applied sciences

MDPI

Article

Full-Band Quasi-Harmonic Analysis and Synthesis of Musical Instrument Sounds with Adaptive Sinusoids

Marcelo Caetano [1,*], George P. Kafentzis [2], Athanasios Mouchtaris [2,3] and Yannis Stylianou [2]

[1] Sound and Music Computing Group, Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), 4200-465 Porto, Portugal

[2] Multimedia Informatics Lab, Department of Computer Science, University of Crete, 700-13 Heraklion, Greece; kafentz@csd.uoc.gr (G.P.K.); mouchtar@ics.forth.gr (A.M.); yannis@csd.uoc.gr (Y.S.)

[3] Signal Processing Laboratory, Institute of Computer Science, Foundation for Technology & Research-Hellas (FORTH), 700-13 Heraklion, Greece

* Correspondence: mcaetano@inesctec.pt; Tel.: +351-22-209-4217

Academic Editor: Vesa Valimaki
Received: 16 February 2016; Accepted: 19 April 2016; Published: 2 May 2016

Abstract: Sinusoids are widely used to represent the oscillatory modes of musical instrument sounds in both analysis and synthesis. However, musical instrument sounds feature transients and instrumental noise that are poorly modeled with quasi-stationary sinusoids, requiring spectral decomposition and further dedicated modeling. In this work, we propose a full-band representation that fits sinusoids across the entire spectrum. We use the extended adaptive Quasi-Harmonic Model (eaQHM) to iteratively estimate amplitude- and frequency-modulated (AM–FM) sinusoids able to capture challenging features such as sharp attacks, transients, and instrumental noise. We use the signal-to-reconstruction-error ratio (SRER) as the objective measure for the analysis and synthesis of 89 musical instrument sounds from different instrumental families. We compare against quasi-stationary sinusoids and exponentially damped sinusoids. First, we show that the SRER increases with adaptation in eaQHM. Then, we show that full-band modeling with eaQHM captures partials at the higher frequency end of the spectrum that are neglected by spectral decomposition. Finally, we demonstrate that a frame size equal to three periods of the fundamental frequency results in the highest SRER with AM–FM sinusoids from eaQHM. A listening test confirmed that the musical instrument sounds resynthesized from full-band analysis with eaQHM are virtually perceptually indistinguishable from the original recordings.

Keywords: musical instruments; analysis and synthesis; sinusoidal modeling; AM–FM sinusoids; adaptive modeling; nonstationary sinusoids; full-band modeling

PACS: 43.75.Zz; 43.75.De; 43.75.Ef; 43.75.Fg; 43.75.Gh; 43.75.Kk; 43.75.Mn; 43.75.Pq; 43.75.Qr

1. Introduction

Sinusoidal models are widely used in the analysis [1,2], synthesis [2,3], and transformation [4,5] of musical instrument sounds. The musical instrument sound is modeled by a waveform consisting of a sum of time-varying sinusoids parameterized by their amplitudes, frequencies, and phases [1–3]. Sinusoidal analysis consists of the estimation of parameters, synthesis comprises techniques to retrieve a waveform from the analysis parameters, and transformations are performed as changes of the parameter values. The time-varying sinusoids, called partials, represent how the oscillatory modes of the musical instrument change with time, resulting in a flexible representation with perceptually meaningful parameters. The parameters completely describe each partial, which can be manipulated independently.

Several important features can be directly estimated from the analysis parameters, such as fundamental frequency, spectral centroid, inharmonicity, spectral flux, onset asynchrony, among many others [2]. The model parameters can also be used in musical instrument classification, recognition, and identification [6], vibrato detection [7], onset detection [8], source separation [9], audio restoration [10], and audio coding [11]. Typical transformations are pitch shifting, time scaling [12], and musical instrument sound morphing [5]. Additionally, the parameters from sinusoidal models can be used to estimate alternative representations of musical instrument sounds, such as spectral envelopes [13] and the source-filter model [14,15].

The quality of the representation is critical and can impact the results for the above applications. In general, sinusoidal models render a close representation of musical instrument sounds because most pitched musical instruments are designed to present very clear modes of vibration [16]. However, sinusoidal models do not result in perfect reconstruction upon resynthesis, leaving a modeling residual that contains whatever was not captured by the sinusoids [17]. Musical instrument sounds have particularly challenging features to represent with sinusoids, such as sharp attacks, transients, inharmonicity, and instrumental noise [16]. Percussive sounds produced by plucking strings (such as harpsichords, harps, and the *pizzicato* playing technique) or striking percussion instruments (such as drums, idiophones, or the piano) feature sharp onsets with highly nonstationary oscillations that die out very quickly, called transients [18]. Flute sounds characteristically comprise partials on top of breathing noise [16]. The reed in woodwind instruments presents a highly nonlinear behavior that also results in attack transients [19], while the stiffness of piano strings results in a slightly inharmonic spectrum [18]. The residual from most sinusoidal representations of musical instrument sounds contains perceptually important information [17]. However, the extent of this information ultimately depends on what the sinusoids are able to capture.

The standard sinusoidal model (SM) [1,20] was developed as a parametric extension of the short-time Fourier transform (STFT) so both analysis and synthesis present the same time-frequency limitations as the Discrete Fourier Transform (DFT) [21]. The parameters are estimated with well-known techniques, such as peak-picking and parabolic interpolation [20,22], and then connected across overlapping frames (partial tracking [23]). Peak-picking is known to bias the estimation of parameters because errors in the estimation of frequencies can bias the estimation of amplitudes [22,24]. Additionally, the inherent time-frequency uncertainty of the DFT further limits the estimation because long analysis windows blur the temporal resolution to improve the frequency resolution and *vice-versa* [21]. The SM uses quasi-stationary sinusoids (QSS) under the assuption that the partials are relatively stable inside each frame. QSS can accurately capture the lower frequencies because these have fewer periods inside each frame and thus less temporal variation. However, higher frequencies have more periods inside each frame with potentially more temporal variation lost by QSS. Additionally, the parameters of QSS are estimated using the center of the frame as the reference and the values are less accurate towards the edges because the DFT has a stationary basis [25]. This results in the loss of sharpness of attack known as pre-echo.

The lack of transients and noise is perceptually noticeable in musical instrument sounds represented with QSS [17,26]. Serra and Smith [1] proposed to decompose the musical instrument sound into a sinusoidal component represented with QSS and a residual component obtained by subtraction of the sinusoidal component from the original recording. This residual is assumed to be noise not captured by the sinusoids and commonly modeled by filtering white noise with a time-varying filter that emulates the spectral characteristics of the residual component [1,17]. However, the residual contains both errors in parameter estimation and transients plus noise missed by the QSS [27].

The time-frequency resolution trade-off imposes severe limits on the detection of transients with the DFT. Transients are essentially localized in time and usually require shorter frames which blur the peaks in the spectrum. Daudet [28] reviews several techniques to detect and extract transients with sinusoidal models. Multi-resolution techniques [29,30] use multiple frame sizes to circumvent the time-frequency uncertainty and to detect modulations at different time scales. Transient modeling synthesis (TMS) [26,27,31] decomposes sounds into sinusoids plus transients plus noise and models each separately. TMS performs sinusoidal plus residual decomposition with QSS and then extracts the transients from the residual.

An alternative to multiresolution techniques is the use of high-resolution techniques based on total least squares [32] such as ESPRIT [33], MUSIC [34], and RELAX [35] to fit exponentially damped sinusoids (EDS). EDS are widely used to represent musical instrument sounds [11,36,37]. EDS are sinusoids with stationary (*i.e.*, constant) frequencies modulated in amplitude by an exponential function. The exponentially decaying amplitude envelope from EDS is considered suitable to represent percussive sounds when the beginning of the frame is synchronized with the onsets [38]. However, EDS requires additional partials when there is no synchronization, which increases the complexity of the representation. ESPRIT decomposes the signal space into sinusoidal and residual, further ranking the sinusoids by decreasing magnitude of eigenvalue (*i.e.*, spectral energy). Therefore, the first K sinusoids maximize the energy upon resynthesis regardless of their frequencies.

Both the SM and EDS rely on sinusoids with stationary frequencies, which are not appropriate to represent nonstationary oscillations [21]. Time-frequency reassignment [39–41] was developed to estimate nonstationary sinusoids. Polynomial phase signals [20,25] such as splines [21] are commonly used as an alternative to stationary sinusoids. McAulay and Quatieri [20] were among the first to interpolate the phase values estimated at the center of the analysis window across frames with cubic polynomials to obtain nonstationary sinusoids inside each frame. Girin *et al.* [42] investigated the impact of the order of the polynomial used to represent the phase and concluded that order five does not improve the modeling performance sufficiently to justify the increased complexity. However, even nonstationary sinusoids leave a residual with perceptually important information that requires further modeling [25].

