AUTOMATIC INSTRUMENT RECOGNITION IN A POLYPHONIC MIXTURE USING SPARSE REPRESENTATIONS

Pierre Leveau[1,2]
[1] GET-ENST (Télécom Paris)
46, rue Dareau
75014 Paris

David Sodoyer[2], **Laurent Daudet**[2]
[2] University Pierre et Marie Curie
Institut Jean Le Rond D'Alembert, LAM team
11, rue de Lourmel
75015 Paris

ABSTRACT

In this paper, we introduce a method to address automatic instrument recognition in polyphonic music. It is based on the decomposition of the music signal with instrument-specific harmonic atoms, yielding an approximate object representation of the signal. A post-processing is then applied to exhibit ensemble saliences that give clues about the number of instruments and their labels. The whole algorithm is then applied on artificial mixes of solo performances. The identification of the number of instrument reaches 73 % on 10-s segments and the fully blind problem of identification of the ensemble label without prior knowledge on the number of instruments is 17 %.

1 INTRODUCTION

Orchestration is a critical information for the automatic indexing of music. It gives an important clue about the music genres, and is often necessary for the query of sound samples for electronic music composing.

Automatic Instrument Recognition has raised some interest over the latest years (see [1] for an overview). Early studies have addressed the recognition of isolated music notes, then of solos phrases. For these two contexts, machines now reach the performance of expert musicians. However, mono-instrument music is only a small part of the overall recorded music, that involves natural or artificial mixes of instruments.

To deal with multi-instrument music, several strategies have been adopted. Template-based approaches have first been proposed [2, 3]. Other approaches adapt "bag-of-frames" approaches to polyphony [4]. Other techniques consist in estimating jointly the instrument sources activated in a probabilistic framework [5], at a heavy computational cost. A recent work [6] presents a representation showing the instrument presence probabilities in the time-pitch plane without note detection. Ensemble classes can also be modeled using standard feature-based representations in addition with a hierarchical taxonomy [7], when the number of instrument combinations is tractable.

In this paper a recent development in the decomposition of music signals is studied for the recognition of music instrument in ensemble music. It relies on principles coming from the sparse approximations domain. To get a useful sparse representation of a signal, two aspects have to be investigated: the building of a signal model (*dictionary design*), and, given a dictionary, the choice of an algorithm and its optimization towards a faster or better approximation. Techniques from the sparse approximation domain have already been used for automatic music transcription in an unsupervised way [8]: the building of the dictionary was done in an data-driven way, prohibiting the signal analysis in view of prior knowledge of the sources. The introduction of prior knowledge about the sources in dictionaries has been presented in [9]: this knowledge is put in the amplitudes of the note partials. In section 2, the decomposition algorithm is briefly described, then a post-processing is introduced to take a decision on the orchestration. The experiments of ensemble recognition are detailed in section 3.

2 ALGORITHM

2.1 Decomposition algorithm

The signal model and decomposition algorithm have been introduced in [9]. For space constraints, only the main features of the algorithm are highlighted here.

2.1.1 Signal Model

The signal is decomposed as a linear combination of short pieces of signal h, called *harmonic atoms*:

$$x(t) = \sum_{n=1}^{N} \alpha_n \, h_{s_n,u_n,f_{0_n},A_n,\Phi_n}(t). \qquad (1)$$

The set of all the atoms available to decompose the signal is called a *dictionary*.

The parameters of these atoms are the scale s_n, the time localization u_n, the fundamental frequency f_{0_n}, the partial amplitudes $A_n = \{a_{m,n}\}_{m=1:M}$ and the partial phases $\Phi_n = \{\phi_{m,n}\}_{m=1:M}$. An atom h is itself defined as a

linear combination of partial atoms:

$$h_{s,u,f_0,A,\Phi}(t) = \sum_{m=1}^{M} a_m \, e^{j\phi_m} g_{s,u,m.f_0}(t) \qquad (2)$$

where the amplitudes of the M partials are constrained to $\sum_{m=1}^{M} a_m^2 = 1$ and the signal g corresponding to each partial is given by a *Gabor* atom:

$$g_{s,u,f} = w\left(\frac{t-u}{s}\right) e^{2j\pi ft} \qquad (3)$$

with w a time and frequency localized window.

In our study, each A vector is linked to an instrument and i a pitch p (integer Midi Code), and is learned from databases of isolated instrument notes or solo performances, as shown in section 2.1.3.

