THE DESAM TOOLBOX: SPECTRAL ANALYSIS OF MUSICAL AUDIO

M. Lagrange[1], *R. Badeau*[2], *B. David*[2], *N. Bertin*[3], *J. Echeveste*[2], *O. Derrien*[4], *S. Marchand*[5], *L. Daudet*[6*]

Ircam[1]
CNRS STMS
Paris, France

Institut Telecom[2]
Telecom ParisTech
CNRS LTCI
Paris, France

METISS Project[3]
IRISA-INRIA
Rennes, France

Laboratoire de Mécanique[4]
et d'Acoustique (LMA)
CNRS - UPR 7051
Marseille, France

LaBRI CNRS[5]
University of
Bordeaux 1
Talence, France

Université Paris Diderot[6]
Institut Langevin
CNRS UMR 7587
Paris, France

mathieu.lagrange@ircam.fr

ABSTRACT

In this paper is presented the DESAM Toolbox, a set of Matlab functions dedicated to the estimation of widely used spectral models for music signals. Although those models can be used in Music Information Retrieval (MIR) tasks, the core functions of the toolbox do not focus on any specific application. It is rather aimed at providing a range of state-of-the-art signal processing tools that decompose music files according to different signal models, giving rise to different "mid-level" representations.

After motivating the need for such a toolbox, this paper offers an overview of the overall organization of the toolbox, and describes all available functionalities.

1. INTRODUCTION

Audio signal processing has made tremendous progress in the last decade, on both quantitative and qualitative levels. It now extends way beyond "traditional" audio engineering community ("hi-fi" hardware and software design, audio effects, audio coding, ...), sharing strong interaction with neighboring fields such as music perception and psychoacoustics, Music Information Retrieval (MIR), auditory displays, tools for composition and home studios, human-machine interfaces, ... to name but a few.

Arguably, one of the most challenging tasks, shared by most of these communities, is to decompose a complex, multi-instrumental, music signal into meaningful entities, or "objects", that not only convey some information from the signal processing point of view, but are also carry intrinsic meaning from a musicological or perceptual prospective (as opposed to *e.g.* time-frequency "atoms" in the analysis/synthesis framework, or signal "features" in the frame-based analysis). This type of processing has sometimes been called "mid-level" representations in the literature, and is also related to CASA framework [1]. Here, "objects" or "sound elements" can be notes, or structural elements of a note, such as harmonic tracks of partials. From these information, one could possibly infer what notes are being played (*i.e.* the score), but also what are the instruments, what type of playing technique is being used or what type of recording techniques have been used.

Obviously, no single technique can perform such a task perfectly on any type of real-life, polyphonic music. But what does "perfect" mean in this context? Obviously, this task will remain an ill-posed problem unless a specific signal model is defined. The goal of the DESAM Toolbox (that draws its name from the collaborative project: Décomposition en Eléments Sonores et Applications Musicales, funded by the French ANR), is to provide a range of state-of-the-art signal processing tools that decompose music files according to different signal models, giving rise to different "mid-level" representations, in other words a set of coefficients of a parametric model. As the main target of this toolbox is a widespread use for rapid prototyping of algorithms, especially in the academic community, as well as an educational purpose, it is composed of a set of MATLAB functions, that can be used for academic research under the only restrictions of a GPL license. The first version is available online [1] and will be maintained and upgraded according to user's feedback.

After motivating the need for such a toolbox in Section 2, the collaborative project that led to the set of tools presented in this toolbox is described in Section 3. The overall organization of the toolbox is then introduced in Section 4. The core tools, respectively based on sinusoidal models and more general spectral models are described in Sections 5 and 6, and the application to automatic music transcription is finally presented in Section 7.

2. RELATION TO OTHER MATLAB TOOLBOXES

Yet, why another audio Matlab toolbox? At time of writing, most widely-used Matlab toolboxes have two targets:

- either related to perception such as the Auditory Toolbox [2] and the Computer Audition Toolbox (CATbox) [3], in which case the main goal is to make some perceptually-relevant pre-processing for further analysis tasks;
- or related to Music Information Retrieval tasks (MIR Toolbox [4], MA Toolbox [5]), in which case the signal parameters are relatively simple, most of the time extracted on a frame-by-frame basis.

We believe there is still a widespread need for tools that implement recent state-of-the-art analysis methods, able not only to analyze audio signals according to various audio signal models - and can hence be used as input parameters for MIR systems - , but also to resynthesize the music from the extracted parameters - intended for audio coding in the engineering point of view, or sparse coding in the neural information processing community. These newer techniques, such as high-resolution methods, or NMF, have already proven their usefulness for a wide range of real-life problems. It is now time to deliver them to the whole community in

[*] The research leading to this paper was supported by the French GIP ANR under contract ANR-06-JCJC-0027-01, Décompositions en Éléments Sonores et Applications Musicales - DESAM. Mathieu Lagrange and Nancy Bertin have contributed to the work presented in this paper when they were working at CNRS LTCI. Olivier Derrien and Laurent Daudet have contributed to this work when they were working at STIC, Toulon and LAM, Paris, respectively.