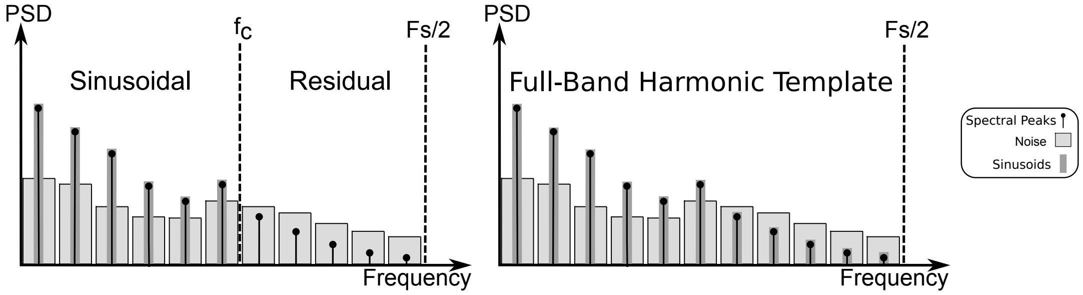
Sinusoidal models rely on spectral decomposition assuming that the lower end of the spectrum can be modeled with sinusoids while the higher end essentially contains noise. The estimation of the separation between the sinusoidal and residual components has proved difficult [27]. Ultimately, spectral decomposition misses partials on the higher end of the spectrum because the separation is artificial, depending on the spectrum estimation method rather than the spectral characteristics of musical instrument sounds. We consider spectral decomposition to be a consequence of artifacts from previous sinusoidal models instead of an acoustic property of musical instruments. Therefore, we propose the full-band modeling of musical instrument sounds with adaptive sinusoids as an alternative to spectral decomposition.

Adaptive sinusoids (AS) are nonstationary sinusoids estimated to fit the signal being analyzed usually via an iterative parameter re-estimation process. AS have been used to model speech [43–46] and musical instrument sounds [25,47]. Pantazis [45,48] developed the adaptive Quasi-Harmonic Model (aQHM), which iteratively adapts the frequency trajectories of all sinusoids at the same time based on the Quasi-Harmonic Model (QHM). Adaptation improves the fit of a spectral template via an iterative least-squares (LS) parameter estimation followed by frequency correction. Later, Kafentzis [43] devised the extended adaptive Quasi-Harmonic Model (eaQHM), capable of adapting both amplitude and frequency trajectories of all sinusoids iteratively. In eaQHM, adaptation is equivalent to the iterative projection of the original waveform onto nonstationary basis functions that are locally adapted to the time-varying characteristics of the sound, capable of modeling sudden changes such as sharp attacks, transients, and instrumental noise. In a previous work [47], we showed that eaQHM is capable of retaining the sharpness of the attack of percussive sounds.

In this work, we propose full-band modeling with eaQHM for a high-quality analysis and synthesis of isolated musical instrument sounds with a single component. We compare our method to QSS estimated with the standard SM [20] and EDS estimated with ESPRIT [36]. In the next section, we discuss the differences in full-band spectral modeling and traditional decomposition for musical instrument sounds. Next, we describe the full-band quasi-harmonic adaptive sinusoidal modeling behind eaQHM. Then, we present the experimental setup, describe the musical instrument sound database used in this work and the analysis parameters. We proceed to the experiments, present the results, and evaluate the performance of QSS, EDS, and eaQHM in modeling musical instrument sounds. Finally, we discuss the results and present conclusions and perspectives for future work.

2. Full-Band Modeling

Spectrum decomposition splits the spectrum of musical instrument sounds into a sinusoidal component and a residual as illustrated in Figure 1a. Spectrum decomposition assumes that there are partials only up to a certain cutoff frequency f_c, above which there is only noise. Figure 1a represents the spectral peaks as spikes on top of colored noise (wide light grey frequency bands) and f_c as the separation between the sinusoidal and residual components. Therefore, f_c determines the number of sinusoids because only the peaks at the lower frequency end of the spectrum are represented with sinusoids (narrow dark grey bars) and the rest is considered wide-band and stochastic noise existing across the whole range of the spectrum. There is noise between the spectral peaks and at the higher end of the spectrum. In a previous study [17], we showed that the residual from the SM is perceptually different from filtered (colored) white noise. Figure 1a shows that there are spectral peaks left in the residual because the spectral peaks above f_c are buried under the estimation noise floor (and sidelobes). Consequently, the residual from sinusoidal models that rely on spectral decomposition such as the SM is perceptually different from filtered white noise.



(a) Spectrum Decomposition **(b)** Full-Band Harmonic Template

Figure 1. Illustration of the spectral decomposition and full-band modeling paradigms.

From an acoustic point of view, the physical behavior of musical instruments can be modeled as the interaction between an excitation and a resonator (the body of the instrument) [16]. This excitation is responsible for the oscillatory modes whose amplitudes are shaped by the frequency response of the resonator. The excitation signal commonly contains discontinuities, resulting in wide-band spectra. For instance, the vibration of the reed in woodwinds can be approximated by a square wave [49], the friction between the bow and the strings results in an excitation similar to a sawtooth wave [16], the strike in percussion instruments can be approximated by a pulse [2], while the vibration of the lips in brass instruments results in a sequence of pulses [50] (somewhat similar to the glottal excitation, which is also wide band [46]).

Figure 1b illustrates a full-band harmonic template spanning the entire frequency range, fitting sinusoids to spectral peaks in the vicinity of harmonics of the fundamental frequency f_0. The spectrum of musical instruments is known to present deviations from perfect harmonicity [16], but quasi-harmonicity is supported by previous studies [51] that found deviations as small as 1%. In this work, the full-band harmonic template becomes quasi-harmonic after the estimation of parameters via least-squares followed by a frequency correction mechanism (see details in Section 3.1). Therefore, full-band spectral modeling assumes that both the excitation and the instrumental noise are wide band.

3. Adaptive Sinusoidal Modeling with eaQHM

In what follows, $x(n)$ is the original sound waveform and $\hat{x}(n)$ is the sinusoidal model with sample index n. Then, the following relation holds:

$$x(n) = \hat{x}(n) + e(n), \tag{1}$$

where $e(n)$ is the modeling error or residual. Each frame of $x(n)$ is

$$x(n,m) = x(n)\,w(n - mH), \quad m = 0, \cdots, M-1, \tag{2}$$

where m is the frame number, M is the number of frames, and H is the hop size. The analysis window $w(n)$ has L samples and it defines the frame size. Typically, $H < L$ such that the frames m overlap.

Figure 2 presents an overview of the modeling steps in eaQHM. The feedback loop illustrates the adaptation cycle, where $\hat{x}(n)$ gets closer to $x(n)$ with each iteration. The iterative process stops when the fit improves by less than a threshold ε. The dark blocks represent parameter estimation based on the quasi-harmonic model (QHM), followed by interpolation of the parameters across frames before additive [1] resynthesis (instead of overlap add (OLA) [52]). The resulting time-varying sinusoids are used as nonstationary basis functions for the next iteration, so the adaptation procedure illustrated in Figure 3 iteratively projects $x(n)$ onto $\hat{x}(n)$. Next, QHM is summarized, followed by parameter interpolation and then eaQHM.

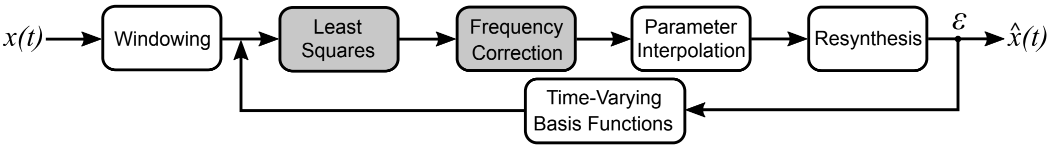


Figure 2. Block diagram depicting the modeling steps in the extended adaptive Quasi-Harmonic Model (eaQHM). The blocks with a dark background correspond to parameter estimation, while the feedback loop illustrates adaptation as iteration cycles around the loop. See text for the explanation of the symbols.

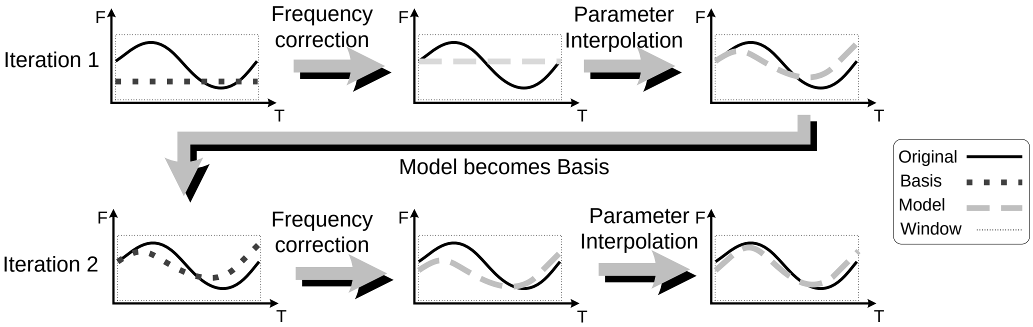


Figure 3. Illustration of the adaptation of the frequency trajectory of a sinusoidal partial inside the analysis window in eaQHM. The figure depicts the first and second iterations of eaQHM around the loop in Figure 2, showing local adaptation as the iterative projection of the original waveform onto the model.