2.1.2 Decomposition algorithm

The Matching Pursuit algorithm [10] is then performed to decompose the signal with this dictionary. In a nutshell, it consists in selecting the atom the most correlated with the signal, to subtract it from the signal, and to iterate on the residual. After several iterations, a decomposition of type (1) is obtained. The set of the selected atoms with their respective weights is called a Book. With the parameters mentioned in Section 3, the runtime takes about 10 times real-time on a 3 GHz monoprocessor, with a Matlab implementation.

2.1.3 Learning

The atoms are first learned on a set of 3 different databases of isolated notes [11, 12, 13], annotated in pitch p and instrument i. For a given time frame of size s, the technique consists in selecting the harmonic comb that best matches the signal, and then in picking the amplitudes of the partials on this comb. For each instrument and note, the amplitude vectors sets are then quantized using a K-means algorithm.

To build dictionaries that are closer to realistic playing conditions, solo performances are then analyzed. In this case, the notes are not localized nor annotated. The Matching Pursuit algorithm with the aforementioned dictionary is thus used because of its capability to automatically adapt to the music notes. The dictionary that we use for this task is built only with the atoms of the corresponding known instrument, whose construction has been described in the previous paragraph. However, the algorithm is modified before the subtraction step: the harmonic comb whose fundamental frequency corresponds to the selected atom is selected to perform the partial amplitudes picking. Moreover, to prevent the selection of harmonic atoms in the residual, the atom is not subtracted: the signal is set to 0 at the extracted atom localization. A quantization step is then performed, as for isolated notes dictionaries.

2.2 Pitch-and-instrument salience

An atom extraction can be seen as a "pitch-and-instrument" salience extractor, since it correlates both a spectral envelope and a harmonic comb with the signal. Given an extracted atom at fundamental frequency f_0, scale s and localization u, we define the f_0-and-instrument salience for instrument i as [1] :

$$S_i = \max_{A \in \mathcal{C}_{i,p}} \{|\langle x, h_{s,u,f_0,A,\Phi_n}\rangle|\} \qquad (4)$$

If an instrument i enveloppe cannot play the pitch p, i.e. $\mathcal{C}_{i,p} = \emptyset$, its salience is set to 0. Although not required for the decomposition, all instrument saliences for every selected atom are kept for the scoring step: they are needed for the ensemble saliences evaluations.

2.3 From Instrument Salience to Ensemble Salience

The scoring algorithm processes the output of the decompositions to have an indication of which instruments are playing. Here, a frame-based scoring is developed: for a given time frame, the score of a given ensemble class depends on which atoms have been extracted and on their f_0-and-instrument salience.

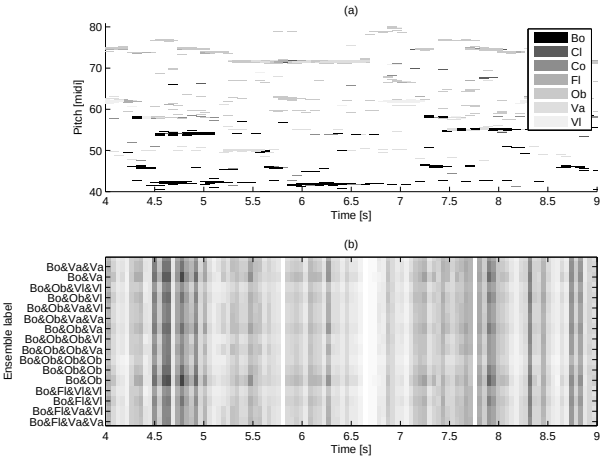

Figure 1. Bassoon (Bo) and Oboe (Ob) duo (synthetic mix): (a) Book representation in the Time-Pitch plane: atoms are represented by rectangles, whose width is the atom scale and height is their amplitudes, (b) Ensemble Saliences for a subset of ensemble labels (high saliences are darker).

Given the decomposition of a music signal, there can be several atoms per time frame since the music is in general polyphonic. The first step to perform is to select which atoms are present for each time frame, the timeline being sampled at the greatest common divider between the Δu corresponding to each scale. Then, the contribution of each atom a on a given time sample is equal to the value

[1] Note that the inner product is not depending on the values of Φ if f_0 is high enough since the partials atoms can be considered as orthogonal: $|\langle x, h_{s,u,f_0,A,\Phi_n}\rangle|^2 = \sum_{m=1}^{M} |\langle x, g_{s,u,m.f_0}\rangle|^2$

at instant u of the weighting window starting at u_a multiplied by the atom weight. Hence, given a time frame u and an ensemble label e, its ensemble salience is the following [2] :

$$\mathcal{S}_e(u) = \frac{\max_{C_e \in \mathcal{C}_e} \sum_{a \in C_e} \mathcal{S}_{ia}(u) w(\frac{u-u_a}{s_a})}{N_e^\beta} \quad (5)$$

where $\mathcal{C}_e$ is the set of all the instrument salience combinations whose time support overlap with u. For example, if two atoms are present at time u, the salience of ensemble *Co&Fl* (Cello and Flute) is the maximum between the sum of the *Fl* salience for the first atom and the *Co* salience for the second one, and sum of the *Co* salience for the first atom and the *Fl* salience for the second one, divided by 2^β. An example of book output and corresponding ensemble saliences is displayed on Figure 1.

