19th INTERNATIONAL CONGRESS ON ACOUSTICS
MADRID, 2-7 SEPTEMBER 2007

TOWARDS POLYPHONIC MUSICAL INSTRUMENTS RECOGNITION

PACS: 43.55.Cs

Richard, Gaël[1]; Leveau, Pierre[1]; [2]; Daudet, Laurent[2]; Essid, Slim[1]; David, Bertrand[1]
[1] GET-ENST (TELECOM-Paris), 37 rue Dareau, 75014 Paris, France
[2] IJLRDA-LAM, University Pierre and Marie Curie-Paris 6, 11 rue de Lourmel, 75015 Paris, France

ABSTRACT
Automatic musical instrument recognition is a relatively new topic in the growing field of Music Information Retrieval. Early studies mostly focused on instrument recognition from recordings of isolated notes. More recently, some studies tackled the problem of musical phrases played in solo (*i.e.* without accompaniment) which better covers the timbre variability of a given instrument. However, the current trend is now to deal with true polyphonic music (*i.e.* involving multiple instruments), which appears to be a far more difficult problem but with more practical applications. The aim of this paper is to provide an overview of the state-of-the-art in automatic musical instrument recognition with a focus on recent and innovative approaches applied to true polyphonic music. It will be shown that the traditional "bag of frames" approaches can obtain interesting results by building efficient automatic taxonomies or by using complementary information to enhance the relevant signal. We however argue that it is important to consider new directions to overcome the limitations of these traditional approaches. One of these promising directions that will be detailed concerns mid-level representations, which are based on the decomposition of the signal into a small number of sound atoms or molecules bearing explicit musical instrument labels.

INTRODUCTION
There is a growing interest for new means of interaction with audio information that is nowadays mostly available in digital format and stored in large databases. There is therefore a strong need for efficient audio indexing techniques that would allow the extraction of a detailed and meaningful symbolic representation directly from a digital audio recording. For music signals, this representation will include information about the metric, the harmony, the melody, the genre or the interpretation style and will ultimately be represented under the form of an enriched music sheet. The availability of such a symbolic representation opens the path for numerous Music Information Retrieval (MIR) applications including content-based search by similarity, cover songs retrieval, automatic post-remixing, ...

Automatic musical instrument recognition is a relatively new topic in the growing field of MIR and aims at recognizing which instruments are played in a given music sample. This topic appears to be an essential step since several applications would directly rely on the knowledge of the instrumentation. Such applications include music synthesis from automatically learned templates, music track separation for personal remixing, parametric low-bit rate audio coding or more generally for music search by content similarity. Early studies mostly focused on instrument recognition from recordings of isolated notes. More recently, some studies tackled the problem of musical phrases played in solo (*i.e.* without accompaniment) which better cover the timbre variability of a given instrument, and is closer to standard playing conditions. However, the current trend is now to deal with true polyphonic music (*i.e.* involving multiple instruments), which appears to be a far more difficult problem but with more practical applications. The aim of this paper is to provide an overview of the state-of-the-art in automatic musical instrument recognition with a focus on

approaches applied to true polyphonic music.

The paper is organized as follows. First, a brief overview of the traditional "bag-of-frames" approach is proposed in next session. Then, in the following section, an alternative and very promising approach based on sparse atomic decomposition of the audio signal is described. Finally, some conclusions are suggested.

TRADITIONAL "BAG-OF FRAMES" APPROACH

The most straightforward and traditional approach for music instrument recognition is based on a training phase in which instruments models are built (see figure 1). For this phase, a training database gathering a representative (and ideally very large) set of signals for each musical instrument is used. A set of features is then computed for each class of the training database, usually on successive short signal segments (called frames) and are then used either as reference templates (during testing phase) or used to build a statistical model for each class. In the recognition phase, the same set of features is computed and is compared either to the templates or to the model to determine what are the most probable instruments currently playing.

