Engineering Applications of Artificial Intelligence 26 (2013) 1671–1680

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Constrained non-negative sparse coding using learnt instrument templates for realtime music transcription ☆

J.J. Carabias-Orti *, F.J. Rodriguez-Serrano, P. Vera-Candeas, F.J. Cañadas-Quesada, N. Ruiz-Reyes

Telecommunication Engineering Department, University of Jaén, Alfonso X el Sabio, 28, 23700 Linares (Jaén), Spain

ARTICLE INFO

Article history:
Received 4 June 2012
Received in revised form
18 March 2013
Accepted 25 March 2013
Available online 16 April 2013

Keywords:
Non-negative sparse coding (NNSC)
Non-negative matrix factorization (NMF)
Beta-divergence
Supervised learning
Instrument spectral patterns
Realtime music transcription

ABSTRACT

In this paper, we present a realtime signal decomposition method with single-pitch and harmonicity constrains using instrument specific information. Although the proposed method is designed for monophonic music transcription, it can be used as a candidate selection technique in combination with other realtime transcription methods to address polyphonic signals. The harmonicity constraint is particularly beneficial for automatic transcription because, in this way, each basis can define a single pitch. Furthermore, restricting the model to have a single-nonzero gain at each frame has been shown to be a very suitable constraint when dealing with monophonic signals. In our method, both harmonicity and single-nonzero gain constraints are enforced in a deterministic manner. A realtime factorization procedure based on Non-negative sparse coding (NNSC) with Beta-divergence and fixed basis functions is proposed. In this paper, the basis functions are learned using a supervised process to obtain spectral patterns for different musical instruments. The proposed method has been tested for music transcription of both monophonic and polyphonic signals and has been compared with other state-of-the-art transcription methods, and in these tests, the proposed method has obtained satisfactory results in terms of accuracy and runtime.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

An audio spectrogram can be decomposed as a linear combination of spectral basis functions. In such a model, the short-term magnitude (or power) spectrum of the signal $x(f,t)$ in frame t and frequency f is modeled as a weighted sum of the basis functions as

$$\hat{x}(f,t) = \sum_{n=1}^{N} g_n(t) b_n(f) \quad (1)$$

where $g_n(t)$ is the gain of the basis function n at frame t, and $b_n(f)$, $n=1,\ldots,N$ are the bases. When dealing with musical instrument sounds in the context of automatic music transcription, ideally, each basis function represents a single pitch, and the corresponding gains contain information about the onset and offset times of notes having that pitch.

There are several methods in the literature that are used to estimate this type of decomposition, such as independent component analysis (ICA) (Plumbley, 2003), sparse coding (Abdallah and Plumbley, 2004), atomic decompositions (Gribonval and Bacry, 2003) and non-negative matrix factorization (NMF) (Lee and Seung, 1999).

Sparse representations have received increasing attention for audio applications such as polyphonic audio transcription (Abdallah and Plumbley, 2004, 2006; Cont, 2006), speech recognition (Gemmeke et al., 2011; Hurmalainen et al., 2011) and sound source separation (Virtanen, 2007; Ozerov and Févotte, 2010).

In sparse coding, the goal is to find a decomposition in which the gains are sparse. Specifically, the probability densities have peaks at zero and have heavy tails (Olshausen and Field, 1997), or in other words, most observations can be encoded with only a few significant nonzero gain values. This assumption fits well with the notion that, in music, only a relatively small fraction of the available notes will be sounded at each frame (Abdallah and Plumbley, 2004).

When dealing with power or magnitude spectrograms, the non-negativity of the parameters is a natural restriction. As a result, it is possible to use the projected steepest descent algorithm, or it is possible to combine NMF and sparse coding, which leads to non-negative sparse coding (NNSC) (Abdallah and Plumbley, 2004; Hoyer, 2004).