[1] http://www.tsi.telecom-paristech.fr/aao/2010/03/29/desam-toolbox

order for them to become "classic" tools, in the same way as the Short Time Fourier Transform or the Additive Synthesis have been for decades. The users will notice that the core algorithm of each of these powerful tools is always strikingly simple and compact, we believe that they provide foundations for a multitude of future improvements.

3. DESAM PROJECT

The DESAM project [2] was a fundamental research project involving four French laboratories:

- CNRS LTCI (*Laboratoire Traitement et Communication de l'Information*), Institut TŐlŐcom - TŐlŐcom Paristech, Paris, France;
- LAM (*Lutheries - Acoustique - Musique*) team, Institut Jean Le Rond d'Alembert, UPMC Univ. Paris 6;
- LaBRI (*Laboratoire Bordelais de Recherche en Informatique*), Bordeaux 1 University;
- STIC (*Laboratoire Sciences et Technologies de l'Information et de la Communication*), Var and Toulon University.

Headed by the LTCI, the project started in November, 2006 and it ended in February, 2010. It was divided in two parts. The first one was devoted to the theoretical and experimental study of parametric and non-parametric techniques for decomposing audio signals into sound elements. The second part focused on some musical applications of these decompositions.

3.1. Decompositions into sound elements

Their pitch and their timbre, specific to the instrument, characterize musical notes. When these notes are well modeled by a mixture of sinusoids, the estimation of frequencies, amplitudes, and their time-variations, are useful to analyze the pitch and timbre of the sound. In this project, we have developed innovative high-resolution (HR) methods for time-frequency analysis, in order to estimate the fine time variations of these two parameters [6, 7, 8]. The modeling of non-stationary sinusoids was further addressed in [9].

Besides, since a musical piece is composed of multiple notes played at different times, it is naturally described as a combination of sound elements (which can be either isolated notes, combinations of notes, or parts of notes). Such a representation is called *sparse*, since a limited number of sound elements permits to describe the whole musical content. Therefore, the first approach that we investigated in order to decompose a sound aims at producing the sparsest representation [10]. The second approach was based on the non-negative matrix factorization (NMF), which we have refined for our needs [11, 12]. It exploits the redundancies in a musical piece (a single tone being generally repeated several times) in order to identify the sound elements via their spectral characteristics and their various occurrences through time.

3.2. Musical Applications

Analyzing a polyphonic recording in order to extract or to modify its musical content (*e.g.* the instruments, the rhythm or the notes) is a difficult exercise, even for an experienced musician. The DESAM project aimed at making a machine capable of performing such tasks. Let us mention some of them:

- The ability of identifying musical instruments from recordings is a key task in music indexation retrieval. An important characteristic of a sound that defines the perception of timbre is its spectral envelope.
- The ability to estimate the pitch of a sound (on a scale from low to high) is critical for identifying musical notes, but remains difficult in a polyphonic recording, because of the overlap of sounds.
- If producing a sound given a musical score happens to be easy both for the musician and computer, the inverse problem, called *automatic transcription*, which aims at recovering a musical score from a recording, proves to be much more complex and requires expert skills.
- Storing and transmitting an increasing volume of musical recordings requires coding this data in a format as compact as possible, making a compromise between the quantity of coded information, and the quality of the reproduced sound.

The decompositions into sound elements provide a representation of the signal as a sum of more elementary entities. From these entities, high level descriptors are extracted and are useful for instrument recognition, rhythm estimation, and multiple pitch estimation. These tasks are all necessary when the design of an automatic transcriber is targeted.

We have thus proposed new methods for estimating and comparing the spectral envelopes of musical sounds [13, 14]. We have also proposed original pitch estimation methods, capable of estimating up to ten simultaneous notes [15, 16], which have been used in an automatic transcription algorithm designed for piano music [17]. An alternate transcription scheme based on NMF has been developed for a larger class of instruments [18]. Besides, the precision of the decomposition permitted a physical analysis of sound production in musical instruments [19], and the development of more effective methods for coding and modifying sounds. Two approaches have been retained for coding. The first one, based on HR methods, permitted to reach very low bit rates [20]. The second one, based on sparse decompositions, was a scalable audio coder which can reach transparency [21]. Signal modifications were performed either by resampling the sinusoidal modeling parameters [22], or by modifying the sound elements of a sparse decomposition [23].

4. OVERVIEW OF THE TOOLBOX

The main contributions of the DESAM project have been integrated in the *DESAM Toolbox*, a software library written in Matlab language and distributed under the terms of the GNU General Public License (GPL). As shown on Figure 1, the toolbox is organized in three main parts. The first two parts are dedicated to the core of the toolbox, namely the representation of audio using sinusoidal or spectral models. The last part is dedicated to an application task: the transcription of polyphonic music.

The part dedicated to sinusoidal models is further separated into 2 subparts. The first one groups methods that deal with the estimation of the parameters of the sinusoidal model over a short time observation interval (*i.e.* frame level). The second one groups methods which address the issue of estimating and/or tracking those parameters over a long period of time (*i.e.* song level).