The β parameter is a sparsity parameter: it balances the weight between the sum of all atom saliences and the number of instrument taken to explain the resulting signal. It can be optimized with respect to the number of instrument detection.

2.4 Voting

Decisions taken on single time frames does not provide useful information as such. However, one can be interested on decisions taken on the whole music signal, or a segment of it. To get a global decision from local ones, voting techniques must be employed. The technique used in this study is derived from a probabilistic framework. Other techniques, like majority-vote, have been tried but they yield to weaker results. First, the ensemble saliences are mapped to ensemble Pseudo Log-Likelihoods (PLL), then a segment PLL for each ensemble label is computed by adding the PLL of each time frames. The mapping of a ensemble salience $\mathcal{S}_e(u)$ to PLL $\mathcal{L}_e(u)$ is achieved with the following formula: $\mathcal{L}_e(u) = (\mathcal{S}_e(u))^\gamma$, where γ weighs the influence of salience amplitudes over the overall score in the segment. Like β in previous Section, the γ coefficient has to be optimized on a development set. The decision over the all segment is obtained by summing all the PLL. It corresponds to an hypothesis of statistical independence between each time frame. This hypothesis is clearly erroneous in music signals (the orchestration does not change at every short time frame), but is commonly taken for fusion of local likelihood.

3 EXPERIMENTS

3.1 Parameters

The parameters used for the decomposition are $s = 46ms$, $\Delta = 23ms$. f_0 is sampled logarithmically with a step of $1/10$ tone. The decompositions are performed until the Signal-To-Residual ratio reaches 20 dB.

The development set and the test set are composed of artificial mixes of solo phrases extracted from commercial CDs, from sources different from the one used for atom learning. The mixes are done by summing the mono-instrument signals of instruments bassoon (Bo), cello (Co), clarinet (Cl), flute (Fl), oboe (Ob), viola (Va) and violin (Vl) after an energy normalization. For each set, 100 10s samples have been made, 25 for each ensemble cardinal.

[2] Using the $L2$ norm $\sqrt{\sum_{a \in C_e} (\mathcal{S}_{ia}(u) w(\frac{u-u_a}{s_a}))^2}$ instead of the $L1$ norm $\sum_{a \in C_e} \mathcal{S}_{ia}(u) w(\frac{u-u_a}{s_a})$ would be more consistent with the optimality criterion of the decomposition, however it leads to weaker results in the studied applications

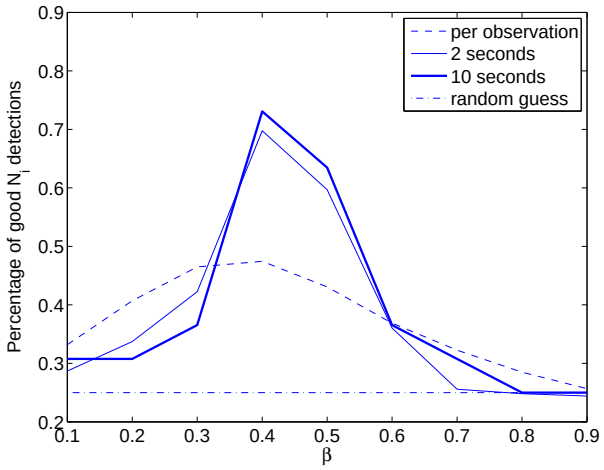

Figure 2. Accuracy of Number of Instrument Detection as a function of β for decision on single times frames, 2 seconds segments and 10 seconds segments.

3.2 Optimization of parameters

The parameters β and γ have to be tuned to maximize the accuracy of the estimation of the number of instruments, which is required to estimated the good instrument label. Optimizing these parameters for instrument label accuracy would overfit the algorithm for the solo recognition, that is the easiest problem. In our experiments on the development set, the best γ parameter has shown to be independant on the decision window: the value $\gamma = 0.8$ gives the best results.