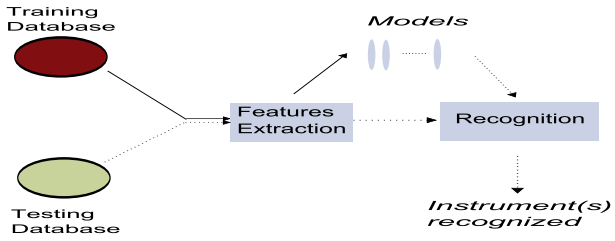

Figure 1: A classic architecture

Feature extraction

The feature extraction module aims at representing the audio signal using a reduced set of features that characterize the signal properties well with a view to discriminating the instruments. The features proposed in the literature can be roughly classified in four categories:

- Temporal features: These features are directly computed on the time domain signal. The advantage of such features is that they are usually straightforward to compute. They include the crest factor, temporal centroid, zero-crossing rate and envelope amplitude modulation
- Cepstral features: Such features are widely used in speech recognition or speaker recognition. This is duly justified by the fact that such features allow to estimate the contribution of the filter (or vocal tract) in a source-filter model of speech production. They are also often used in audio indexing applications since many audio sources also obey a source filter model. The usual features include the Mel-Frequency Cepstral Coefficients (MFCC), the Linear-Predictive Cepstral Coefficients and their first and second derivatives.
- Spectral features: These features are usually computed on the spectrum (magnitude of the Fourier Transform) of the time domain signal. They include the first four spectral statistical moments, namely the spectral centroid, the spectral width, the spectral asymmetry defined from the spectral skewness, and the spectral kurtosis describing the peakedness/flatness of the spectrum. A number of spectral features were also defined in the framework of MPEG7 [14]. Frequency derivative of the constant-Q coefficients, describing spectral irregularity or smoothness were also reported to be successful by Brown [2]; Octave Band Signal Intensities were later proposed by S. Essid & al. ([9]), to capture in a rough manner the power distribution of the different harmonics of a musical sound without recurring to pitch-detection techniques.
- Perceptual features: Relative specific loudness, sharpness and spread can also be used.

To obtain centered and unit variance features, a linear transformation is often applied to each computed feature. This normalization procedure is usually more robust to outliers than a mapping of the feature dynamic range to [-1 : 1]. More details on most of these common features can be found in [24]. When a large number of features is chosen, it becomes necessary to use feature selection techniques to reduce the size of the feature set. It is then possible to select the most effective features for the problem at hand, that is in this case to select the features that best discriminate the instruments considered.

Classification

Different classification strategies were proposed in the context of automatic musical instrument recognition (see for example [13] for a review). Early studies were based on the K-Nearest Neighbors algorithm for musical recognition on isolated notes ([16], [7]) and more recently on musical phrases ([22]). The current trend is however to use more sophisticated modelling approaches (Discriminant analysis in [1], neural networks in [19], Gaussian mixture models in [2],[9] and Support Vector machines in [23], [9]). An interesting research direction consists in integrating a model of temporal dynamics by means of Hidden Markov Models, for example (see [18]).

From isolated notes to polyphonic signals

Most of the studies in automatic instrument recognition only tackle the problem of monophonic sources (e.g. only one instrument played at a time). On the contrary, most real musical sources correspond to a mixture of several instruments. In most cases, the methods designed for the monophonic case will not directly work for the polyphonic case. In fact, the feature extraction process is highly non-linear and cannot exploit a simple additivity of the different sources. However, several approaches were already proposed to cope with this limitation. For example, in [4], the missing feature theory is applied by ignoring the features that are perturbed by other sources. In a later work [5], it is proposed to extract the predominant harmonic combs related to the predominant fundamental frequency. These harmonic combs are then used as features for automatic instrument recognition in polyphonic music. In a sense, this approach uses the concept of prior source separation before conducting the individual instrument identification. Such a concept is further developed in [15], where the polyphonic signals are first preprocessed by a source separation approach based on independent subspace analysis. The resulting signals are then classified by traditional classifiers such as K nearest neighbors (K-NN) or Gaussian Mixture models (GMM). This concept is also particularly well adapted for percussive or drum signals transcription as demonstrated in [10]. Some other approaches will rely on more specific models, such as pitch dependent instrument models [17] while some other will pay particular attention to limit overtraining by learning note models on polyphonic or mixed signals [6]. In [25], instrument spectrum models are put in a probabilistic framework to perform both instrument identification and transcription of the input. This method also has the advantage to take the additivity of the music signals into account. Another interesting direction consists in learning a model for any possible combination of instruments for a given music genre. For example, in a jazz trio (bass, piano, drum), a total of seven classes will be defined and learned. To overcome the potential complexity burden, [8] further proposed to follow a hierarchical methodology where the taxonomy is automatically learned from labelled data. This method was successfully applied to jazz ensembles. However, this approach needs to have labelled data with information for any given time about which instruments are currently playing, and gathering such data may be rather tedious.

