Real-time Transcription of Music Signals

Arshia Cont

REAL-TIME TRANSCRIPTION OF MUSIC SIGNALS: MIREX2007 SUBMISSION DESCRIPTION

Arshia Cont
University of California in San Diego, CA. And
Ircam-Centre Pompidou, Paris, France.
http://cosmal.ucsd.edu/arshia/

ABSTRACT

This paper briefly describes the underlying methodology beneath our submissions to the first Multi-F0 Estimation and Tracking Evaluation Task at MIREX 2007 [1] . The two systems described here are based on simple facts about music pitch structures that is briefly described in this abstract and share variations of an on-line machine learning approach presented previously in [1]. The systems were designed as a compromise between the speed of performance and precision, and achieves real-time performance and compete with systems very far from on-line considerations.

1 INTRODUCTION

The task of estimating multiple fundamental frequencies of audio and speech signals has attained substantial effort from the research community in the recent years. More interestingly, proposed algorithms in the literature undergo a wide variety of methods spanning from pure signal processing models to machine learning methods. For an excellent overview of different methods for multiple-f_0 estimation, we refer the curious reader to [2].

The submitted algorithm is quite different from others in the literature both in its purpose and approach. It is destined not for continuous multiple f_0 recognition but rather for projection of the ongoing spectrum to learned pitch templates. The decomposition algorithm on the other hand, does not compromise signal processing models for pitches and consists of an algorithm for efficient decomposition of a spectrum using known pitch structures and based on sparse non-negative constraints.

An important motivation behind the submitted algorithm is the simple intuition that humans tend to use a reconstructive scheme during detection of multiple pitches or multiple instruments and based on their history of timbral familiarity and music education. That is to say, in music dictation practices, well-trained musicians tend to transcribe music by conscious (or unconscious) addition of familiar pitches produced by musical instruments. The main idea here is that during detection of musical pitches and instruments, there is no direct assumption of *independence* associated with familiar patterns used for reconstruction and we rely more on *reconstruction* using superpositions.

Considering these facts, we can generally formulate our problem by *non-negative* factors. Non-negativity in this case simply means that we do not *subtract* pitch patterns in order to determine the correct combination but rather, we somehow manage to directly point to the correct combination of patterns that reconstruct the target by simple linear superposition. Mathematically speaking, given V as a non-negative representational scheme of the realtime audio signal in $\mathbb{R}_+^N$, we would like to achieve

$$V \approx WH \tag{1}$$

where W is a non-negative $\mathbb{R}_+^{N\times r}$ matrix holding r templates corresponding to objects to be detected and H is a simple non-negative $r \times 1$ vector holding the contribution of each template in W for reconstructing V. During realtime detection, we are already in possession of W and we tend to obtain H indicating the presence of each template in the audio buffer that is arriving online to the system in V. Given this formulation, there are three main issues to be addressed :

1. What is an efficient representation for V ?
2. How to learn templates in W using V ?
3. How to obtain acceptable results in H in realtime ?

In the framework of MIREX 2007 Multiple-f_0 estimation and tracking, the answer to the *estimation* task simply lies in the decoded H vector for each analysis frame of real-time audio. For polyphonic pitch *tracking*, a parallel tracking module is added on top of H vectors through time to grab the onset and offset times of each event. The estimation part is an enhancement of a previous development fully described in [1]. In this abstract we define the underlying principles of the algorithm and invite the curious reader to the follow the mentioned paper, [3] and further publications.

[1] http://www.music-ir.org/mirex2007/index.php/Multiple_Fundamental_Frequency_Estimation_&_Tracking

2 GENERAL ARCHITECTURE

The proposed method relies on unsupervised learning algorithms that are used for knowledge representation and discovery. During realtime observation, the algorithm tries to reconstruct the ongoing audio using previously learned pitch structures of an instrument, as a linear combination with non-negative weights. This implies an offline learning of pitch structures of all the pitches of an instrument which will be used as templates during learning.

The real-time estimation algorithm features a novel machine learning procedure based on *Sparse Non-Negative Constraints*. This overall architecture is similar to the system proposed in [4] with a crucial difference for music signals. Instead of using a regular Non-Negative Matrix Factorization (NMF) [5] algorithm for real-time determination of pitch, we use a modified NMF algorithm with sparseness constraints as outlined casually in this abstract.

2.1 Representational Front-end

The *additive* characteristic of NMF is an essential factor for any kind of representation used for V which, in the case of multiple pitch observation, implies that the spectral representation used for V should demonstrate a harmonic stack of pitch templates added together for a given chord.

The signal processing front end used for this observation is the result of a fixed point analysis of frequency to instantaneous frequency mapping of the ongoing audio spectrum [6]. The short-time Fourier transform (STFT) is an efficient tool for instantaneous frequency (IF) estimation [7]. As a result, vector V would be non-negative amplitudes of the fixed-point instantaneous frequency representing harmonic stacks at each analysis frame with the rest of the spectrum zeroed out.

