Signal Processing 130 (2017) 105–117

Contents lists available at ScienceDirect

Signal Processing

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

Sparse modeling of chroma features ☆

Ted Kronvall *, Maria Juhlin, Johan Swärd, Stefan I. Adalbjörnsson *, Andreas Jakobsson

Mathematical Statistics, Centre for Mathematical Sciences, Lund University, P.O. Box 118, SE-221 00 Lund, Sweden

ARTICLE INFO

Article history:
Received 25 August 2015
Received in revised form
16 June 2016
Accepted 21 June 2016
Available online 24 June 2016

Keywords:
Chroma
Multi-pitch estimation
Sparse modeling
Amplitude modulation
Block sparsity
ADMM

ABSTRACT

This work treats the estimation of chroma features for harmonic audio signals using a sparse reconstruction framework. Chroma has been used for decades as a key tool in audio analysis, and is typically formed using a periodogram-based approach that maps the fundamental frequency of a musical tone to its corresponding chroma. Such an approach often leads to problems with tone ambiguity. We address this ambiguity via sparse modeling, allowing us to appropriately penalize ambiguous estimates while taking the harmonic structure of tonal audio into account. Furthermore, we also allow for signals to have time-varying envelopes. Using a spline-based amplitude modulation of the chroma dictionary, the presented estimator is able to model longer frames than what is conventional for audio, as well as to model highly time-localized signals, and signals containing sudden bursts, such as trumpet or trombone signals. Thus, we may retain more signal information as compared to alternative methods. The performances of the proposed methods are evaluated by analyzing the average estimation errors for synthetic signals, as compared to the Cramér–Rao lower bound, and by visual inspection for estimates of real instrument signals. The results show strong visual clarity, as compared to other commonly used methods.

1. Introduction

Music is an art-form that people have enjoyed for millennia. Perhaps music is even enjoyed more today, as the development of personalized computers and smart telephones have enabled ubiquitous music listening, automatic identification of songs, and even the chance for anyone to be a self-made DJ. When listening, learning, composing, mixing, and identifying music, there are a number of musical features one may utilize (see, e.g., [3]). One of the fundamental building blocks in music, the musical note, is a periodic sound, typically characterized by its pitch, timbre, intensity, and duration. For transcription purposes, i.e., to separate one tone from another, pitch serves as the common descriptor, and we will herein use pitch and tone interchangeably. From a conventional perspective, pitch is measured on an ordinal scale, at which a pitch is humanly perceived as either higher, lower, or the same as another pitch. However, from the perspective of scientific audio analysis, a pitch is quantified using an interval scale, in which its spectral distribution of energy is modeled. A single pitch may be seen as a superposition of several narrowband spectral peaks, which are approximately integer multiples of a fundamental frequency. Thus, we refer to the group of frequencies as the pitch, and to each frequency component as a partial harmonic, or, alternatively, just as a partial. As for the fundamental frequency, it is either the lowest partial, or, if that partial is missing, the smallest spectral distance between adjacent partials. The number of harmonics in a certain pitch, as well as the relative power between these, varies greatly over time and between different sounds. Identifying pitches in a way similar to our human perception has proved to be a difficult estimation problem. Partly, this difficulty is due to coinciding frequency components between certain pitches. For instance, two pitches, where one has exactly twice the fundamental frequency of the other, are referred to as being octave equivalent, as the relative distance by a factor of two is called an octave. These will typically share a large number of partials, often making an estimation procedure ambiguous between octaves. To further complicate matters, other pairs of pitches may also have many coinciding partials. These are typically found together in audio, an aspect which is referred to as harmony, since they are perceptually pleasant to hear [4]. Jointly estimating several pitches in a signal, i.e., multi-pitch estimation, has been thoroughly examined in the literature (see, e.g., [5–7], and the references therein). However, separating intricate combinations of frequency components into multiple pitches often proves difficult. Typically, issues arise when the complexity of the audio signal increases, such that there are simultaneously two or more pitches with overlapping spectral content present, for instance played by two or more

☆This work was supported in part by the Swedish Research Council, Carl Trygger's foundation, and the Royal Physiographic Society in Lund. This work has been published in part at the ICASSP 2015 [1] and EUSIPCO 2015 [2] conferences.

* Corresponding authors.
E-mail addresses: ted@maths.lth.se (T. Kronvall), juhlin@maths.lth.se (M. Juhlin), js@maths.lth.se (J. Swärd), sia@maths.lth.se (S.I. Adalbjörnsson), aj@maths.lth.se (A. Jakobsson).