The quick overview given above shows that the "bag-of-frames" approach is popular and it often provides satisfying performance. However, despite the promising results obtained, it is also important to investigate other paths which would at the same time allow for training the individual instrument models on isolated notes or solos, be able to cope with the large variability of sounds produced by any given instrument and finally limit the perturbation of the other sources (instrument) in the transcription process. One of this research directions relies on a signal model that is supposed to well represent the studied classes of instrument. Such methods aim at representing the signal as an explicit linear combination of sound sources, which can be adapted to better fit the analyzed signal. These approaches are at the intersection of source separation, automatic transcription and musical instrument identification and are often referred to as mid-level representation since a further post-processing is usually required to tackle any of the mentioned applications. One of these approaches, based on a decomposition of the signal into elementary

sound templates, is briefly recalled below but the interested reader is invited to consult [20] and [21] for a more in-depth presentation.

SPARSE ATOMIC DECOMPOSITION OF POLYPHONIC MUSIC SIGNALS

Principle

Since "bag-of-frames" methods are limited in describing music as a sum of musical sources, approaches that extract additive features have to be developed. The method described below fulfills this goal by getting a structured representation of the musical sounds, where each element of the representation can be considered as a sound object.

The basic concept of the sparse approximation derives from the "atomic" decomposition of audio signals. In this case, the polyphonic signal is decomposed into a small number of sound atoms or molecules, where each atom consists of windowed harmonic sinusoidal partials and each molecule of several atoms spanning successive time windows. Each atom is associated with a specific instrument and a specific pitch by learning the amplitudes of its partials on separate solo data.

The overall goal of sparse decomposition is then to approximate a given signal $x(t)$ as a linear combination of atoms $h_\lambda(t)$ taken from a fixed *dictionary* $\mathcal{D} = \{h_\lambda(t)\}_\lambda$

$$x(t) = \sum_{\lambda \in \Lambda_x} \alpha_\lambda h_\lambda(t). \qquad \text{(Eq. 1)}$$

Various dictionaries have been used for audio signals so far, including harmonic atoms [11], local cosine and wavelet bases [3], and *chirped Gabor* atoms [12]. In the context of musical instrument recognition, it is possible to choose, as mentioned above, specific harmonic atoms that carry instrument specific information.

Decomposition in harmonic atoms

More precisely, such a harmonic atom is defined as a sum of M windowed sinusoidal partials at harmonic frequencies with constant amplitudes but linearly varying fundamental frequency. Using chirped Gabor atoms to represent the partials, each harmonic atom is in turn expressed as

$$h_{s,u,f_0,c_0,A,\Phi}(t) = \sum_{m=1}^{M} a_m \, e^{j\phi_m} g_{s,u,m\times f_0,m\times c_0}(t) \qquad \text{(Eq. 2)}$$

where $g_{s,u,f,c}(t) = \frac{1}{\sqrt{s}}\, w\left(\frac{t-u}{s}\right) e^{2j\pi\left(f(t-u)+\frac{c}{2}(t-u)^2\right)}$ represent the chirped Gabor atoms and where s is the scale parameter, u the time localization, f_0 the fundamental frequency, c_0 the fundamental chirp rate, $A = \{a_m\}_{m=1\ldots M}$ the vector of partials amplitudes and $\Phi = \{\phi_m\}_{m=1\ldots M}$ the vector of partials phases.