Depending on the learning process, the basis can be fixed or adaptive. In a fixed basis, the basis is learned by training the system on isolated notes, while in an adaptive basis, the basis is

* Corresponding author. Tel.: +34 953648581; fax: +34 953648508.
E-mail addresses: carabias@ujaen.es, carabiasjulio@gmail.com (J.J. Carabias-Orti).

☆Supported by the Andalusian Business, Science and Innovation Council under project P10-TIC-6762, (FEDER) the Spanish Ministry of Science and Innovation under Project TEC2009-14414-C03-02, and the University of Jaen under Project R1/12/2010/64.

0952-1976/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.engappai.2013.03.010

learned directly from the signal to be analyzed. As demonstrated in Carabias-Orti et al. (2011), the fixed basis spectra have proven to provide a good generalization of the model parameters when the music scene in the training and test signals do not differ too much.

In this paper, we propose a signal decomposition method that can be used for monophonic music transcription and as a candidate selection technique in polyphonic transcriptors. This approach is based on NNSC which is constrained to explicitly assume the signal to be monophonic with only one possible state (note) at each frame. This extreme sparsity constraint has been used before by other signal decomposition methods within a probabilistic framework. For example, Benaroya et al. (2006) proposed a method for sound source separation in which each source Short Time Fourier Transform (STFT) is modeled by a Gaussian Mixture Model (GMM) modulated by a frame-dependent amplitude parameter accounting for nonstationarity, which leads to the Gaussian Scaled Mixture Model (GSMM) where the source is implicitly assumed to be monophonic with many possible states. This method has also been used in Durrieu et al. (2010) for main melody tracking in polyphonic audio signals. Ozerov et al. (2009) proposed a method called the Factorial Scaled Hidden Markov Model (FS-HMM) that generalized GSMM and NMF with Itakura Saito divergence (IS-NMF) and incorporates temporal continuity using Markov modeling.

Conversely, we propose a novel method that enforces single-pitch and harmonicity constrains in a deterministic manner, performs the decomposition based on NNSC with Beta-divergence (Févotte et al., 2009), and uses instrument specific information, which is learned in a supervised way (i.e. using a training stage). For the testing stage, the simplicity of the method allows for the direct computation of the factorization, which leads to very efficient runtimes in comparison with other signal decomposition methods found in the literature, such as GSMM or FS-HMM. In fact, the runtimes obtained and the possibility to analyze each frame independently makes our method more suitable for realtime applications where very low latency is required. Moreover, we propose to use our method as a candidate selection stage in combination with a realtime signal decomposition method to address polyphonic music transcription, which is a more complex scenario. To evaluate the reliability of our method, we have applied it to the music transcription of monaural monophonic and polyphonic signals, and we have obtained satisfactory results in comparison with other signal decomposition methods and selected state-of-the-art transcription methods.

The structure of the rest of the paper is as follows. In Section 2, we review the harmonicity and sparsity constrained signal decomposition methods proposed in previous studies. In Section 3, a novel theoretical approach to constrain a signal model to be harmonic and have a single nonzero gain is explained, and the algorithm to perform the decomposition for music transcription is detailed. In Section 4, we propose a realtime polyphonic transcription system in which the novel method explained in the previous section is used as a candidate selection technique. In Section 5, the proposed method is applied to music transcription of monophonic and polyphonic signals and is compared with other state-of-the-art transcription methods. Finally, we summarize the work and discuss future perspectives in Section 6.