[2] http://perso.telecom-paristech.fr/rbadeau/desam

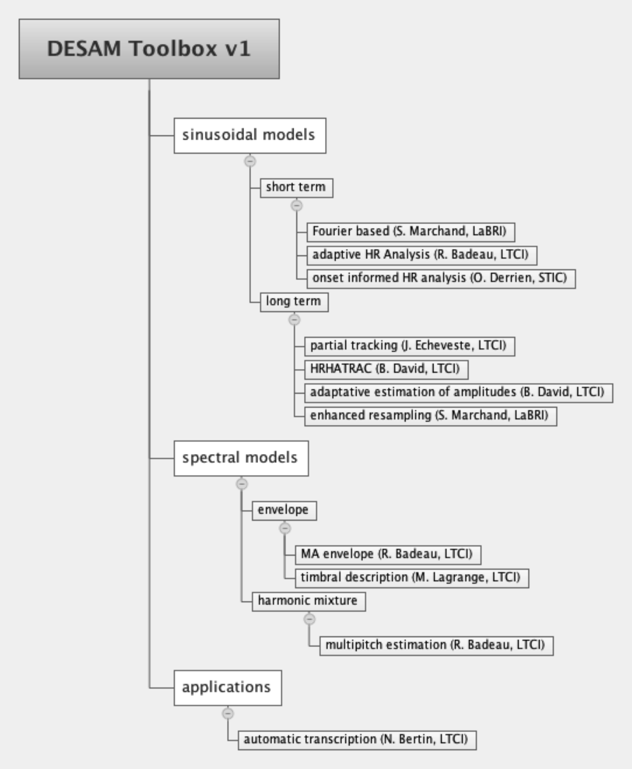
DESAM Toolbox v1
sinusoidal models
short term
Fourier based (S. Marchand, LaBRI)
adaptive HR Analysis (R. Badeau, LTCI)
onset informed HR analysis (O. Derrien, STIC)
long term
partial tracking (J. Echeveste, LTCI)
HRHATRAC (B. David, LTCI)
adaptative estimation of amplitudes (B. David, LTCI)
enhanced resampling (S. Marchand, LaBRI)
spectral models
envelope
MA envelope (R. Badeau, LTCI)
timbral description (M. Lagrange, LTCI)
harmonic mixture
multipitch estimation (R. Badeau, LTCI)
applications
automatic transcription (N. Bertin, LTCI)

Figure 1: *Overview of the DESAM Toolbox.*

Another complementary way of describing the spectral content of audio signals, explored in the second part, is to introduce constraints on spectral shapes or envelopes. Once estimated, they can be considered for deriving timbral features or estimating the fundamental frequencies of polyphonic sounds.

5. SINUSOIDAL MODELS

5.1. Short-term Models

Given a short observation window of a given audio signal, one would like to estimate the frequency and the amplitude of a given number of sinusoidal components. The DESAM Toolbox provides several and complementary ways of achieving this task. A first approach is able to estimate those parameters as well as their first derivative by considering the Fourier spectrum as its underlying representation. In order to get rid of frequency resolution constraints, a second approach considers High Resolution (HR) methods.

5.1.1. Fourier-based Methods

The DESAM Toolbox includes efficient non-stationary sinusoidal estimation using enhanced versions of either the reassignment method or the derivative method (see [9]).

Both methods have been proven to be equivalent in theory and practice and to achieve nearly-optimal results in terms of estimation precision (provided that the frequency resolution is sufficient to isolate the spectral peaks corresponding to the sinusoids), see [24, 25, 9, 26].

Although the resolution is still limited to the width of a bin of the discrete Fourier transform (see Figure 2 for an example of this frequency overlap phenomenon), the precision is close to the optimal. Moreover, the non-stationary model is more general than the one used in HR analysis, since it considers more general frequency modulations:

$$s(t) = \sum_{p=1}^{P} a_p(t) \exp(j\phi_p(t)),$$

where P is the number of partials, and

$$\begin{array}{rcl} a_p(t) & = & a_p \exp(\mu_p t), \\ \phi_p(t) & = & \phi_p + \omega_p t + \frac{1}{2}\psi_p t^2. \end{array}$$

For the reassignment method, the syntax is:

```
[a, mu, phi, omega, psi, delta_t] =
        reassignment (x, Fs, m)
```

The input parameters are:

- `x`, the signal frame to be analyzed;
- `Fs`, the sampling frequency (in Hz);
- `m`, the bin index where to pick the spectral peak (optional);

and the output values are `a`, `mu`, `phi`, `omega` and `psi`, corresponding respectively to the estimated amplitude, amplitude modulation, phase, frequency and frequency modulation of the spectral peak. `delta_t` is the reassigned time.

For the derivative method, the syntax is quite similar:

```
[a, mu, phi, omega, psi] =
derivative (x, d1, d2, Fs, m)
```

except that this function requires the first and second derivatives `d1` and `d2` of the signal `x`, which can be computed using the function

```
drv = discrete_derivative (src, Fs)
```

where `src` is the source signal and `drv` is its derivative. The test function of the discrete differentiation generates the Figure 1 of [9]. The `test_global.m` procedure generates the full tests (Figures 2–6 of [9]) with heavy computations requiring some time, though. For now, the implementation of the reassignment method is more efficient. Thus, we recommend this method for the estimation of the sinusoidal parameters. See the `test_example.m`[3] script for a small example using this method.