3.3 Ensemble recognition

For these values, the instrument recognition rates for decisions on 10 s segments are depicted on Figure 3. It shows that the problem of finding an instrument among the mix is correctly addressed when the number of instrument is known (from 70 % to 100 %, depending on the ensemble type), and a less accurately when it is not known (from 54 % to 84 %). However, as the required number of instruments increases, the method fails at correctly identifying them alltogether. Dealing with ensembles of more than three instruments needs more refined techniques both at decomposition step and post-processing or more prior information, since the problem is significantly more difficult (results for random draw is at less than 1 %).

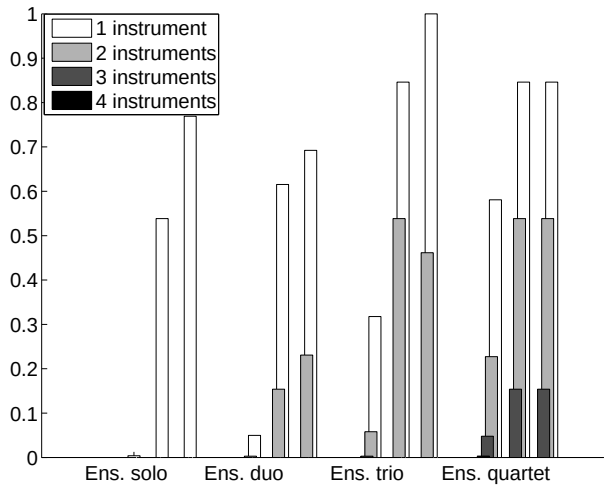

Figure 3. Ensemble recognition results for each subset (solos, duos, trios, quartets). For each ensemble, the three groups of bars depict respectively the results of a random draw, the results of our algorithm with no knowledge on the number of instruments playing, and the results knowing the number of instruments playing.

4 CONCLUSION

In this paper, we have developed a novel approach to the highly complex problem of identifying the instruments playing in ensemble music. The approach consists in getting a knowledge-assisted mid-level representation of the signal, then in performing a post-processing using ensemble saliences based on individual instrument saliences derived from the representation. The results are encouraging for the estimation of the number of instrument, but weak for the ensemble classification, which is a much more difficult problem without prior information on ensemble labels occurences.

Future work will be dedicated to the improvement of the decomposition step by first refining atom parameters, in order to better fit the underlying signal structures, and then by grouping atoms into molecules to catch temporal dependencies. The joint estimation of atom combinations will also be investigated using more elaborated sparse decomposition algorithms. Finally, the post-processing will be improved by using melodic line tracking techniques, to disambiguate highly polyphonic mixes.

5 ACKNOWLEDGEMENTS

L. Daudet is partially supported by French ANR under contract ANR-06- JCJC-0027-01- DESAM.

6 REFERENCES

[1] P. Herrera, A. Klapuri, and M. Davy. *Signal Processing Methods for Music Transcription*, chapter 6 - Automatic Classification of Pitched Musical Instrument Sounds. Springer, 2006.

[2] K. Kashino and H. Murase. A sound source identification system for ensemble music based on template adaptation and music stream extraction. *Speech Communication*, 27:337–349, March 1999.

[3] T. Kinoshita, S. Sakai, and H. Tanaka. Musical sound source identification based on frequency component adaptation. In *Proc. IJCAI Worshop on Computational Auditory Scene Analysis*, pages 18–24, 1999.

[4] J. Eggink and G. J. Brown. Instrument recognition in accompanied sonatas and concertos. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, 2004.

[5] E. Vincent. Musical source separation using time-frequency source priors. *IEEE Trans. on Audio, Speech and Language Processing*, 14(1):91–98, January 2006.

[6] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and G. Okuno. Instrogram: A new musical instrument recognition technique without using onset detection nor f0 estimation. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, volume 5, pages 229–232, 2006.

[7] S. Essid, G. Richard, and B. David. Instrument recognition in polyphonic music based on automatic taxonomies. *IEEE Trans. on Audio, Speech and Language Processing*, 14(1):68–80, January 2006.

[8] P. Smaragdis and J.C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. of IEEE Int. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 177–180, 2003.

[9] P. Leveau, E. Vincent, G. Richard, and L. Daudet. Mid-level sparse representations for timbre identification: design of an instrument-specific harmonic dictionary. In *1st Workshop on Learning the Semantics of Audio Signals*, dec 2006.

[10] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. on Signal Processing*, 41(12):3397–3415, December 1993.

[11] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC Musical Instrument Sound Database. Distributed online at http://staff.aist.go.jp/m.goto/RWC-MDB/.

[12] Iowa database, http://theremin.music.uiowa.edu/mis.html.

[13] Studio online database, http://forumnet.ircam.fr/402.html?l=1.