2.2 Learning Pitch Templates

The system knows the pitch structures of all pitches of an instrument for use during realtime observation. Here we briefly mention how we learn different pitch templates for an instrument. As a reminder, W contains pitch structures of all pitches of a given instrument. For example, for an acoustic piano, matrix W would contain all 88 pitches as 88 different columns. To this end, training is done using databases of instrumental sounds (e.g. [8]) and an off-line training learns different pitch structures of an instrument by browsing all sounds produced by the given instrument in the database and stores them in matrix W for future use.

For each audio file in the database, training is an iterative NMF algorithm with a symmetric kullback-leibler divergence for reconstruction error. The regular NMF learning is moreover enhanced by a harmonic constraint at each iteration, enforcing learning of harmonic templates of each given note in the database.

2.3 Sparsity of the Solution

Despite perceptual advantages of an NMF approach over ICA algorithms for multiple-pitch detection, since pitch templates are not mathematically independent, for a given spectrum (in V) there may exist many possible solutions (H) using templates in W. More specifically for our problem, a given piano chord can be reconstructed by the templates of its original pitches as well as octaves, dominant and other pitches with harmonic relations to the original ones.

To overcome this problem, we use the strong assumption that the correct solution for a given spectrum (in V) uses a minimum of templates in W, or in other words, the solution has the minimum number of non-zero elements in H. This assumption is hard to be proofed for every music instrument and highly depends on the template presentations in W, but is easily imaginable as harmonic structure of a music note can be minimally expressed (in the mean squared sense) using the original note than a combination of its octaves and dominant.

Fortunately, this assumption has been heavily studied in the field of *sparse coding*. The concept of sparse coding refers to a representational scheme where only a few units out of a large population are effectively used to represent typical data vectors.

These concerns led us to a modified non-negative decomposition algorithm with sparsity controls. In the submitted system, sparsity is assured by a mixture of two commonly used norms in the sparse coding literature : ℓ_ϵ (approximated here by a $\tanh$ function) and ℓ_2 norms. The overall algorithm is a gradient-descent update where at each iteration the result is being optimized by projections to the intersection of two $\tanh$ and ℓ_2 hyper-planes. In this application, ℓ_ϵ is controlled by the user and ℓ_2 is provided by the signal's on-going average spectral power that assure time-continuity of the results. This outlines our first submission. The second submission is basically the same algorithm, but the ℓ_ϵ norm is replaced with a combination of ℓ_1 and ℓ_ϵ to relax the constraints on certain iterations.

3 CONCLUSION

In this abstract, we briefly described the underlying concept to our submissions to the first Multi-f_0 estimation and tracking evaluation contest. The proposed method was designed to meet the *real-time* constraint usually met in computer music and MIR tasks. Therefore the proposed system is a compromise between speed of computation and precision and would be competing with systems much further than real-time performance. The training of our system is done over real recordings of music, and thus we also expect worse performance on synthesized scores in general.

The system is currently being released for computer music real-time programming environments such as `Pure`

Data [2] and MaxMSP [3]. Progress can be checked at the author's webpage.

For further information regarding details, future developments, and extension of the presented algorithms, we refer the curious reader to [1, 3] and future publications.

4 REFERENCES

[1] Arshia Cont. Realtime multiple pitch observation using sparse non-negative constraints. In *International Symposium on Music Information Retrieval (ISMIR)*. Victoria, CA., October 2006.

[2] A. de Cheveigné. Multiple f0 estimation. In D.-L. Wang and G.J. Brown, editors, *Computational Auditory Scene Analysis : Principles, Algorithms and Applications*, pages 45–72. IEEE Press / Wiley, 2006.

[3] Arshia Cont, Shlomo Dubnov, and David Wessel. Realtime multiple-pitch and multiple-instrument recognition for music signals using sparse non-negative constraints. In *Proceedings of Digital Audio Effects Conference (DAFx)*. Bordeaux, October 2007.

[4] Fei Sha and Lawrence Saul. Real-time pitch determination of one or more voices by nonnegative matrix factorization. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA, 2005.

[5] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press, 2001.

[6] Hideki Kawahara, H. Katayose, Alain de Cheveigné, and R.D. Patterson. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity. In *Eurospeech*, volume 6, pages 2781–2784, 1999.

[7] Toshihiko Abe, Takao Kobayashi, and Satoshi Imai. Harmonic tracking and pitch extraction based on instantaneous frequency. In *IEEE ICASSP*, pages 756–759. Tokyo, 1995.

[8] Guillaume Ballet, Riccardo Borghesi, Peter Hoffmann, and Fabien Lévy. Studio online 3.0 : An internet "killer application" for remote access to ircam sounds and processing tools. In *Journée d'Informatique Musicale (JIM)*, paris, 1999.

[2] http://crca.ucsd.edu/~msp/software.html

[3] http://www.cycling75.com/