5.1.2. Simple High Resolution (HR) analysis/synthesis

An analysis/synthesis scheme for musical signals has been proposed in [27]. It is based on the Exponentially Damped Sinusoids (EDS) model:

$$s(t) = \sum_{p=1}^{P} \alpha_p \, z_p^t,$$

where P is the number of partials, the complex amplitudes α_p are of the form $\alpha_p = a_p \exp(j\phi_p)$ (where $a_p > 0$ is the real amplitude and $\phi_p \in \mathbb{R}$ is the phase of the partial p), and the complex

[3] Directory "sinusoidalModels/shortTerm/phaseBased"

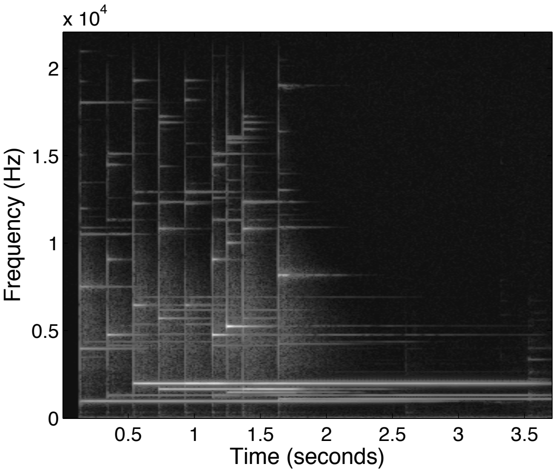

Figure 2: *Fourier Spectrogram of glockenspiel tones.*

poles z_p are of the form $z_p = \exp(\delta_p + j2\pi f_p)$ (where $\delta_p \in \mathbb{R}$ is the damping factor and $f_p \in \mathbb{R}$ is the frequency of the partial p).

This model is estimated by means of a HR subspace method called ESPRIT, which overcomes the spectral resolution limit of the Fourier transform, and achieves an accurate estimation of the sinusoidal components of the signal. Initially, the signal is pre-accentuated. The noise autoregressive envelope is estimated, and the signal is whitened with a FIR Filter [27].

The function `HR_analysis.m`[4] decomposes the audio signal into a sum of exponentially damped sinusoids, on a single time-segment, and the noise component is discarded. The function `HR_synthesis.m` re-synthesizes the audio signal corresponding to these EDS components. An optional wide-band whitening filter is proposed in the analysis function. The syntax is:

```
[poles, amplitudes] = HR_analysis(signal,
              order, whitening)

signal = HR_synthesis(poles, amplitudes,
              signal_length)
```

The parameters are:

- `signal`, signal to be analyzed or synthesized;
- `order`, model order;
- `whitening`, flag for optional whitening;
- `poles`, vector containing the signal poles;
- `amplitudes`, vector containing the complex amplitudes;
- `signal_length`, length of the signal to be synthesized (in samples).

5.1.3. Subband analysis and dynamic segmentation for HR audio coding

For an efficient audio coding application, the HR analysis/synthesis can be performed in frequency subbands and on a dynamic frame-by-frame basis. In order to reduce the complexity, we use a perfect-reconstruction filter bank, which keeps only the positive-frequencies in the signal. A dynamic frame-by-frame segmentation is performed according to an onset detection algorithm [28]. The corresponding script is `demo.m`. The audio codec described in [20] is based on this analysis/synthesis scheme.

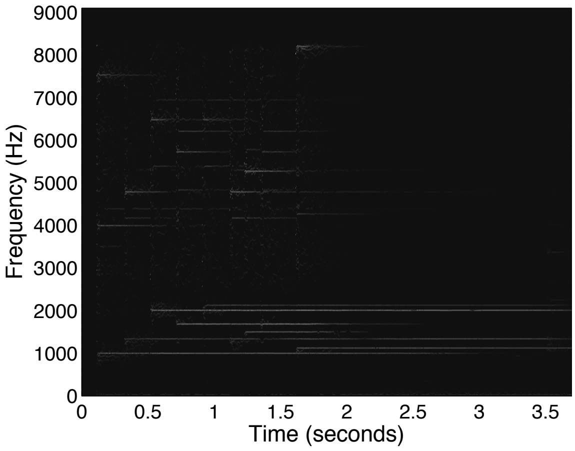

Figure 3: *Estimating the frequencies of glockenspiel tones using the fast HR subspace tracking method. Notice how the frequency overlap is reduced compared to a conventional Fourier-based spectrogram.*

5.1.4. Adaptive subband analysis and fast HR subspace tracking

An adaptive version of this method, using a classical filter bank (with subband-by-subband whitening) and a fixed frame-length analysis/synthesis, is also described in [27]. This scheme leads to a new representation, called the HR-ogram, where the signal components are represented as points in the time-frequency plane, see Figure 3. The stochastic part is then defined as the residual of this decomposition. The deterministic and stochastic parts can thus be processed separately, leading to high quality audio effects.