http://dx.doi.org/10.1016/j.sigpro.2016.06.020
0165-1684/

instruments. In the Western musicological system, the frequency interval corresponding to an octave is discretized into 12 intervals, called semi-tones. By gathering all pitches with octave equivalence to their respective semi-tone, these form 12 groups of pitches, called chroma. As octave equivalent pitches share a large number of harmonics, the notion of chroma is thus a method for grouping together those pitches which are perceived as most similar. Therefore, chroma features are widely used in applications such as cover song detection, transcription, and recommender systems (see, e.g., [8–10]). Most methods for chroma estimation begin by obtaining estimates of the pitches in a signal, which are then mapped into their respective chroma. Some of these take the harmonic structure into account, and others do not. The commonly used method by Ellis [11] is formed via a time-smoothed version of the short-time Fourier transform (STFT), whereas the CP and CENS methods by Müller and Ewert [12] use a filterbank approach, and the method in [13] uses a sparse methodology. Neither of these take the harmonic structure of pitches into account. Other approaches instead allow for the harmonic structure, such as the method in [14], using a non-negative least squares approach, the method presented in [15], which uses a comb filtering technique, or the method in [16], in which post-processing on the periodogram is performed. Most existing methods have in common that their estimates are not directly formed from the actual data, but rather on a representation of these measurements, such as the STFT or the magnitude of the periodogram. Herein, we propose to estimate the chroma using a sparse model reconstruction framework in the time-domain, where explicit model orders are not required. The estimate is found as the solution to a convex optimization problem. The solution is obtained as a linear combination of an over-complete chroma-based set of Fourier basis functions. Overfitting is avoided by introducing convex penalties, thus promoting solutions having the sought chroma structure. The model orders are thus set implicitly, using tuning parameters, which may be obtained using cross-validation, or by utilizing some simple heuristics. In this paper, we generalize upon the work in [7], taking into account the chroma structure, as well as allowing the frequency components to have time-varying amplitudes. The proposed extension increases robustness, as it allows for highly non-stationary signals, or signals with sudden bursts, like trumpets, whose nature may easily be misinterpreted when using ordinary chroma selection techniques. As in [17], the extended model uses a spline basis to detail the time-varying envelope of the signal, thereby enabling the amplitudes to evolve smoothly with time. The theoretical performance of the proposed estimator is verified using synthetic signals, which are compared to the Cramér–Rao lower bound (CRLB) for the chroma signal model. The practical use of the proposed estimator is illustrated using some excerpts from a recorded trumpet signal, showing an increased visual performance, as compared to some typical reference methods.

2. The chroma signal model

A sound signal typically contains a wide band of frequencies. However, for tonal audio, it is well known that a predominant part of the spectral energy is confined to a small number of frequency locations. In this section, we will therefore describe a framework for quantifying a musical tone from these frequencies. From the harmonic model, we know that the ideal frequency locations are placed at integer multiples of the lowest partial, which is defined as the fundamental frequency. However, as many sound sources produce tones far from ideal, such as, e.g., missing partials, or only having partials at odd integer multiples, we will use a more rigorous definition as to avoid ambiguity. Thus, if we assume to have a tone whose frequency components are placed at some integer multiples of a frequency, we say that the fundamental frequency is the largest frequency possible to which the components can still be placed at integer multiples. Without this definition, we may take any fundamental frequency, divide it by two any number of times, and say that it is still the fundamental frequency, only that the partials' multiples are restricted to some high-numbered subset of the even number set. And not only is this a mathematical issue, it is also a practical estimation issue, in which the halving frequency may be chosen instead of the fundamental. To that end, we state the tone's signal model using its chroma, which collects all tones that are halvings or doublings of each other. In fact, we state two different signal models, one which assumes that the signal has frame-wise constant amplitudes, and one which has amplitudes well modeled using weights for a set of B-spline functions.

But first, in order to formalize the definition of a fundamental frequency, let $\psi(f, \ell)$ denote the function which describes the frequency of the ℓth component. If this function is known, the entire group of components, or partials, representing a musical tone may be described by their fundamental frequency, f. Many oscillating sources, such as, for instance, the human vocal tract and stringed, or wind, instruments, emit tonal audio where the partials are integer multiples of the fundamental, i.e.,

$$\psi(f, \ell) = f\ell, \quad \ell \in \mathcal{L} \subseteq \mathbb{N} \tag{1}$$

where $\mathcal{L}$ denotes the index set of partials present in the signal. However, for an arbitrary $\mathcal{L}$, the definition in (1) is not sufficient to uniquely describe a pitch, as the set of frequencies may map to infinitely many combinations of f and $\mathcal{L}$. For example, for any $n \in \mathbb{N}$, the two pitches

$$\boldsymbol{\psi} = \left\{ \psi(f, \ell) \colon f \in \mathbb{R}, \ell \in \mathcal{L} \subseteq \mathbb{N} \right\} \tag{2}$$

$$\boldsymbol{\psi}' = \left\{ \psi(f', \ell') \colon f' = \frac{f}{n}, \ell' \in \mathcal{L}' = \{n\ell \colon \ell \in \mathcal{L}\} \right\} \tag{3}$$

have identical frequency components. Therefore, some constraints need to be imposed on $\mathcal{L}$. A common assumption for pitches is spectral smoothness of the harmonics, i.e., that adjacent harmonics should be of comparable magnitude [18]. This implies that $\mathcal{L}$ typically has few missing harmonics, and that n is as small as possible. However, in some signals, the first harmonic might be missing, so rather than defining the pitch as the signal's smallest frequency component, we define the fundamental frequency more rigorously.

Definition 1 (*Fundamental frequency*). If the set of frequencies in a pitch may be described by (2), then for any $n \in \mathbb{Q}$, the fundamental frequency is the largest $f' = f/n$ which fulfill (3), i.e., which ensures that $\mathcal{L}' = \{n\ell, \ell \in \mathcal{L}\} \subseteq \mathbb{N}$.

The index set therefore plays a vital role in the definition of the pitch frequency. Furthermore, because of the harmonic structure, many different pitches will have coinciding partials. To illustrate this, consider two pitches

$$\boldsymbol{\psi} = \left\{ \psi(f, \ell) \colon f \in \mathbb{R}, \ell \in \{1, 2, \ldots, L\} \right\} \tag{4}$$

$$\boldsymbol{\psi}' = \left\{ \psi(f', \ell') \colon f' = \frac{f}{n}, \ell' \in \{1, 2, \ldots, nL\} \right\} \tag{5}$$

which consist of all harmonics from $\ell = 1$ up to L and nL, respectively. Here, n may be a rational number, as long as (5) is fulfilled. Indeed, both pitches are unique according to our definition. Still, they will share a large number of harmonics, in fact L of them, as $\boldsymbol{\psi}$ forms a perfect subset of $\boldsymbol{\psi}'$, i.e., $\boldsymbol{\psi} \in \boldsymbol{\psi}'$. As sounds, they