The most distinctive feature of this signal model is that the vector of partials amplitudes A is learnt on solo training data, so as to represent a single instrument i among the multiple instruments possibly present in the polyphonic signal. More precisely, the frequency range is partitioned into several *pitch classes* p and each vector A is associated with a single instrument/pitch class $\mathcal{C}_{ip}$. Note that, each instrument/pitch class $\mathcal{C}_{ip}$ may be represented by several amplitude vectors denoted by A_{ipk}, with $k = 1 \ldots K_{ip}$. This allows to better cope with the intrinsic variability of sounds produced by each instrument.

The atoms are iteratively extracted following a Matching Pursuit technique. First, the most prominent atom (*i.e.* the most correlated with the signal) is extracted and subtracted from the original signal and the procedure is iterated until a predefined number of atoms have been extracted or until a pre-defined Signal to Noise Ratio of the representation is reached. To increase the quality of the representation without enlarging too much the size of the dictionary, it is possible to adaptively tune some of the atom parameters. For example, in [21], it is proposed to locally optimize the fundamental frequency f_0 and fundamental chirp rate c_0 by a conjugate gradient technique.

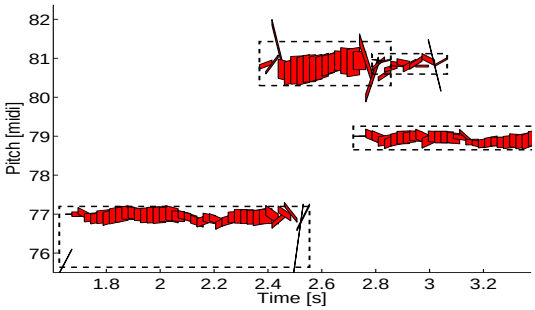

Figure 2: Representation of a music signal involving two instruments as a collection of harmonic molecules. Each atom is represented by a parallelogram, and each molecule by a dashed-line rectangle covering several atoms.

Decomposition in molecules

It is however important to note that the instrument-specific harmonic atoms are time-localized and therefore do not capture long-term temporal content. This is the main motivation for the development of a structured representation where the signal is decomposed on a set of instrument-specific harmonic molecules, where a molecule $\mathcal{M}$ is a group of instrument-specific harmonic atoms $h_\lambda(t)$. Figure 2 displays an example of molecules for a music signal involving two instruments.

It is shown in [21] that this approach is efficient for musical instruments identification in solos and obtains very promising results in polyphonic music involving 5 or less instruments. The average recognition rate on solos ranged from 56% to 86% depending on the instrument, which is nearly as efficient as more traditional and heavily studied bag-of-frames approaches. For duos, the preliminary results obtained show the potential of this approach for polyphonic signals (the correct pair of instruments is recognized with a rate between 23% and 69% depending of the pair and it is observed that at least one of the instrument is correctly recognized with a rate of about 97 %). It is also important to underline the potential of this alternative approach for a large range of applications including audio signal modification, wav-to-midi transcription and music visualization.

CONCLUSIONS

We provided in this paper a brief overview of the domain of music instrument recognition and presented a promising alternative approach based on the decomposition of the audio signal in harmonic atoms or molecules. If the mainstream approaches (e.g. “bag-of frames” approaches) are powerful and allow for reaching high recognition performances in solos, it is important to investigate new directions to solve the more complex, but more useful, situation where several instruments are played simultaneously (polyphony).

Acknowledgments

This research was supported by the European Commission under contract *FP6-027026-K-SPACE* and by the National Project *ANR-Musicdiscover*.