3.1. The Quasi-Harmonic Model (QHM)

QHM [48] projects $x(n,m)$ onto a template of sinusoids $e^{j2\pi n\hat{f}_k/f_s}$ with constant frequencies $\hat{f}_k$ and sampling frequency f_s. QHM estimates the parameters of $\hat{x}(n,m)$ using

$$\hat{x}\left(n,m\right)=\sum_{k=-K}^{K}\left(\mathbf{a}_k+n\mathbf{b}_k\right)e^{j2\pi\hat{f}_k n/f_s},\tag{3}$$

where k is the partial number, K is the number of real sinusoids, $\mathbf{a}_k$ the complex amplitude and $\mathbf{b}_k$ is the complex slope of the k^{th} sinusoid. The term $n\mathbf{b}_k$ arises from the derivative of $e^{j2\pi n\hat{f}_k/f_s}$ with respect to frequency. The constant frequencies $\hat{f}_k$ define the spectral template used by QHM to fit the analysis parameters $\mathbf{a}_k$ and $\mathbf{b}_k$ by least-squares (LS) [44,45]. In principle, any set of frequencies $\hat{f}_k$ can be used because the estimation of $\mathbf{a}_k$ and $\mathbf{b}_k$ also provides a means of correcting the initial frequency values $\hat{f}_k$ by making $\hat{f}_k$ converge to nearby frequencies f_k present in the signal frame. The mismatch between f_k and $\hat{f}_k$ leads to an estimation error $\eta_k = f_k - \hat{f}_k$. Pantazis *et al.* [48] showed that QHM provides an estimate of η_k given by

$$\hat{\eta}_k=\frac{f_s}{2\pi}\frac{\operatorname{Re}\left\{\mathbf{a}_k\right\}\operatorname{Im}\left\{\mathbf{b}_k\right\}-\operatorname{Im}\left\{\mathbf{a}_k\right\}\operatorname{Re}\left\{\mathbf{b}_k\right\}}{\left|\mathbf{a}_k\right|^2},\tag{4}$$

which corresponds to the frequency correction block in Figure 2. Then $\hat{x}\left(n,m\right)$ is locally synthesized as

$$\hat{x}\left(n,m\right)=\sum_{k=-K}^{K}\hat{a}_k e^{j\left(2\pi\hat{F}_k n/f_s+\hat{\phi}_k\right)},\tag{5}$$

where $\hat{a}_k=\left|\mathbf{a}_k\right|$, $\hat{F}_k=\hat{f}_k+\hat{\eta}_k$, and $\hat{\phi}_k=\angle\mathbf{a}_k$ are constant inside the frame m.

The full-band harmonic spectral template shown in Figure 1b is obtained by setting $\hat{f}_k = kf_0$ with k an integer and $1\leq k\leq {}^{f_s}/_{2f_0}$. The f_0 is not necessary to estimate the parameters, but it improves the fit because the initial full-band harmonic template approximates better the spectrum of isolated quasi-harmonic sounds. QHM assumes that the sound being analyzed contains a single source, so, for isolated notes from pitched musical instruments, a constant f_0 is used across all frames m.

3.2. Parameter Interpolation across Frames

The model parameters $\hat{a}_k$, $\hat{F}_k$, and $\hat{\phi}_k$ from Equation (5) are estimated as samples at the frame rate ${}^{1}/_{H}$ of the amplitude- and frequency-modulation (AM–FM) functions $\hat{a}_k\left(n\right)$ and $\hat{\phi}_k\left(n\right) = {}^{2\pi}/_{f_s}\hat{F}_k\left(n\right)+\hat{\phi}_k$, which describe, respectively, the long-term amplitude and frequency temporal variation of each sinusoid k. For each frame m, $\hat{a}_k\left(\tau,m\right)$ and $\hat{F}_k\left(\tau,m\right)$ are estimated using the sample index at the center of the frame $n=\tau$ as reference. Resynthesis of $\hat{x}\left(n,m\right)$ requires $\hat{a}_k\left(n,m\right)$ and $\hat{F}_k\left(n,m\right)$ at the signal sampling rate f_s. Equation (5) uses constant values, resulting in locally stationary sinusoids with constant amplitudes and frequencies inside each frame m.

However, the parameter values might vary across frames, resulting in discontinuities such as $\hat{a}_k\left(\tau,m\right)\neq\hat{a}_k\left(\tau,m+1\right)$ due to temporal variations happening at the frame rate ${}^{1}/_{H}$. OLA resynthesis [52] uses the analysis window $w\left(n\right)$ to taper discontinuities at the frame boundaries by resynthesizing $\hat{x}\left(n,m\right)=\hat{x}\left(n\right)w\left(n\right)$ for each m similarly to Equation (2) and then overlap-adding $\hat{x}\left(n,m\right)$ across m to obtain $\hat{x}\left(n\right)$.

Additive synthesis is an alternative to OLA that results in smoother temporal variation [20] by first interpolating $\hat{a}_k\left(\tau,m\right)$ and $\hat{\phi}_k\left(\tau,m\right)$ across m and then summing over k. In this case, $\hat{a}_k\left(n\right)$ is obtained by linear interpolation of $\hat{a}_k\left(\tau,m\right)$ and $\hat{a}_k\left(\tau,m+1\right)$. Recursive calculation across m results in a piece-wise linear approximation of $\hat{a}_k\left(n\right)$. $\hat{F}_k\left(n\right)$ is estimated via piece-wise polynomial interpolation of $\hat{F}_k\left(\tau,m\right)$ across m with quadratic splines, and $\hat{\phi}_k\left(n\right)$ is obtained integrating $\hat{F}_k\left(n\right)$ in two steps because $\hat{\phi}_k\left(\tau,m\right)$ is wrapped around 2π across m. First, $\bar{\phi}_k\left(n\right)$ is calculated as

$$\bar{\phi}_k\left(n\right)=\hat{\phi}_k\left(\tau,m\right)+\frac{2\pi}{f_s}\sum_{u=m}^{m+1}\hat{F}_k\left(u\right).\tag{6}$$

The calculation of $\bar{\phi}_k\left(n\right)$ using Equation (6) does not guarantee that $\bar{\phi}_k\left(\tau, m+1\right) = \hat{\phi}_k\left(\tau, m+1\right) + 2\pi P$, with P the closest integer to unwrap the phase (see details in [45]). Thus, $\hat{\phi}_k\left(n\right)$ is calculated as

$$\hat{\phi}_k\left(n\right) = \hat{\phi}_k\left(\tau, m\right) + \frac{2\pi}{f_s}\sum_{u=m}^{m+1}\left[\hat{F}_k\left(u\right) + \gamma \sin\left(\frac{\pi\left(u - m\tau\right)}{\left(m+1\right)\tau - m\tau}\right)\right], \tag{7}$$

where the term given by the sine function ensures continuity with $\hat{\phi}_k\left(\tau, m+1\right)$ when γ is

$$\gamma = \frac{\pi}{2}\left[\frac{\hat{\phi}_k\left(\tau, m+1\right) + P - \bar{\phi}_k\left(\tau, m+1\right)}{\left(m+1\right)\tau - m\tau}\right], \tag{8}$$

with P given by $|\hat{\phi}_k\left(\tau, m+1\right) - \bar{\phi}_k\left(\tau, m+1\right)|$ (see [45]).

3.3. The Extended Adaptive Quasi-Harmonic Model (eaQHM)

Pantazis *et al.* [45] proposed adapting the phase of the sinusoids. The adaptive procedure applies LS, frequency correction, and frequency interpolation iteratively (see Figure 2), projecting $x\left(n,m\right)$ onto $\hat{x}\left(n,m\right)$. Figure 3 shows the first and second iterations to illustrate adaptation of one sinusoid. Kafentzis *et al.* [43] adapted both the instantaneous amplitude and the instantaneous phase of $\hat{x}\left(n,m\right)$ with a similar iterative procedure in eaQHM. The analysis stage uses

$$\hat{x}\left(n,m\right) = \sum_{k=-K}^{K}\left(\mathbf{a}_k + n\mathbf{b}_k\right)\hat{A}_k\left(n,m\right)e^{j\hat{\Phi}_k\left(n,m\right)}, \tag{9}$$

where $\hat{A}_k\left(n,m\right)$ and $\hat{\Phi}_k\left(n,m\right)$ are functions of the time-varying instantaneous amplitude and phase of each sinusoid, respectively [43,45], obtained from the parameter interpolation step and defined as

$$\hat{A}_k\left(n,m\right) = \frac{\hat{a}_k\left(n\right)}{\hat{a}_k\left(\tau,m\right)}, \tag{10a}$$

$$\hat{\Phi}_k\left(n,m\right) = \hat{\phi}_k\left(n\right) - \hat{\phi}_k\left(\tau,m\right), \tag{10b}$$

where $\hat{a}_k\left(n\right)$ is the piece-wise linear amplitude and $\hat{\phi}_k\left(n\right)$ is estimated using Equation (7). Finally, eaQHM models $x\left(n\right)$ as a set of amplitude and frequency modulated nonstationary sinusoids given by

$$\hat{x}_i\left(n\right) = \sum_{k=-K}^{K}\hat{a}_{k,i-1}\left(n\right)e^{j\hat{\phi}_{k,i-1}\left(n\right)}, \tag{11}$$

where $\hat{a}_{k,i-1}\left(n\right)$ and $\hat{\phi}_{k,i-1}\left(n\right)$ are the instantaneous amplitude and phase from the previous iteration $i-1$. Adaptation results from the iterative projection of $x\left(n\right)$ onto $\hat{x}\left(n\right)$ from $i-1$ as the model $\hat{x}\left(n\right)$ are used as nonstationary basis functions locally adapted to the time-varying behavior of $x\left(n\right)$. Note that Equation (9) is simply Equation (3) with a nonstationary basis $\hat{A}_k\left(n,m\right)e^{j\hat{\Phi}_k\left(n,m\right)}$. In fact, Equation (9) represents the next parameter estimation step, which will be again followed by frequency correction as in Figure 2. The convergence criterion for eaQHM is either a maximum number of iterations i or an adaptation threshold ε calculated as

$$\frac{\mathrm{SRER}^{i-1} - \mathrm{SRER}^{i}}{\mathrm{SRER}^{i-1}} < \varepsilon, \tag{12}$$

where the signal-to-reconstruction-error ratio (SRER) is calculated as

$$\mathrm{SRER} = 20\log_{10}\frac{\mathrm{RMS}\left(x\right)}{\mathrm{RMS}\left(x - \hat{x}\right)} = 20\log_{10}\frac{\mathrm{RMS}\left(x\right)}{\mathrm{RMS}\left(e\right)}. \tag{13}$$

The SRER measures the fit between the model $\hat{x}(n)$ and the original recording $x(n)$ by dividing the total energy in $x(n)$ by the energy in the residual $e(n)$. The higher the SRER, the better the fit. Note that ε stops adaptation whenever the fit does not improve from iteration $i-1$ to i regardless of the absolute SRER value. Thus, even sounds from the same instruments can reach different SRER.