The program included in the DESAM Toolbox[5] decomposes the audio signal into a sum of exponentially damped sinusoids and autoregressive noise. Then the ESTER method [29] is used to estimate the number of sinusoids, and a fast adaptive algorithm (called Sequential Iteration) performs the subspace tracking [30]. Sinusoids/noise separation is achieved by projection onto the signal subspace and the noise subspace [31], and by reconstruction with the synthesis filter bank. The whole processing is described in Part III of reference [27], and was presented at Acoustics'08 [6]. The main function is `analyse.m`, whose syntax is the following:

```
[z,alpha,x] = analyse(s,Fs)
```

The input and output parameters are:

- `s`, signal to be analyzed;
- `Fs`, sampling frequency (preferably 44100 Hz);
- `z`, matrix containing the signal poles;
- `alpha`, matrix containing the complex amplitudes;

[4] Directory "sinusoidalModels/shortTerm/highResolution_lma".

[5] Directory "sinusoidalModels/shortTerm/highResolution_telecomParisTech".

- x, sinusoidal part of the signal.

This code was successfully applied to the decomposition of sounds from various musical instruments.

5.2. Long-term Models

5.2.1. Conventional Partial Tracking

Following pioneering work of MacAulay & al [32] and Serra & al [33], most partial tracking algorithms link together spectral peaks estimated using short-term methods such as the ones described in the previous section in order to form partials, *i.e.* sinusoidal oscillators whose parameters evolve slowly with time.

The approaches cited above, respectively termed Maq and Serra consider heuristics such as frequency proximity in order to enforce the slow variation of the parameters. In [34], more sophisticated approaches were developed in order to ensure those constraints, respectively by considering the predictability of the evolution and by enforcing that the spectral content of those evolution is of low frequency. Those approaches are respectively termed LP (for Linear Prediction) and HF (for High Frequency analysis), see Figure 4 for some results.

The syntax for the Maq tracking method is:

```
[P, Z,tag] = peaks2partialsMaq(A, F, tag,
                 Z, deltaF)
```

where

- A, F are respectively the amplitudes and the frequencies of the spectral peaks;
- tag is the partial index value assigned to the next created partial;
- Z is the state of the active tracks;
- deltaF is the maximal frequency deviation allowed between 2 successive peaks in any partial.
- P is the partial's label assigned to the spectral peaks.

The 3 other methods follow a similar syntax. One can notice that this syntax allows the trackers to be used in a streaming fashion as shown in the demo.m[6] script where the spectral data to be processed is split in two successive sets of peaks.

The four methods have been successfully tested in a small set of audio data. However, it should not be considered as authoritative implementation of the above described algorithms at this stage (as more extensive testing is still required).

5.2.2. Tracking of frequency components: the HRHATRAC algorithm [7]

HRHATRAC stands for *High Resolution HArmonics TRACking* and denotes an algorithm aiming at modeling musical sounds as multiple, slowly varying, spectral trajectories surrounded by an additive noise. HRHATRAC combines the efficiency of one of the most recent subspace tracking algorithm [35] with a gradient update for adapting the signal poles estimates. Hence it is able to update the frequency of each component from a time-instant to the next. It leads to a representation of the sinusoidal content of the signal in terms of spectral trajectories (slowly varying frequency components).

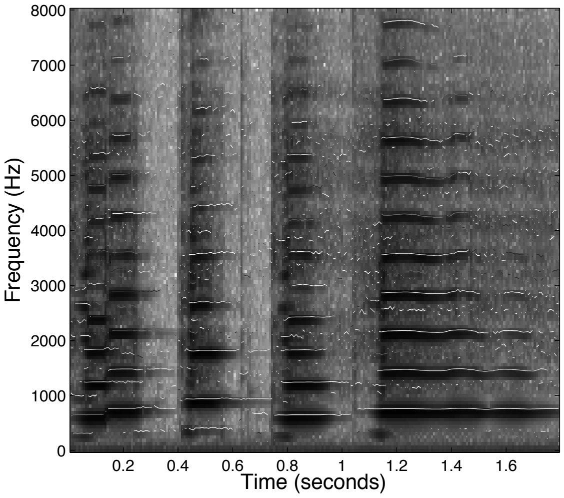

Figure 4: *Tracking the harmonics of a clarinet using the* LP *method.*

The corresponding DESAM Toolbox function hrhatrack returns for the kth frequency component its instantaneous frequency $f_p(t) = (2\pi)^{-1}\phi'_p(t)$ and instantaneous amplitude modulation $a_p(t)$. The sinusoidal part of the signal can thus be obtained as $\sum_p a_p(t)\exp(j\phi_p(t))$. The main function is hrhatrack.m[7], whose syntax is the following:

```
        [freqs,amps] =
hrhatrack(s,Ns,P,beta,muL,muV)
```

The input and output parameters are:

- s, signal to be analyzed;
- Ns, size of the signal subspace. Ideally, the number of complex frequency components, if it is known. If not known, overestimate Ns (typically by a factor 1.5 to 2);
- P, size of the autocovariance matrix ($P \times P$), default P = 3*Ns;
- beta, forgetting factor for the updated covariance matrix, default: 0.99;
- muL,muV, gradient steps for the eigenvalues and the eigenvectors updates respectively, default 0.9;
- freqs,amps, $f_p(t)$ and $a_p(t)$.