References

[1] G. Agostini, M. Longari, E. Pollastri. Musical instrument timbres classification with spectral features. In *Int. Workshop on Multimedia Sig. Proc.*. Cannes, France (2001)

[2] J. C. Brown, O. Houix, S. McAdams. Feature dependence in the automatic identification of musical woodwind instruments. Journal of the Acous. Soc. of Am. **109 (2000)** 1064–1072

[3] L. Daudet. Sparse and structured decompositions of signals with the molecular matching pursuit. IEEE Trans. on Audio, Speech and Language Proc. **(2006)** 1808–1816

[4] J. Eggink, G. J. Brown. Application of missing feature theory to the recognition of musical instruments in polyphonic audio. In *Proc. of Int. Conf. on Music Inf. Ret. (ISMIR)* (2003)

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

[6] D. Ellis, G. Poliner. Classification-based melody transcription. Machine Learning **65 (2006), no. 2** 439–456

[7] A. Eronen. Comparison of features for musical instrument recognition. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (2001) New Paltz, New York

[8] S. Essid, G. Richard, B. David. Instrument recognition in polyphonic music based on automatic taxonomies. IEEE Trans. on Audio, Speech and Language Proc. **14 (2006), no. 1** 68–80

[9] S. Essid, G. Richard, B. David. Musical instrument recognition by pairwise classification strategies. IEEE Trans. on Audio, Speech and Language Proc. **(2006)**

[10] O. Gillet, G. Richard. Transcription and separation of drum signals from polyphonic music. accepted in IEEE Trans. on Audio, Speech and Language Proc. **(2007)**

[11] R. Gribonval, E. Bacry. Harmonic decomposition of audio signals with matching pursuit. IEEE Trans. on Sig. Proc. **51 (2003), no. 1** 101–111

[12] R. Gribonval. Fast matching pursuit with a multiscale dictionary of gaussian chirps. IEEE Trans. on Sig. Proc. **49 (2001), no. 5** 994–1001

[13] P. Herrera, A. Klapuri, M. Davy. *Chap.6 Automatic Classification of Pitched Musical Instrument Sounds*. in Signal Processing methods for Music transcription, Edited by A. Klapuri and M. Davy, Springer (2006)

[14] ISO/IEC 15938-4:2002. Information technology – Multimedia content description interface – Part 4: Audio (2002)

[15] P. Jinachitra. Polyphonic instrument identification using independent subspace analysis. In *Proc. of Int. Conf. on Multimedia and Expo (ICME)* (2004)

[16] I. Kaminskyj, A. Materka. Automatic source identification of monophonic musical instrument sounds. In *IEEE International Conference on Neural Networks* (1995)

[17] T. Kitahara, M. Goto, K. Komatani, T. Ogata, H. Okuno. Instrument identification in polyphonic music: feature weighting with mixed sounds, pitch-dependent timber modeling, and use of musical context. In *Proc. of Int. Conf. on Music Inf. Ret. (ISMIR)* (2005)

[18] T. Kitahara, M. Goto, H. G. Okuno. Musical instrument identification based on f0-dependent multivariate normal distribution. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Hong Kong (2003)

[19] B. Kostek, A. Czyzewski. Automatic recognition of musical instrument sounds - further developments. In *110th AES convention*. The Netherlands (2001)

[20] P. Leveau, E. Vincent, G. Richard, 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*. Athènes, Grèce (2006)

[21] P. Leveau, E. Vincent, G. Richard, L. Daudet. Instrument-specific harmonic atoms for mid-level music representation. accepted in IEEE Trans. on Audio, Speech and Language Proc.

[22] A. Livshin, X. Rodet. Musical instrument identification in continuous recordings. In *7th International Conference on Digital Audio Effects (DAFX-4)*. Naples, Italy (2004)

[23] J. Marques, P. J. Moreno. A study of musical instrument classification using gaussian mixture models and support vector machines. Technical report, Compaq Comp. Corp. (1999)

[24] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, IRCAM (2004)

[25] E. Vincent. Musical source separation using time-frequency source priors. IEEE Trans. on Audio, Speech and Language Proc. **14 (2006), no. 1** 91–98