4. Experimental Setup

We now investigate the full-band representation of musical instrument sounds and the nonstationarity of the adaptive AM–FM sinusoids from eaQHM. We aim to show that spectral decomposition fails to capture partials at the higher end of the spectrum so full-band quasi-harmonic modeling increases the quality of analysis and resynthesis by capturing sinusoids across the full range of the spectrum. Additionally, we aim to show that adaptive AM–FM sinusoids from eaQHM capture nonstationary partials inside the frame. We compare full-band modeling with eaQHM against the SM [1,20] and EDS estimated with ESPRIT [36] using the same number of partials K. We assume that the musical instrument sounds under investigation can be well represented as quasi-harmonic. Thus, we set $K_{\max}$ to the highest harmonic number k below Nyquist frequency $f_s/2$ or equivalently the highest integer K that satisfies $Kf_0 \leq f_s/2$. The fundamental frequency f_0 of all sounds was estimated using the sawtooth waveform inspired pitch estimator (SWIPE) [53] because in the experiments the frame size L, the maximum number of partials $K_{\max}$, and the full-band harmonic template depend on f_0. In the SM, K is the number of spectral peaks modeled by sinusoids. For EDS, ESPRIT uses K to determine the separation between the dimension of the signal space (sinusoidal component) and of the residual.

The SM is considered the baseline for comparison due to the quasi-stationary nature of the sinusoids and the need for spectral decomposition. EDS estimated with ESPRIT is considered the state-of-the-art due to the accurate analysis and synthesis and constant frequency of EDS inside the frame m. We present a comparison of the local and global SRER as a function of K and L for the SM and EDS against eaQHM in two experiments. In experiment 1, we vary K from 2 to $K_{\max}$ and record the SRER. In experiment 2, we vary L from $3T_0f_s$ to $8T_0f_s$ samples and record the SRER, where $T_0 = 1/f_0$ is the fundamental period. The local SRER is calculated within the first frame $m=0$, where we expect the attack transients to be. The first frame is centered at the onset with $\tau = 0$ (and the first half is zero-padded), so artifacts such as pre-echo (in the first half of the frame) are also expected to be captured by the local SRER. The global SRER is calculated across all frames, thus considering the whole sound signal $\hat{x}(n)$. Next, we describe the musical instrument sounds modeled and the selection of parameter values for the algorithms.

4.1. The Musical Instrument Sound Dataset

In total, 92 musical instrument sounds were selected. “Popular” and “Keyboard” musical instruments are from the RWC Music Database: Musical Instrument Sound [54]. All other sounds are from the Vienna Symphonic Library [55] database of musical instrument samples. Table 1 lists the musical instrument sounds used. The recordings were chosen to represent the range of musical instruments commonly found in traditional Western orchestras and in popular recordings. Some instruments feature different registers (alto, baritone, bass, *etc*). All sounds used belong to the same pitch class (C), ranging in pitch height from C2 ($f0 \approx 65$ Hz) to C6 ($f0 \approx 1046$ Hz). The dynamics is indicated as *forte* (“f”) or *fortissimo* (“ff”), and the duration of most sounds is less than 2 s. Normal attack (“na”) and no vibrato (“nv”) were chosen whenever available. Presence of vibrato (“vib”), progressive attack (“pa”), and slow attack (“sa”) are indicated, as well as different playing modes such as *staccato* (“stacc”), *sforzando* (“sforz”), and *pizzicato* (“pz”), achieved by plucking string instruments. Extended techniques were also included, such as *tongue ram* (“tr”) for the flute, *près de la table* (“pdlt”) for the harp, muted (“mu”) strings, and bowed idiophones (vibraphone, xylophone, *etc.*) for short (“sh”) and long (“lg”) sounds.

Different mallet materials such as metal ("met"), plastic ("pl"), and wood ("wo") and hardness such as soft ("so"), medium ("med"), and hard ("ha") are indicated.

Table 1. Musical instrument sounds used in all experiments. See text in Section 4.1 for a description of the terms in brackets. Sounds **in bold** were used in the listening test described in Section 6. The quasi-harmonic model (QHM) failed for the sounds *in italics* marked *.

Family	Musical Instrument Sounds
Brass	Bass Trombone (C3 f nv na), Bass Trombone (C3 f stac), Bass Trumpet (C3 f na vib), Cimbasso (C3 f nv na), Cimbasso (C3 f stac), *Contrabass Trombone** (C2♯ f stac), Contrabass Tuba (C3 f na), Contrabass Tuba (C3 f stac), Cornet (C4 f), French Horn (C3 f nv na), **French Horn** (C3 f stac), Piccolo Trumpet (C5 f nv na), Piccolo Trumpet (C5 f stac), Tenor Trombone (C3 f na vib), **Tenor Trombone** (C3 f nv sa), Tenor Trombone (C3 f stac), **C Trumpet** (C4 f nv na), C Trumpet (C4 f stac), Tuba (C3 f vib na), Tuba (C3 f stac), Wagner Tuba (C3 f na), Wagner Tuba (C3 f stac)
Woodwinds	Alto Flute (C4 f vib na), Bass Clarinet (C3 f na), **Bass Clarinet** (C3 f sforz), Bass Clarinet (C3 f stac), Bassoon (C3 f na), Bassoon (C3 f stac), Clarinet (C4 f na), Clarinet (C4 f stac), *Contra Bassoon** (C2 f stac), *Contra Bassoon** (C2 f sforz), English Horn (C4 f na), English Horn (C4 f stac), **Flute** (C4 f nv na), Flute (C4 f stac), Flute (C4 f tr), Flute (C4 f vib na), **Oboe 1** (C4 f stac), **Oboe 2** (C4 f nv na), Oboe (C4 f pa), Piccolo Flute (C6♯ f vib sforz), Piccolo Flute (C6 f nv ha ff)
Plucked Strings	Cello (C3 f pz vib), **Harp** (C3 f), Harp (C3 f pdlt), Harp (C3 f mu), Viola (C3 f pz vib), Violin (C4 f pz mu)
Bowed Strings	**Cello** (C3 f vib), Cello (C3 f stac), **Viola** (C3 f vib), Viola (C4 f stac), Violin (C4 f), Violin (C4♯ ff vib), Violin (C4 f stac)
Struck Percussion	**Glockenspiel** (C4 f), Glockenspiel (C6 f wo), Glockenspiel (C6 f pl), Glockenspiel (C6 f met), Marimba (C4 f), Vibraphone (C4 f ha 0), Vibraphone (C4 f ha fa), Vibraphone (C4 f ha sl), Vibraphone (C4 f med 0), Vibraphone (C4 f med fa), Vibraphone (C4 f med 0 mu), **Vibraphone** (C4 f med sl), Vibraphone (C4 f so 0), Vibraphone (C4 f so fa), Xylophone (C5 f GA L), Xylophone (C5 met), Xylophone (C5 f HO L), Xylophone (C5 f mP L), **Xylophone** (C5 f wP L)
Bowed Percussion	Vibraphone (C4 f sh vib), Vibraphone (C4 f sh nv), Vibraphone (C4 f lg nv)
Popular	**Accordion** (C3♯ f), **Acoustic Guitar** (C3 f), Baritone Sax (C3 f), **Bass Harmonica** (C3♯ f), Chromatic Harmonica (C4 f), Classic Guitar (C3 f), Mandolin (C4 f), **Pan Flute** (C5 f), **Tenor Sax** (C3♯ f), **Ukulele** (C4 f)
Keyboard	**Celesta** (C3 f na nv), Celesta (C3 f stac), Clavinet (C3 f), **Piano** (C3 f)

In what follows, we will present the results for 89 sounds because QHM failed to adapt for the three sounds marked * in Table 1. The estimation of parameters for QHM uses LS [45]. The matrix inversion fails numerically when the matrix is close to singular (see [44]). The fundamental frequency (C2 $\approx$ 65 Hz) of these sounds determines a full-band harmonic spectral template whose frequencies are separated by C2, which results in singular matrices.

4.2. Analysis Parameters

The parameter estimation for the SM follows [20] with a Hann window for analysis, and phase interpolation across frames via cubic splines followed by additive resynthesis. The estimation of parameters for EDS uses ESPRIT with a rectangular window for analysis and OLA resynthesis [36]. Parameter estimation in eaQHM used a Hann window for analysis and additive resynthesis following Equation (11). In all experiments, ε in Equation (12) is set to 0.01 and $f_s = 16$ kHz for all sounds. The step size for analysis (and OLA synthesis) was $H = 16$ samples for all algorithms, corresponding to 1 ms. The frame size is $L = qT_0f_s$ samples with q an integer. The size of the FFT for the SM is kept constant at $N = 4096$ samples with zero padding.

5. Results and Discussion

5.1. Adaptation Cycles in eaQHM

Figure 4 shows the global and local SRER as a function of the number of adaptation cycles (iterations). Each plot was averaged across the sounds indicated, while the plot "all instruments" is an average of the previously shown. The SRER increases quickly after a few iterations, slowly converging to a final value considerably higher than before adaptation. Iteration 0 corresponds to QHM initialized with the full-band harmonic template, thus Figure 4 demonstrates that the adaptation of the sinusoids by eaQHM increases the SRER when compared to QHM.