A demo script is also included: demo_hrhatrack.m.

5.2.3. Adaptive estimation scheme for instantaneous amplitudes

In [8] an adaptive estimation scheme is proposed for sample-wise update of the instantaneous amplitudes of known frequency components. The resulting decomposition of the signal in terms of sinusoidal content + noise is derived simultaneously. A fast sequential Least Squares algorithm is used which, for a given basis of r distinct frequency components, recursively derives the Least

[6] Directory "sinusoidalModels/longTerm/telecomParisTech_echeveste"

[7] Directory "sinusoidalModels/longTerm/telecomParisTech_david"

Squares estimates of the associated amplitudes and phases. While a direct calculation would achieve a $O(nr^2)$ (n being the total number of samples) complexity, the main cost of our implementation is only of $4r$ multiplications *per* sample.

The main function is `fastls.m`[8], whose syntax is the following:

```
[b,xc,er] = fastls(s,z,n)
```

The input and output parameters are:

- `s`, signal to be analyzed;
- `z`, z = [z1, z2, z3,...], complex poles;
- `n`, length of the analysis snapshot;
- `b`, matrix of the instantaneous amplitudes sequence;
- `xc`, sinusoidal part of the signal;
- `er`, residual (`s = xc + er`).

A demo script is also included: `demo1_fastls.m`.

5.2.4. Enhanced resampling

Once estimated using the above described methods, the parameters of the partials can be resampled using the method proposed in [22] which is suitable for general signals (non necessarily zero-centered or uniformly-sampled) and used for time scaling purposes as in [36]. The syntax is:

```
dst = enhanced_resample (src,
      src_param, dst_param)
```

where `src` and `dst` are the source and destination (resampled) signals, respectively. The other input parameters `src_param` and `dst_param` are either the sampling rate given as a scalar (uniform sampling case) or the sampling times given in a vector (non-uniform sampling case).

This function resamples the source signal according to the specified rates or times (see the `test_global.m`[9] function for a small example).

6. SPECTRAL MODELS AND APPLICATIONS

In this section, tools related to the estimation of the spectral envelope of a sound spectrum are presented, with applications to the study of the timber and the multipitch analysis.

6.1. ARMA envelope estimation

We proposed in [13] new algorithms for estimating autoregressive (AR), moving average (MA), and ARMA models in the spectral domain. These algorithms were derived from a maximum likelihood approach, where spectral weights are introduced in order to selectively enhance the accuracy on a predefined set of frequencies, while ignoring the other ones. This is of particular interest for modeling the spectral envelope of harmonic signals, whose spectrum only contains a discrete set of relevant coefficients. In the simple case of AR modeling, we proved that the proposed algorithm converges to the optimal solution, and that the convergence rate is enhanced by remapping the poles at each iteration. In the context of speech processing, our simulation results showed that the proposed method provides a more accurate ARMA modeling of nasal vowels than the Durbin method.

The main demonstration function is `test.m`[10], which performs the numerical simulations presented in [13], and plots the corresponding figure.

6.2. Timbral descriptors

Most of the approaches dealing with timbre description consider a concise encoding of the spectral envelope like the Mel-Frequency Cepstral Coefficients (MFCCs). Alternatively, we studied in [37] descriptors which are encoding the pseudo-periodic evolution of some relevant part of the envelope through time and demonstrated that those new features can conveniently be combined with the previous ones in order to improve the description capabilities of the features set.

Also, it is commonly considered that all the data is described using the same set of features. However, one can easily notice major discrepancies in some realistic settings between the database and the request side, for example regarding audio quality. Considering the matching scheme proposed in [14], a smooth description of the spectral content is required on the database side in order to maximize generalization properties. To that end, we studied various envelopes based on AR modeling and the True Envelope approach. On the request side, the audio quality can be much lower, only some prominent spectral peaks are considered.

The script `demo.m`[11] computes a large set of those spectro-temporal features over a predefined signal and displays its self-similarity matrix. Features based on MFCC computation rely on the Matlab code published by Dan Ellis [38].

6.3. EM-based Multiple pitch estimation

The problem of multi-pitch estimation consists in estimating the fundamental frequencies of multiple harmonic sources, with possibly overlapping partials, from their mixture. In [16], we introduced a novel approach for multi-pitch estimation, based on the statistical framework of the expectation-maximization algorithm, which aims at maximizing the likelihood of the observed spectrum. The proposed method was particularly promising, due to its robustness to overlapping partials, and its capacity to simplify the multi-pitch estimation task into successive single-pitch and spectral envelope estimations. It requires a proper initialization, involving a first stage of basic multi-pitch estimation for instance, and could advantageously make use of heuristics, in order to avoid staying trapped in local maxima. The effectiveness of this approach was confirmed by our simulations in the context of musical chord identification, performed on audio-like synthetic signals.

The main function is `test.m`[12], which performs the numerical simulations presented in [16], and plots the corresponding figures.

