those of the frames. This limitation might explain why we observe oscillations on Fig.1(a). Increasing the overlap between adjacent frames makes little sense, as this would substantially increase the number of required computations and eventually we would be comparing signals only differing by two sample values. Our proposed solution (Eq.4) is based on an extrapolation of the signal segment $s\left[\Delta, \Delta + d_g\right]$ using bi-directional Linear Prediction [6]. When the extrapolated segment is windowed with $W\left[t + n\right]$, it results in the extrapolated grain $\overleftrightarrow{g}_n$. We compute the similarity measure with all extrapolated grains $\overleftrightarrow{g}_n$ and retain the value of the optimal time-shift n_{opt} that leads to maximum similarity:

$$\Gamma_{LP}\left(g, g'\right) = \max_n \left[\Gamma_{SP}\left(\overleftrightarrow{g}_n, g'\right)\right] \quad (4)$$

Fig.1(b) shows the values of Γ_{LP} for the clarinet signal used previously. The oscillations are heavily reduced as expected.

3. A GRANULAR MODEL

Using the similarity measure $\Gamma\left(g_i, g_j\right)$ that we have presented, we can cluster the grains g_i into K classes C^k (see Sec. 4). In essence, our approach relies on the assumption that all grains inside one class of similarity are approximate versions of a unique root grain γ^k, the *prototype* of all these versions. The following section presents a mathematical formalization of this idea.

3.1. General Granular Model

We define the following model of a grain:

$$g_i = F\left(\gamma^k, \theta_i\right) + e_i \quad (5)$$

- γ^k is the *prototype* of the k^{th} class, chosen from C^k;
- F is a function verifying: $\Gamma\left(F\left(g, \theta\right), g\right) = 1, \forall\left(g, \theta\right)$;
- θ is the parameter vector of F;
- $\widetilde{g}_i = F\left(\gamma^k, \theta_i\right)$ is an approximation of g_i ;
- $e_i = g_i - \widetilde{g}_i$ is the approximation-error for grain g_i.

F should be set by the user to reflect some invariance properties of Γ .We will now discuss some methods to compute the prototypes γ_k and the parameters θ_i.

3.2. Signal decomposition

Eq.5 shows that every grain shall be represented with only a single prototype, which is moreover a grain chosen inside the corresponding class. This is for the sake of simplicity, and we are aware that this choice is not suitable when dealing with polyphonic signals *(i.e.* several sounds overlapping in time, prominent echo or reverberation). The sum of all the approximate grains $\widetilde{g}_i$ gives an approximate reconstruction $\widetilde{s} = \sum_{i=1}^{n_g} \widetilde{g}_i$ of the main signal s. Let us define

[1] The soundfile is available at *http://www.irisa.fr/metiss/lmcdonag*

$\Theta = \{\theta_i\}_{i=1..n_g}$ and $P = \{\gamma^k\}_{k=1..K}$ .We search to minimize the norm of the global reconstruction-error $\|s - \widetilde{s}\|$, which is equivalent to solving the optimization problem:

$$\min_{P,\Theta} \left\| s - \sum_{i=1}^{n_g} F\left(\gamma^k, \theta_i\right) \right\|$$

Unfortunately, solving this problem does not guarantee a sparse decomposition. For example, Eq.1 is a solution but is not sparse at all ($K = n_g$). We must introduce some more constraints, explicitely by arbitrarily fixing the number of classes K, or not. As the global minimization over all possible combinations of P and Θ is computationally prohibitive, we use algorithms where the minimization is done partially, in two successive stages.

4. LEARNING OF THE PROTOTYPES

The learning algorithm performs a classification of the grains according to the values of the similarity measure and finds a set of prototypes for the classes. Finally the signal is reconstructed approximately using the prototypes and optimized parameter values. Note that grains with norm below the noise floor are removed prior to classification. The performance of the algorithm and its variants will be assessed with their corresponding *Signal-to-Noise Ratios* (SNR).

4.1. EM Classification procedure

Given a grain g and a measure Γ, the closest neighbour $\overline{\gamma}^k(g)$ of g amongst a collection of prototypes P is defined as:

$$\overline{\gamma}^k(g) = \arg\max_{\gamma \in P} \left[\max_{\theta} \Gamma\left(g, F\left(\gamma, \theta\right)\right) \right]$$

Algorithm 1 EM-like Classification [7]

Randomly choose $\{i_1, .., i_K\}$ different indexes. Initialize the prototypes P with the set $\{g_{i_1}, .., g_{i_K}\}$. Iterate steps 1 & 2 over i.

1. For every g, find its closest neighbor $\overline{\gamma}^k(g)$ amongst P. Assign g to class C^k.
2. For each C^k, find g_0 having greatest average similarity with all $g \in C^k$. Replace γ^k with g_0.

Remark: the number of classes K must be set by the user before runtime, results are dependent upon initialization.

4.2. A direct procedure: Row Scanning

A property of the similarity measures is a form of *pseudo-transitivity*:

$$\left|\Gamma\left(g, g'\right)\right|, \left|\Gamma\left(g', g''\right)\right| \geq \varepsilon \rightarrow \left|\Gamma\left(g, g''\right)\right| \geq \alpha \cdot \varepsilon \qquad (6)$$

We can only give a theoretical lower bound for α, which is $\left|\Gamma\left(g, g''\right)\varepsilon^{-1}\right|$. In practice though, we observed that α is often close to 1.This means that when one finds three grains, two pairs of which have a high Γ value, the remaining pair roughly have the same (high) Γ. This observation lead us to write and experiment another classification algorithm (Alg.2), which scans the rows of $\Gamma\left(g_i, g_j\right)$ to form classes amongst which all grain verify Eq.6. Two threshold values α_0 and ε_0 must be provided before execution.