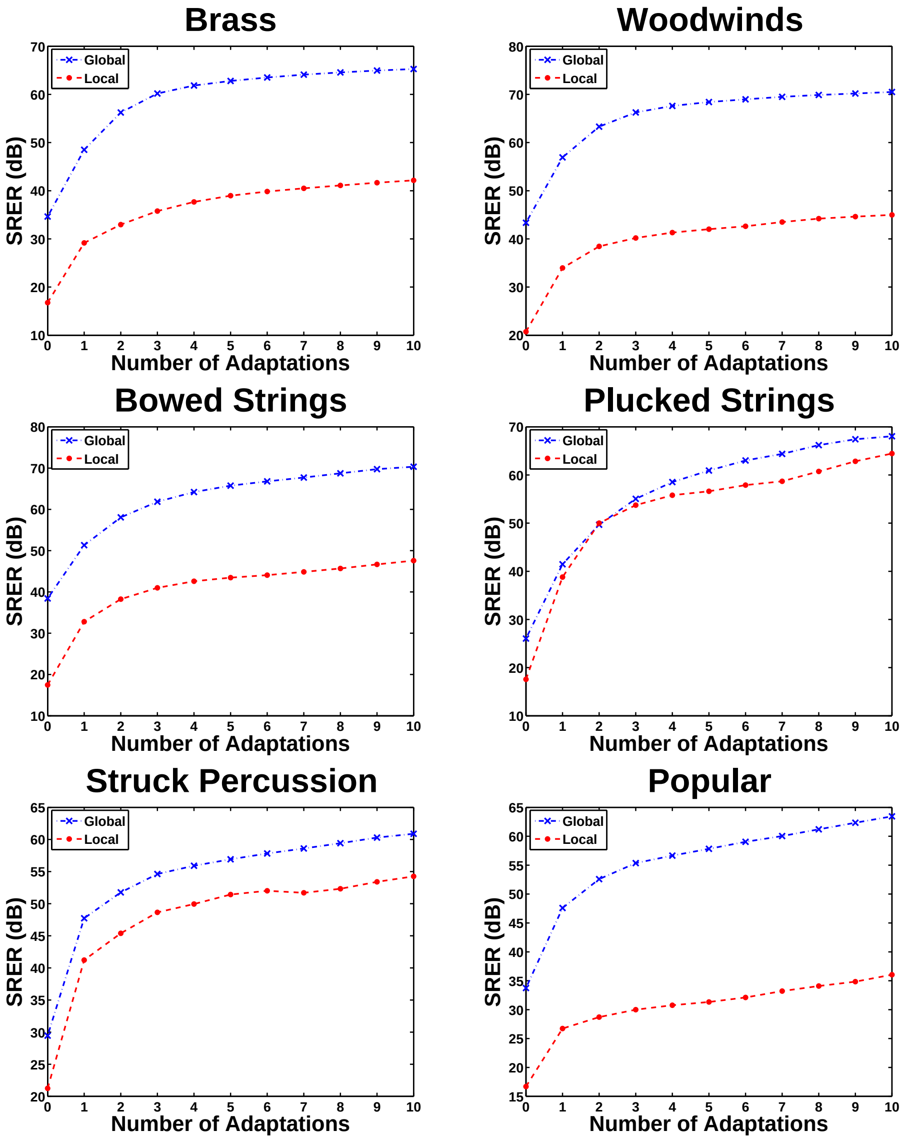


Figure 4. *Cont.*

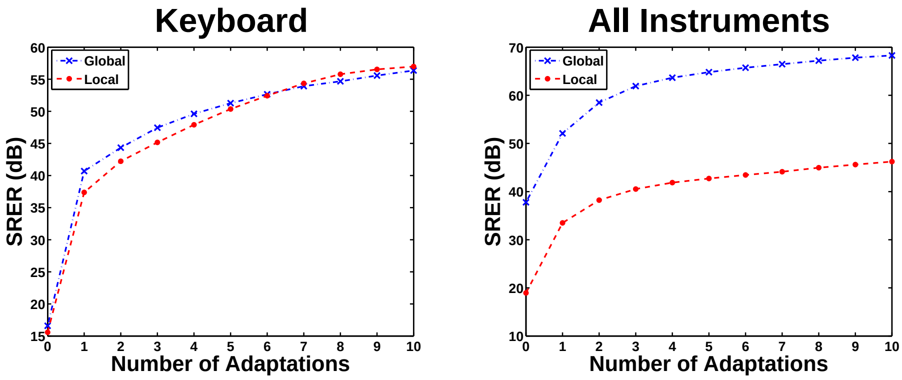


Figure 4. Plot of the signal-to-reconstruction-error ratio (SRER) as a function of number of adaptations to illustrate how adaptation increases the SRER in eaQHM. Iteration 0 corresponds to QHM initialized with the full-band harmonic spectral template.

5.2. Experiment 1: Variation Across K (Constant $L = 3T_0 f_s$)

We ran each algorithm varying K (the frame size was kept at $L = 3T_0 f_s$) and recorded the resulting local and global SRER values. We started from $K = 2$ and increased K by two partials up to $K_{\max}$. Figure 5 shows the local and global SRER (averaged across sounds) as a function of K for the SM, EDS, and eaQHM. Sounds with different f_0 values have different $K_{\max}$. Figure 5 shows that the addition of partials for the SM does not result in an increase in SRER after a certain K. EDS tends to continuously increase the SRER with more partials that capture more spectral energy. Finally, eaQHM increases the SRER up to $K_{\max}$.

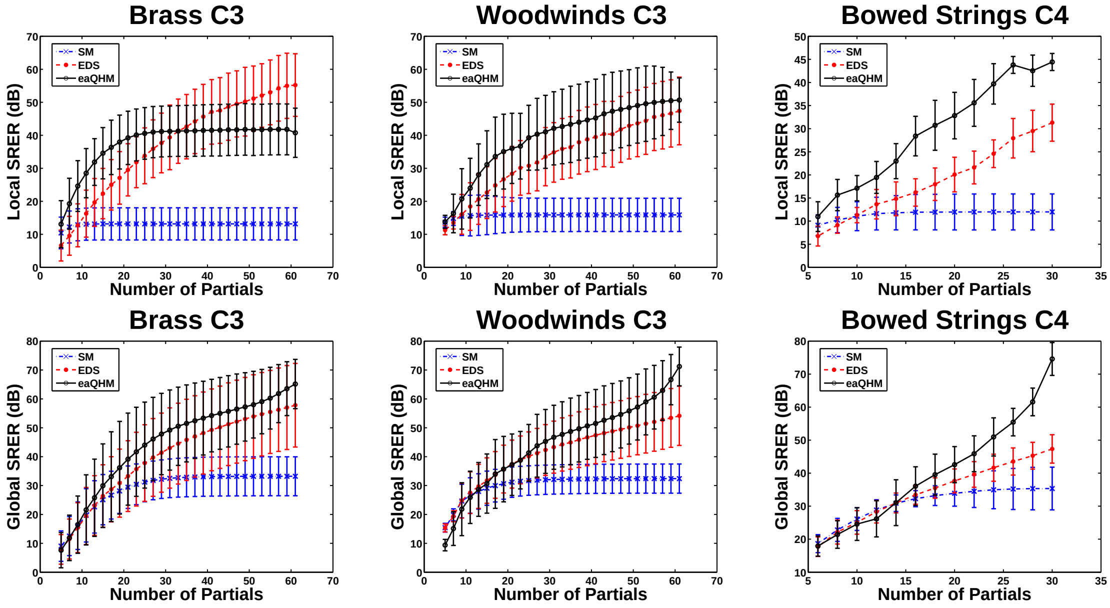


Figure 5. *Cont.*

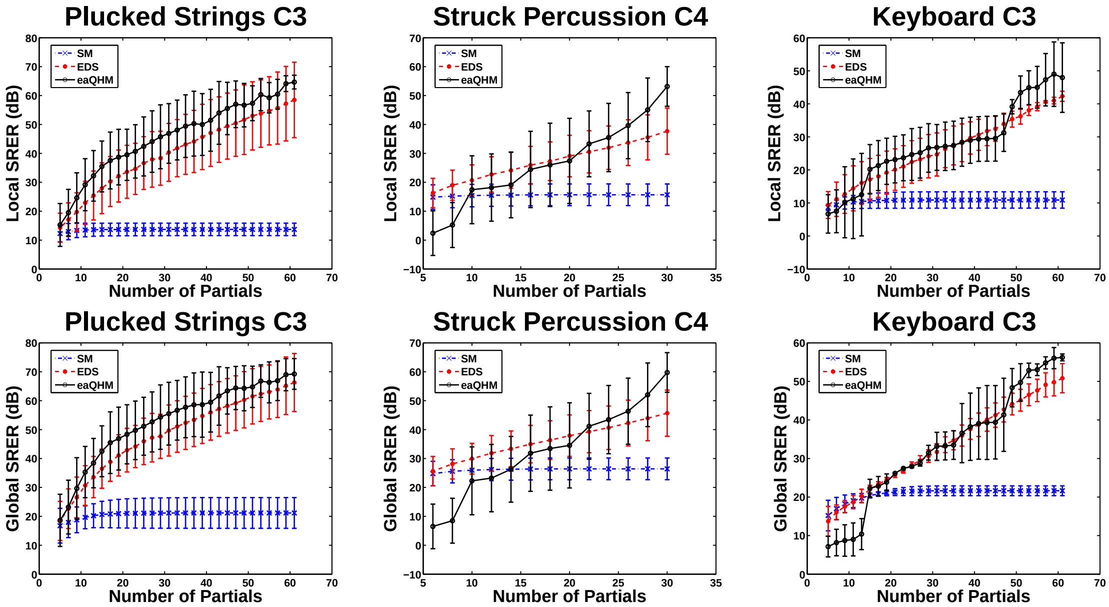


Figure 5. Comparison between local and global SRER as a function of the number of partials for the three models (the standard sinusoidal model (SM), exponentially damped sinusoids (EDS), and eaQHM). The bars around the mean are the standard deviation across different sounds from the family indicated. The distributions are not symmetrical as suggested by the bars.