7. AUTOMATIC MUSIC TRANSCRIPTION

In reference [18], we presented a new method for automatic music transcription. The proposed approach was based on a Bayesian model which introduces harmonicity and temporal smoothness

[8] Directory "sinusoidalModels/longTerm/amplitudeEstimation"

[9] Directory "sinusoidalModels/longTerm/parametersResampling".

[10] Directory "spectralModels/envelope/telecomParisTech_badeau".

[11] Directory "spectralModels/envelope/telecomParisTech_lagrange"

[12] Directory "spectralModels/harmonicMixture/telecomParisTech_badeau".

constraints into the non-negative matrix factorization of time-frequency representations, providing a meaningful mid-level representation of the data. The model involves superimposed Gaussian components having a harmonic structure, while temporal continuity was enforced through an inverse-Gamma Markov chain prior. We then presented two algorithms which perform the maximum a posteriori (MAP) estimation of the model parameters. The first one [18] is a space-alternating generalized expectation-maximization (SAGE) algorithm, whereas the second one [39] is a novel multiplicative algorithm designed for minimizing an objective function composed of the Itakura-Saito divergence plus a prior penalty term. The proposed algorithms outperformed other benchmarked NMF approaches in a task of polyphonic music transcription, evaluated on a realistic piano music database.

Thus the program included in the DESAM Toolbox[13] computes the fast Bayesian constrained Itakura-Saito NMF of a magnitude ERB (equivalent rectangular bandwidth) transform, with basis spectra representing partial clusters and fundamental frequencies on the MIDI scale tuned at 440 Hz.

The main function is `bertin_multipitch.m`, whose syntax is the following:

```
bertin_multipitch(wavfile,
framewise_f0file,notewise_f0file)
```

The parameters are:

- `wavfile`, input wave file to be transcribed;
- `framewise_f0file`, output file with framewise fundamental frequency (f0) transcription every 10 ms;
- `notewise_f0file`, output file with transcription of the onset, offset and f0 of each note.

8. CONCLUSION

The initial version of the DESAM Toolbox was described. It introduces a rather large set of processing tools for estimating parameters of widely used spectral models. This toolbox is therefore aimed at the researchers community interested in the modeling of musical audio but could be of potential interest for anyone interested in audio and acoustics.

The Matlab code is distributed under the GNU Public License (GPL) and anyone willing to add new methods and/or improve the existing ones is encouraged to contact the maintainer of the toolbox (Mathieu Lagrange).

Further developments will include a more coherent Application Programming Interface (API) for methods tackling similar problems, Octave compatibility, and the development of experimental testbeds for evaluating and comparing the proposed methods.

9. REFERENCES

[1] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound.* The MIT Press, 1990.

[2] M. Slaney, "Auditory toolbox version2," Interval Research Corporation, Tech. Rep., 1998.

[3] S. Dubnov and M. Yazdani, "Computer audition toolbox (catbox)," 2010, online web resource. [Online]. Available: http://cosmal.ucsd.edu/cal/projects/CATbox/catbox.htm

[4] O. Lartillot and P. Toiviainen, "A Matlab toolbox for musical feature extraction from audio," in *International Conference on Digital Audio Effects (DAFx-07)*, 2007.

[5] E. Pampalk, "A Matlab Toolbox to Compute Similarity from Audio," in *Proceedings of the ISMIR International Conference on Music Information Retrieval (ISMIR'04)*, 2004.

[6] R. Badeau and B. David, "Adaptive subspace methods for high resolution analysis of music signals," in *Acoustics'08*, Paris, France, June-July 2008.

[7] B. David, R. Badeau, and G. Richard, "HRHATRAC algorithm for spectral line tracking of musical signals," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP'06)*, vol. III, Toulouse, France, May 2006, pp. 45–48.

[8] B. David and R. Badeau, "Fast sequential LS estimation for sinusoidal modeling and decomposition of audio signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, Oct. 2007, pp. 211–214.

[9] S. Marchand and P. Depalle, "Generalization of the derivative analysis method to non-stationary sinusoidal modeling," in *11th International Conference on Digital Audio Effects (DAFx-08)*, Espoo, Finland, September 2008, pp. 281–288.

[10] L. Daudet, "Audio sparse decompositions in parallel - Let the greed be shared!" *IEEE Signal Processing Magazine, Special Issue on Signal Processing on Platforms with Multiple Cores: Part 2 – Design and Applications*, vol. 27, no. 2, pp. 90–96, March 2010.

[11] N. Bertin, C. Févotte, and R. Badeau, "A tempering approach for Itakura-Saito non-negative matrix factorization. With application to music transcription." in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'09)*, Taipei, Taiwan, April 2009, pp. 1545–1548.

[12] R. Hennequin, R. Badeau, and B. David, "NMF with time-frequency activations to model non-stationary audio events," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'10)*, Dallas, Texas, USA, March 2010, pp. 445–448.

[13] R. Badeau and B. David, "Weighted maximum likelihood autoregressive and moving average spectrum modeling," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'08)*, Las Vegas, Nevada, USA, March-April 2008, pp. 3761–3764.