2. Theoretical background

2.1. Basic Harmonic Constrained (BHC) method

As automatic music transcription is the application of the method used in this work, this method is constrained to be harmonic. This restriction has been used in other works devoted to music transcription (Carabias-Orti et al., 2011; Bertin et al., 2010; Vincent et al., 2010; Raczyński et al., 2007). The harmonicity constraint is particularly beneficial for music transcription because in this way each basis can define a single fundamental frequency. This constraint is introduced in the model presented in Eq. (1) requiring that a distinct basis function represents each note of the instrument.

$$b_n(f) = \sum_{m=1}^{M} a_n[m]G(f - mf_0(n)) \tag{2}$$

where $b_n(f)$, $n = 1, \ldots, N$ are the bases for each note n, $m = 1, \ldots, M$ is the number of harmonics, $a_n[m]$ is the amplitude of harmonic m for note n, $f_0(n)$ is the fundamental frequency of note n, $G(f)$ is the magnitude spectrum of the window function, and the spectrum of a harmonic component at frequency $mf_0(n)$ is approximated by $G(f - mf_0(n))$.

The model for the magnitude spectra of a music signal is then obtained as

$$\hat{x}(f,t) = \sum_{n=1}^{N}\sum_{m=1}^{M} g_n(t)a_n[m]G(f - mf_0(n)) \tag{3}$$

where the time gains $g_n(t)$ and the harmonic amplitudes $a_n[m]$ are the parameters of the method to be estimated.

To obtain the factorization of Eq. (3), the reconstruction error between the observed spectrogram $x(f,t)$ and the modeled spectrogram $\hat{x}(f,t)$ is minimized. In several recent works (Vincent et al., 2010; Févotte et al., 2009; Févotte and Idier, 2011), the cost function to be minimized is the Beta-divergence,

$$D_\beta(x(f,t)|\hat{x}(f,t)) = \begin{cases} \sum_{f,t}\frac{1}{\beta(\beta-1)}(x(f,t)^\beta + (\beta-1)\hat{x}(f,t)^\beta - \beta x(f,t)\hat{x}(f,t)^{\beta-1}) & \beta\in(0,1)\cup(1,2] \\ \sum_{f,t} x(f,t)\log\frac{x(f,t)}{\hat{x}(f,t)} - x(f,t) + \hat{x}(f,t) & \beta = 1 \\ \sum_{f,t}\frac{x(f,t)}{\hat{x}(f,t)} + \log\frac{x(f,t)}{\hat{x}(f,t)} - 1 & \beta = 0 \end{cases} \tag{4}$$

The Beta-divergence includes in its definition the most popular cost functions. When $\beta = 2$, the Beta-divergence is equivalent to the Euclidean (EUC) distance. The Kullback–Leibler (KL) divergence is obtained when $\beta = 1$, and the Itakura–Saito (IS) divergence is computed when $\beta = 0$.

2.2. BHC with Sparse Constraint (BHC-SC) method

Sparsity is a natural restriction applied to the gains that enforces the signal model to have only a few nonzero gains $g_n(t)$ at each frame t. This assumption fits well with the concept that, in music, only a relatively small fraction of the available notes will be sounded at each frame. In the special case of monophonic transcription, there should be only a single nonzero gain at each frame. Other works devoted to transcription and source separation have used sparseness in their signal models (Abdallah and Plumbley, 2004; Gemmeke et al., 2011; Virtanen, 2007; Hoyer, 2004).

For signal models that minimize a divergence, the sparsity is typically introduced as a regularization penalty term (Gemmeke et al., 2011). This penalty term helps to discard those solutions in which most of their gains are set to nonzero values. The cost function is then defined as

$$D(x(f,t)|\hat{x}(f,t)) = D_\beta(x(f,t)|\hat{x}(f,t)) + \lambda\sum_{n,t}\phi(g_n(t)) \tag{5}$$

where ϕ is a function that penalizes nonzero gains and λ is a parameter that controls the importance of the regularized term. Although there are several definitions for the penalty terms in the literature, in the experimental setup we have used the L1 norm $\phi(x) = \|x\|_1$ as proposed in Virtanen (2007), Olshausen and Field (1997), Candés and Wakin (2008) because it has proven to be less sensitive to variations in parameter λ (Virtanen, 2007).

Download English Version:

https://daneshyari.com/en/article/380710

Download Persian Version:

https://daneshyari.com/article/380710

Daneshyari.com