Algorithm 2 Row Scanning

Set $G = \{g_i\}_{i=1..n_g}$. Until $G = \emptyset$, do:

1. Pick a g^p from G. Remove it from G. Create a new class C^p and initialize it with g^p.
2. Scan all $g \in G$. If $\Gamma(g, g^p) \geq \varepsilon$, put g into C^p.
3. For every $g \in C^p$, check that every other $g' \in C^p$ satisfies $\Gamma(g, g') \geq \alpha \cdot \varepsilon$, otherwise remove g from C^p.
4. Substract C_p from G - Go to step 1.

Remarks the number of classes is computed by the algorithm, only α and ε need to be chosen in advance by the user. Judging by our tests, the value of ε does not seem critical provided $0.5 \leq \varepsilon < 1$. The degree of *homogeneity* inside classes is controlled by α. A high α imposes low dispersion of Γ-values inside one class, ultimately leading to $K \approx n_g$ when $\alpha = \varepsilon = 1$. An iterative procedure may also be used to automatically adjust α and ε for attaining a specified reconstruction-error value. Although we have not fully tested this, we believe the low complexity of the algorithm permits repeated execution in reasonable time.

5. EXPERIMENTS

Tests were performed on recorded and synthetized monophonic audio signals, using Matlab software. We used the *Signal-to-Noise Ratio* to measure and compare the performance of algorithms 1 and 2:

$$\mathrm{SNR}_{dB} = -20 \log_{10} \frac{\|s - \widetilde{s}\|}{\|s\|}$$

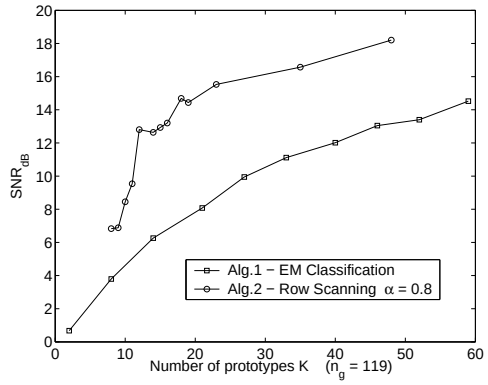


Fig. 2. SNR *vs.* dictionary size

Table 1. Dependency of K to ε

ε	.6	.7	.8	.9	.93	.95	.98	.999..
K	10	11	15	18	23	35	48	n_g

The following results were obtained with the aforementioned clarinet signal, which consists of 11 notes played successively. Nine grains having norms $\|g_i\| < 10^{-3}$ were removed prior to classification. Fig.2 summarizes the SNR figures obtained with both algorithms and various dictionary sizes K. For Alg.1, SNR is averaged over 10 trials to minimize the influence of initialization. With Alg.2, the number of prototypes K is dependent on ε (*cf.* Tab.1). Overall, the SNR is increasing with K, but with Alg.2 it may decrease locally, which happens with this particular signal when K is equal to the number of different notes in the melody.

5.1. Segmentation based on musical content

The classification process provides us with data readily useable for segmentation purposes. The proposed method may be used for analyzing musical signals and derive a representation similar to 'piano-roll' editors commonly found in MIDI sequencer software, as can be seen on Fig. 3. The figure displays the assignment of each grain g_i to a class $C^{k(i)}$, obtained using Alg.2 with $\varepsilon = 0.7$ and $\alpha = 0.8$, with i on the horizontal axis and k on the vertical axis. The class $k = 0$ represents the grains with norm below the noise-floor which were removed prior to running the algorithm. Time-locations of the prototypes are depicted with grey vertical lines. The analyzed signal is shown at the bottom.

6. CONCLUSIONS

Although the method has not been tested extensively with a large corpus of audio signals, the results we have obtained with the method encourage us to develop it and test it further.We shall conclude by listing the work in progress and the planned future developments:

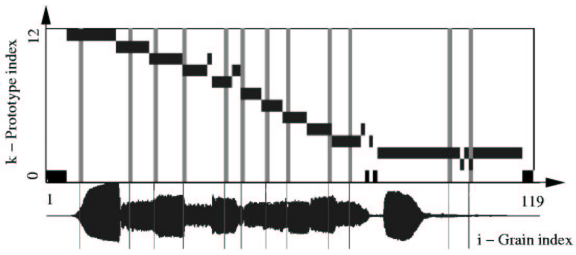


Fig. 3. Grain affectations to classes

- Currently the reconstruction-error is minimized on a per-grain basis, which is not optimal because of the overlap between adjacent grains. We are investigating ways to remedy this in order to achieve a better reconstruction. The prototype model and learning must also be improved.
- An extension of the approach to signals with *polyphonic* and *multi-timbral* content is being considered.
- An on-line version of the algorithm is in development, with streaming-media applications in mind.
- Use mp3 compression as a pre-analysis tool. Working directly on mp3-quantized DCT frames would be a simple and efficient way to implement some form of psycho-acoustic invariance criterion in the similarity measure.
- Investigate a musical granular re-synthesis method.

7. REFERENCES

[1] D. Gabor, "Theory of Communication," *J. IEEE*, 1946.

[2] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1998.

[3] C. Roads, *The Computer Music Tutorial*, MIT Press.

[4] E. Moulines and F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis using Diphones," *Speech Comm.*, 1990.

[5] Pierre Boulez, "*Dialogue de l'ombre double*," in "Pierre Boulez". Erato Disques, CD 2292-45648-2.

[6] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978.

[7] A. Dempster, N. Laird, and D.Rubin, "Maximum Likelihood from incomplete data via the EM algorithm," *J. Royal Star. Soc.*, 1977.