The SM, EDS, and eaQHM use different analysis and different synthesis methods, which partially explains the different behavior under variation of K. More importantly, the addition of partials for each algorithm uses different criteria. Both the SM and EDS use spectral energy as a criterion, while eaQHM uses the frequencies of the sinusoids assuming quasi-harmonicity. In the SM, a new sinusoid is selected as the next spectral peak (increasing frequency) with spectral energy above a selected threshold regardless of the frequency of the peak. In fact, the frequency is estimated from the peak afterwards. For EDS, K determines the number of sinusoids used upon resynthesis. However, ESPRIT ranks the sinusoids by decreasing eigenvalue rather than the frequency, adding partials with high spectral energy that will increase the fit of the reconstruction. The frequencies of the new partials are not constrained by harmonicity. Finally, eaQHM uses the spectral template to search for nearby spectral peaks with LS and frequency correction. The sinusoids will converge to spectral peaks in the neighborhood of the harmonic template with K harmonically related partials starting from f_0. Therefore, $K_{\max}$ in eaQHM corresponds to full-band analysis and synthesis but not necessarily for the SM or EDS.

5.3. Experiment 2: Variation Across L (Constant $K = K_{\max}$)

We ran each algorithm varying L from $3T_0f_s$ to $8T_0f_s$ with a constant number of partials $K_{\max}$ and measured the resulting local and global SRER. In the literature [46], $L = 3T_0f_s$ is considered a reasonable value for speech and audio signals when using the SM. We are unaware of a systematic investigation of how L affects modeling accuracy for EDS. Figure 6 shows the local and global SRER (averaged across sounds) as a function of L expressed as q times T_0f_s, so sounds with different f_0 values have different frame size L in samples.

Figure 6 shows that the SRER decreases with L for all algorithms. The SM seldom outperforms EDS or eaQHM, but it is more robust against variations of L. For the SM, L affects both spectral estimation and temporal representation. In the FFT, L determines the trade-off between temporal and spectral resolution, which affects the performance of the peak picking algorithm for parameter estimation. The temporal representation is affected because the parameters are an average across L referenced to the center of the frame. In turn, ESPRIT estimates EDS with constant frequency inside the frames referenced to the beginning of the frame, thus L affects the temporal modeling accuracy more than the spectral estimation. However, the addition of sinusoids might compensate for the stationary frequency of EDS inside the frame. Finally, the SRER for eaQHM decreases considerably when L increases because L adversely affects the frequency correction and interpolation mechanisms. Frequency correction is applied at the center of the analysis frame and eaQHM uses spline interpolation to capture frequency modulations across frames. Thus, adaptation improves the fit more slowly for longer L, generally reaching a lower absolute SRER value.

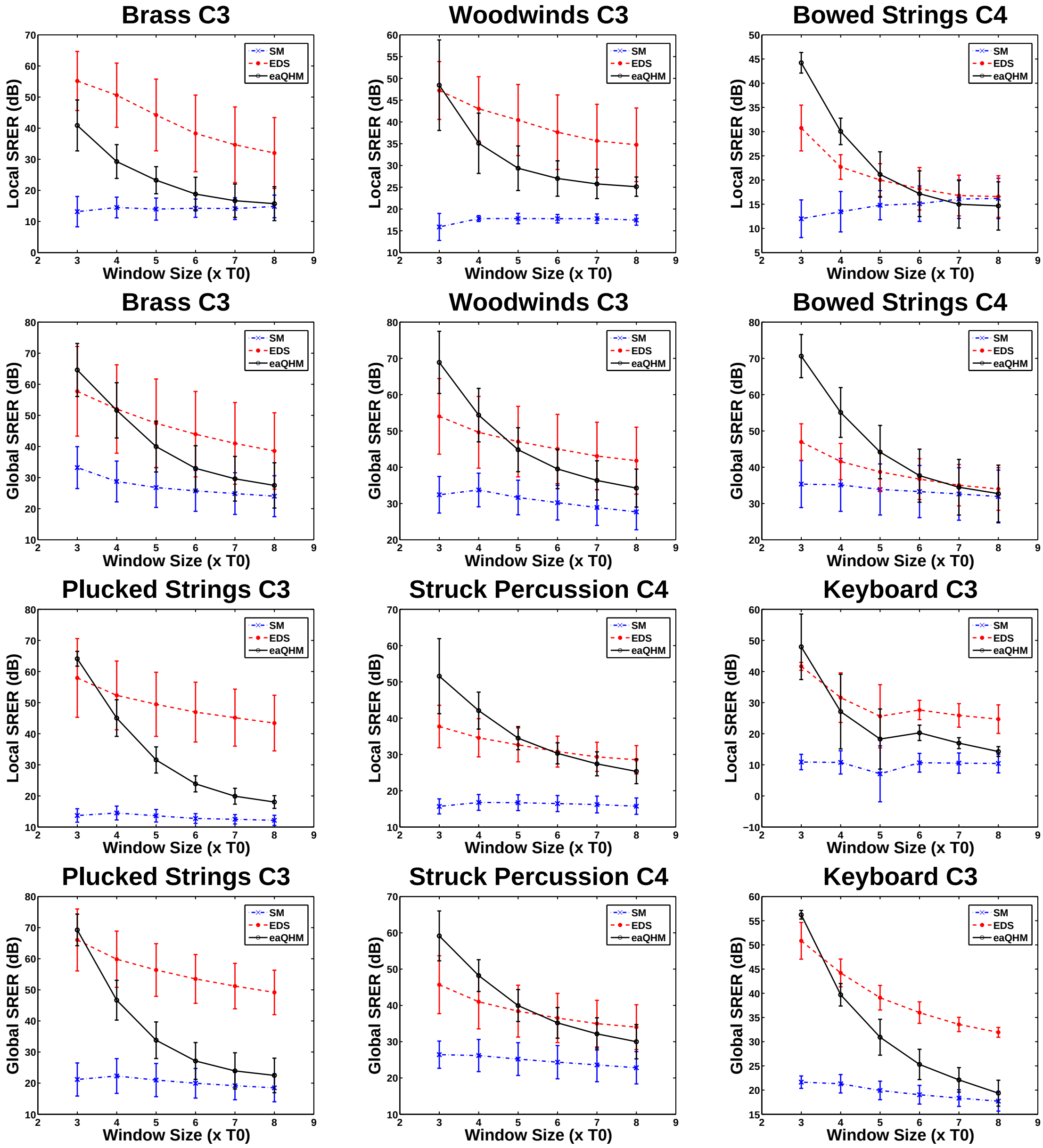


Figure 6. Comparison between local and global SRER as a function of the size of the frame for the three models (SM, EDS, and eaQHM). The bars around the mean are the standard deviation across different sounds from the family indicated. The distributions are not symmetrical as suggested by the bars.

5.4. Full-Band Quasi-Harmonic Analysis with AM–FM Sinusoids

To simplify the comparison and reduce the information, we present the differences of SRER instead of absolute SRER values. For each sound, we subtract the absolute SRER values (in dB) for the SM and EDS from that of eaQHM to obtain the differences of SRER. The local value measures the fit for the attack and the global value measures the overall fit. Positive values indicate that eaQHM results in higher SRER than the other method for that particular sound, while a negative value means the opposite. The different SRER values are averaged across all musical instruments that belong to the family indicated. Table 2 shows the comparison of eaQHM against EDS and the SM with $K = K_{\max}$ and $L = 3T_0 f_s$ clustered by instrumental family. The distributions are not symmetrical around the mean as suggested by the standard deviation.

Table 2. Local and global difference of signal-to-reconstruction-error ratio (SRER) comparing eaQHM with exponentially damped sinusoids (EDS) and eaQHM with the standard sinusoidal model (SM) for the frame size $L = 3T_0 f_s$ and number of partials $K = K_{\max}$. The three C2 sounds are not included.

	SRER (eaQHM-EDS)		**SRER (eaQHM-SM)**	
Family	**Local (dB)**	**Global (dB)**	**Local (dB)**	**Global (dB)**
Brass	$-\mathbf{9.4} \pm 7.0$	12.5 ± 6.8	27.3 ± 5.8	31.9 ± 4.0
Woodwinds	7.8 ± 3.9	22.0 ± 5.9	30.9 ± 7.5	36.1 ± 4.7
Bowed Strings	12.2 ± 4.2	24.1 ± 6.7	35.0 ± 4.7	40.0 ± 4.7
Plucked Strings	8.3 ± 5.0	4.7 ± 3.4	49.5 ± 4.3	46.6 ± 5.1
Bowed Percussion	$-\mathbf{2.7} \pm 2.5$	16.3 ± 2.2	12.7 ± 2.6	37.6 ± 3.6
Struck Percussion	10.5 ± 4.8	10.1 ± 2.6	28.6 ± 13.3	26.0 ± 11.3
Popular	6.3 ± 3.3	11.9 ± 7.0	26.5 ± 10.8	27.5 ± 11.6
Keyboard	5.7 ± 3.4	5.4 ± 4.3	37.0 ± 8.0	34.6 ± 2.0
Total	5.3 ± 2.4	13.2 ± 3.3	31.0 ± 7.1	35.0 ± 5.9

Thus, Table 2 summarizes the result of full-band quasi-harmonic analysis with adaptive AM–FM sinusoids from eaQHM comparing with the SM and EDS under the same conditions, namely the same number of sinusoids $K = K_{\max}$ and frame size $L = 3T_0 f_s$. When eaQHM is compared to the SM, both local and global difference SRER are positive for all families. This means that full-band quasi-harmonic modeling with eaQHM results in a better fit for the analysis and synthesis of musical instrument sounds.