[14] M. Lagrange, R. Badeau, and G. Richard, "Robust similarity metrics between audio signals based on asymmetrical spectral envelope matching," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'10)*, Dallas, Texas, USA, March 2010, pp. 405–408.

[15] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech and Language Processing*, 2010, to be published. [Online]. Available: http://www.ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5356234&isnumber=4358086

[16] R. Badeau, V. Emiya, and B. David, "Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra," in *International*

[13] Directory "applications/multiPitch/telecomParisTech_bertin".

Conference on Acoustics, Speech, and Signal Processing (ICASSP'09), Taipei, Taiwan, April 2009, pp. 3073–3076.

[17] V. Emiya, R. Badeau, and B. David, "Automatic transcription of piano music based on HMM tracking of jointly-estimated pitches," in *16th European Signal Processing Conference (EUSIPCO)*, Lausanne, Sweden, August 2008.

[18] N. Bertin, R. Badeau, and E. Vincent, "Enforcing Harmonicity and Smoothness in Bayesian Non-negative Matrix Factorization Applied to Polyphonic Music Transcription," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 538–549, March 2010.

[19] J.-L. Le Carrou, F. Gautier, and R. Badeau, "Sympathetic string modes in the concert harp," *Acta Acustica united with Acustica*, vol. 95, no. 4, pp. 744–752, July-August 2009.

[20] O. Derrien, G. Richard, and R. Badeau, "Damped sinusoids and subspace based approach for lossy audio coding," in *Acoustics'08*, Paris, France, June-July 2008.

[21] E. Ravelli, G. Richard, and L. Daudet, "Union of MDCT bases for audio coding," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 8, pp. 1361–1372, November 2008.

[22] M. Raspaud and S. Marchand, "Enhanced resampling for sinusoidal modeling parameters," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'07)*, New Paltz, New York, USA, October 2007, pp. 327–330.

[23] O. Derrien, "Time-Scaling Of Audio Signals With Multi-Scale Gabor Analysis," in *10th Conference on Digital Audio Effects (DAFX'07)*, Bordeaux, France, September 2007, pp. 1–6.

[24] S. Marchand and M. Lagrange, "On the Equivalence of Phase-Based Methods for the Estimation of Instantaneous Frequency," in *Proceedings of the 14th European Conference on Signal Processing (EUSIPCO'2006)*. Florence, Italy: EURASIP, September 2006.

[25] M. Lagrange and S. Marchand, "Estimating the Instantaneous Frequency of Sinusoidal Components Using Phase-Based Methods," *Journal of the Audio Engineering Society*, vol. 55, no. 5, pp. 385–399, May 2007.

[26] B. Hamilton, P. Depalle, and S. Marchand, "Theoretical and Practical Comparisons of the Reassignment Method and the Derivative Method for the Estimation of the Frequency Slope," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'09)*, New Paltz, New York, USA, October 2009.

[27] R. Badeau, "High resolution methods for estimating and tracking modulated sinusoids. Application to music signals." Ph.D. dissertation, École Nationale Supérieure des Télécommunications, Paris, France, April 2005, ParisTech 2006 PhD Award. [Online]. Available: http://www.perso.enst.fr/rbadeau/unrestricted/Thesis.pdf

[28] C. Duxbury, M. Sandler, and M. Davies, "A hybrid approach to musical note onset detection," in *5th International Conference on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, September 2002.

[29] R. Badeau, B. David, and G. Richard, "A new perturbation analysis for signal enumeration in rotational invariance techniques," *IEEE Transactions on Signal Processing*, vol. 54, no. 2, pp. 450–458, February 2006.

[30] R. Badeau, R. Boyer, and B. David, "EDS parametric modeling and tracking of audio signals," in *5th International Conference on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, September 2002, pp. 139–144.

[31] B. David, G. Richard, and R. Badeau, "An EDS modelling tool for tracking and modifying musical signals," in *Stockholm Music Acoustics Conference (SMAC 2003)*, vol. 2, Stockholm, Sweden, August 2003, pp. 715–718.

[32] R. J. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, August 1986.

[33] X. Serra and J. O. Smith, "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[34] M. Lagrange, S. Marchand, and J. Rault, "Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1625–1634, May 2007.

[35] R. Badeau, B. David, and G. Richard, "Fast Approximated Power Iteration Subspace Tracking," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 2931–2941, Aug. 2005.

[36] M. Raspaud, S. Marchand, and L. Girin, "A Generalized Polynomial and Sinusoidal Model for Partial Tracking and Time Stretching," in *Proceedings of the Digital Audio Effects (DAFx'05) Conference*, Madrid, Spain, September 2005, pp. 24–29.

[37] M. Lagrange, M. Raspaud, R. Badeau, and G. Richard, "Explicit modeling of temporal dynamics within musical signals for acoustical unit similarity," *Pattern Recognition Letters*, 2010, to be published.

[38] D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource. [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

[39] N. Bertin, R. Badeau, and E. Vincent, "Fast Bayesian NMF algorithms enforcing harmonicity and temporal continuity in polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, October 2009, pp. 29–32.