When eaQHM is compared to EDS, all global difference SRER are positive and all local difference SRER are positive except for *Brass* and *Bowed Percussion*. Thus, EDS can fit the attack of *Brass* and *Bowed Percussion* better than eaQHM. The exponential amplitude envelope of EDS is considered suitable to model percussive sounds with sharp attacks such as harps, pianos, and marimbas [36,37]. The musical instrument families that contain percussive sounds are *Plucked strings*, *Struck percussion*, and *Keyboard*. Table 2 shows that eaQHM outperformed EDS locally and globally for all percussive sounds. The ability to adapt the amplitude of the sinusoidal partials to the local characteristics of the waveform makes eaQHM extremely flexible to fit both percussive and nonpercussive musical instrument sounds. On the other hand, both *Brass* and *Bowed Percussion* present slow attacks typically lasting longer than one frame L. Note that $L/f_s = 3T_0 \approx 22$ ms for C3 ($f_0 \approx 131$ Hz) while *Bowed Percussion* can have attacks longer than 100 ms. Therefore, one frame $L = 3T_0 f_s$ does not measure the fit for the entire duration of the attack.

Note that the local SRER is important because the global SRER measures the overall fit without indication of *where* the differences lie in the waveform. For musical instrument sounds, differences in the attack impact the results differently than elsewhere because the attack is among the most important perceptual features in dissimilarity judgment [56–58]. Consequently, when comparing two models with the global SRER, it is only safe to say that a higher SRER indicates that resynthesis results in a waveform that is closer to the original recording.

5.5. Full-Band Modeling and Quasi-Harmonicity

Time-frequency transforms such as the STFT represent L samples in a frame with N DFT coefficients provided that $N \geq L$. Note that $N \in \mathbb{C}$, corresponding to $p = 2N$ real numbers. There is signal expansion whenever the representation uses p parameters to represent L samples and $p > L$. Sinusoidal models represent L samples in a frame with K sinusoids. In turn, each sinusoid is described by p parameters, requiring pK parameters to represent L samples. Therefore, there is a maximum number of sinusoids to represent a frame without signal expansion. For example, white noise has a flat spectrum across that would take a large number of sinusoids close together in frequency resulting in signal expansion.

The pK parameters to represent L samples can be interpreted as the degrees of freedom of the fit. As a general rule, more parameters mean greater flexibility of representation (hence potentially a better fit), but with the risk of over-fitting. Table 3 shows a comparison of the number of real parameters p (per sinusoid k per frame m) for the analysis and synthesis stages of the SM, EDS, and eaQHM. Note that eaQHM and EDS require more parameters than the SM at the *analysis* stage, but eaQHM and the SM require fewer parameters than EDS for the *synthesis* stage. The difference is due to the resynthesis strategy used by each algorithm. EDS uses OLA resynthesis, which requires all analysis parameters for resynthesis, while both eaQHM and the SM use additive resynthesis.

Table 3. Comparison of the number of real parameters p per sinusoid k per frame m for the analysis and synthesis stages of the SM, EDS, and eaQHM. The table presents the number of real parameters p to estimate and to resynthesize each sinusoid inside a frame.

	Number of Real Parameters p Per Sinusoid k Per Frame m		
	SM	**EDS**	**eaQHM**
Analysis	$p = 3$	$p = 4$	$p = 4$
Synthesis	$p = 3$	$p = 4$	$p = 3$

Harmonicity of the partials guarantees that there are no signal expansions in full-band modeling with sinusoids. Consider $L = qT_0 f_s$ with q an integer and $T_0 = {}^1/_{f_0}$. Using $K_{\max} \approx {}^{f_s}/_{2f_0}$ quasi-harmonic partials and p parameters per partial, it takes at most $pK_{\max} = {}^{(pf_s)}/_{2f_0}$ numbers to represent $L = qT_0 f_s = {}^{(qf_s)}/_{f_0}$ samples, which gives the ratio $r = {}^{(pK_{\max})}/_{L} = {}^{p}/_{2q}$. Table 3 shows that analysis with eaQHM requires $p = 4$ real parameters. Thus, a frame size with $q > 2$ is enough to guarantee no signal expansion. This result is due to the full-band paradigm using $K_{\max}$ harmonically related partials, not a particular model. The advantage of full-band modeling results from the use of one single component instead of decomposition.

Table 4 compares the complexity of SM, EDS, and eaQHM in Big-O notation. The complexity of SM is $\mathcal{O}\left(N \log N\right)$, which is the complexity of the FFT algorithm for size N inputs. ESPRIT estimates the parameters of EDS with singular value decomposition (SVD), whose algorithmic complexity is $\mathcal{O}(L^2 + K^3)$ for an L by K matrix (frame size *versus* the number of sinusoids). Adaptation in eaQHM is an iterative fit where each iteration i requires running the model again as described in Section 3. For each iteration i, eaQHM estimates the parameters with least squares (LS) via calculation of the pseudoinverse matrix using QR decomposition. The algorithmic complexity of QR decomposition is $\mathcal{O}\left(K^3\right)$ for a square matrix of size K (the number of sinusoids).

Adaptation of the sinusoids in eaQHM can result in over-fitting. The amplitude and frequency modulations capture temporal variations inside the frame such as transients and instrumental noise around the partials. However, adaptation must not capture noise resulting from sources such as quantization, which is extraneous to the sound. Ideally, the residual should contain only external additive noise without any perceptually important information from the sound [17].

Table 4. Comparison of algorithmic complexity in Big-O notation. The table presents the complexity as a function of the size of the input N, L, and K and the number of iterations i. See text for details.

	Algorithmic Complexity		
	SM	**EDS**	**eaQHM**
Complexity	$\mathcal{O}\left(N \log N\right)$	$\mathcal{O}(L^2 + K^3)$	$\mathcal{O}\left(iK^3\right)$

6. Evaluation of Perceptual Transparency with a Listening Test

We performed a listening test to validate the full-band representation of musical instrument sounds with eaQHM. The aim of the test was to evaluate whether full-band modeling with eaQHM resulted in resynthesized musical instrument sounds that are perceptually indistinguishable from the original recordings. The 21 sounds **in bold** in Table 1 were selected for the listening test, which presented pairs *original* and *resynthesis*. The participants were instructed to listen to each pair as many times as necessary and to answer the question "Can you tell the difference between the two sounds in each pair?" Full-band (FB) resynthesis with eaQHM (using a harmonic template with $K = K_{\text{max}}$ sinusoids) was used for all 21 musical instrument sounds. For nine of these sounds, half-band (HB) resynthesis with eaQHM (using a harmonic template with $K = K_{\text{max}}/2$ sinusoids) was also included as control group to test the aptitude of the listeners and compare against the FB version. All HB versions were placed at random positions among the FB, so the test presented 30 pairs overall. The listening test can be accessed at [59].

In total, 20 people aged between 26 and 40 took the test. The participants declared themselves as experienced with listening tests and familiar with signal processing techniques. Figure 7 shows the result of the listening test as the percentage of the people who answered "no" to the question, indicating that they cannot tell the difference between the original recording and the resynthesis. In general, the result of the listening test shows that full-band modeling with eaQHM results in perceptually indistinguishable resynthesis for most musical instrument sounds tested. The figure indicates that 10 out of the 21 FB sounds tested were rated perceptually identical to the original by 100% of the listeners. As expected, most HB sounds fall under 30% (except *Tenor Trombone*) and most FB sounds lie above 70% (except *Pan Flute*). Table 1 shows that *Tenor Trombone* is played at C3 and *Pan Flute* at C5. The Tenor Trombone sound is not bright, which indicates that there is little spectral energy at the higher frequency end of the spectrum. Thus, the HB version synthesized with fewer partials than K_{max} was perceived as identical to the original by some listeners. The Pan Flute sound contains a characteristic breathing noise captured as AM–FM elements in eaQHM. However, the breathing noise in the full-band version sounds brighter than the original recording and most listeners were able to tell the difference.

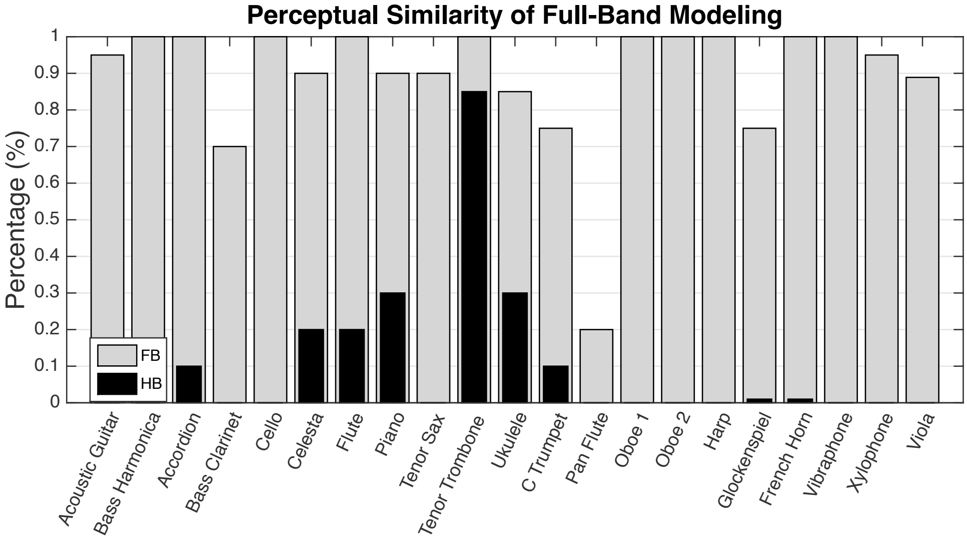


Figure 7. Result of the listening test on perceptual similarity of full-band (FB) and half-band (HB) resynthesis with eaQHM compared to the original recording. The sounds used in the listening test appear **in bold** in Table 1.

7. Conclusions

We proposed the full-band quasi-harmonic modeling of musical instrument sounds with adaptive AM–FM sinusoids from eaQHM as an alternative to spectrum decomposition. We used the SRER to measure the fit of the sinusoidal model to the original recording of 89 percussive and nonpercussive musical instruments sounds from different families. We showed that full-band modeling with eaQHM results in higher global SRER values when compared to the standard SM and to EDS estimated with ESPRIT for $K_{\max}$ sinusoids and frame size $L = 3T_0 f_s$. EDS resulted in higher local SRER than eaQHM for two of nine instrumental families, namely *Brass* and *Bowed Percussion*. A listening test confirmed that full-band modeling with eaQHM resulted in perceptually indistinguishable resynthesis for most musical instrument sounds tested.

Future work should investigate a method to prevent over-fitting with eaQHM. Additionally, the use of least-squares to estimate the parameters leads to matrices that are badly conditioned for sounds with low fundamental frequencies. A more robust estimation method to prevent bad-conditioning would improve the stability of eaQHM. Currently, eaQHM can only estimate the parameters of isolated sounds. We intend to develop a method for polyphonic instruments and music. Future work also involves using eaQHM in musical instrument sound transformation, estimation of musical expressivity features such as vibrato, and solo instrumental music. The companion webpage [60] contains sound examples. Finally, the proposal of a full-band representation of musical instrument sounds with adaptive sinusoids motivates further investigation on full-band extensions of other sinusoidal methods, such as SM and EDS used here.

Acknowledgments: This work was partly supported by project "NORTE-01-0145-FEDER-000020" financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF) and by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant 644283. The latter project also supplied funds for covering the costs to publish in open access.

Author Contributions: Marcelo Caetano conceived and designed the experiments, analyzed the data, and wrote the manuscript. George P. Kafentzis performed the experiments, helped analyze the results, and revised the manuscript. Athanasios Mouchtaris supervised the research and revised the manuscript. Yannis Stylianou supervised the research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Serra, X.; Smith, J.O. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Comput. Music J.* **1990**, *14*, 49–56.
2. Beauchamp, J.W. Analysis and synthesis of musical instrument sounds. In *Analysis, Synthesis, and Perception of Musical Sounds*; Beauchamp, J.W., Ed.; Modern Acoustics and Signal Processing; Springer: New York, NY, USA, 2007; pp. 1–89.
3. Quatieri, T.; McAuley, R. Audio signal processing based on sinusoidal analysis/synthesis. In *Applications of Digital Signal Processing to Audio and Acoustics;* Kahrs, M., Brandenburg, K., Eds.; Kluwer Academic Publishers: Berlin/Heidelberg, Germany, 2002; Chapter 9, pp. 343–416.
4. Serra, X.; Bonada, J. Sound Transformations based on the SMS high level attributes. *Proc. Digit. Audio Eff. Workshop* **1998**, *5*. Available online: http://mtg.upf.edu/files/publications/dafx98-1.pdf (accessed on 26 April 2016).
5. Caetano, M.; Rodet, X. Musical Instrument sound morphing guided by perceptually motivated features. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 1666–1675.
6. Barbedo, J.; Tzanetakis, G. Musical instrument classification using individual partials. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 111–122.
7. Herrera, P.; Bonada, J. Vibrato Extraction and parameterization in the spectral modeling synthesis framework. *Proc. Digit. Audio Eff. Workshop* **1998**, *99*. Available online: http://www.mtg.upf.edu/files/publications/dafx98-perfe.pdf (accessed on 26 April 2016).

8. Glover, J.; Lazzarini, V.; Timoney, J. Real-time detection of musical onsets with linear prediction and sinusoidal modeling. *EURASIP J. Adv. Signal Process.* **2011**, doi:10.1186/1687-6180-2011-68.
9. Virtanen, T.; Klapuri, A. Separation of harmonic sound sources using sinusoidal modeling. In Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Istanbul, Turkey, 5–9 June 2000; Volume 2, pp. II765–II768.
10. Lagrange, M.; Marchand, S.; Rault, J.B. Long interpolation of audio signals using linear prediction in sinusoidal modeling. *J. Audio Eng. Soc.* **2005**, *53*, 891–905.
11. Hermus, K.; Verhelst, W.; Lemmerling, P.; Wambacq, P.; Huffel, S.V. Perceptual audio modeling with exponentially damped sinusoids. *Signal Process.* **2005**, *85*, 163–176.
12. Nsabimana, F.; Zolzer, U. Audio signal decomposition for pitch and time scaling. In Proceedings of the International Symposium on Communications, Control, and Signal Processing (ISCCSP), St Julians, Malta, 12–14 March 2008; pp. 1285–1290.
13. El-Jaroudi, A.; Makhoul, J. Discrete all-pole modeling. *IEEE Trans. Commun. Technol.* **1969**, *39*, 481–488.
14. Caetano, M.; Rodet, X. A source-filter model for musical instrument sound transformation. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 137–140.
15. Wen, X.; Sandler, M. Source-Filter Modeling in the Sinusoidal Domain. *J. Audio Eng. Soc.* **2010**, *58*, 795–808.
16. Fletcher, N.H.; Rossing, T.D. *The Physics of Musical Instruments*, 2nd ed.; Springer: New York, NY, USA, 1998.
17. Caetano, M.; Kafentzis, G.P.; Degottex, G.; Mouchtaris, A.; Stylianou, Y. Evaluating how well filtered white noise models the residual from sinusoidal modeling of musical instrument sounds. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2013; pp. 1–4.
18. Bader, R.; Hansen, U. Modeling of musical instruments. In *Handbook of Signal Processing in Acoustics*; Havelock, D., Kuwano, S., Vorländer, M., Eds.; Springer: New York, NY, USA, 2009; pp. 419–446.
19. Fletcher, N.H. The nonlinear physics of musical instruments. *Rep. Prog. Phys.* **1999**, *62*, 723–764.
20. McAulay, R.J.; Quatieri, T.F. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust. Speech Signal Process.* **1986**, *34*, 744–754.
21. Green, R.A.; Haq, A. B-spline enhanced time-spectrum analysis. *Signal Process.* **2005**, *85*, 681–692.
22. Belega, D.; Petri, D. Frequency estimation by two- or three-point interpolated Fourier algorithms based on cosine windows. *Signal Process.* **2015**, *117*, 115–125.
23. Prudat, Y.; Vesin, J.M. Multi-signal extension of adaptive frequency tracking algorithms. *Signal Process.* **2009**, *89*, 96–973.
24. Candan, Ç. Fine resolution frequency estimation from three DFT samples: Case of windowed data. *Signal Process.* **2015**, *114*, 245–250.
25. Röbel, A. Adaptive additive modeling with continuous parameter trajectories. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1440–1453.
26. Verma, T.S.; Meng, T.H.Y. Extending spectral modeling synthesis with transient modeling synthesis. *Comput. Music J.* **2000**, *24*, 47–59.
27. Laurenti, N.; De Poli, G.; Montagner, D. A nonlinear method for stochastic spectrum estimation in the modeling of musical sounds. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 531–541.
28. Daudet, L. A review on techniques for the extraction of transients in musical signals. *Proc. Int. Symp. Comput. Music Model. Retr.* **2006**, *3902*, 219–232.
29. Jang, H.; Park, J.S. Multiresolution sinusoidal model with dynamic segmentation for timescale modification of polyphonic audio signals. *IEEE Trans. Speech Audio Process.* **2005**, *13*, 254–262.
30. Beltrán, J.R.; de León, J.P. Estimation of the instantaneous amplitude and the instantaneous frequency of audio signals using complex wavelets. *Signal Process.* **2010**, *90*, 3093–3109.
31. Levine, S.N.; Smith, J.O. A compact and malleable sines+transients+noise model for sound. In *Analysis, Synthesis, and Perception of Musical Sounds*; Beauchamp, J.W., Ed.; Modern Acoustics and Signal Processing; Springer: New York, NY, USA, 2007; pp. 145–174.
32. Markovsky, I.; Huffel, S.V. Overview of total least-squares methods. *Signal Process.* **2007**, *87*, 2283–2302.
33. Roy, R.; Kailath, T. ESPRIT-estimation of signal parameters via rotational invariance techniques. *IEEE Trans. Acoust. Speech Signal Process* **1989**, *37*, 984–995.

34. Van Huffel, S.; Park, H.; Rosen, J. Formulation and solution of structured total least norm problems for parameter estimation. *IEEE Trans. Signal Process.* **1996**, *44*, 2464–2474.
35. Liu, Z.S.; Li, J.; Stoica, P. RELAX-based estimation of damped sinusoidal signal parameters. *Signal Process.* **1997**, *62*, 311–321.
36. Nieuwenhuijse, J.; Heusens, R.; Deprettere, E.F. Robust exponential modeling of audio signals. In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Seattle, WA, USA, 12–15 May 1998; Volume 6, pp. 3581–3584.
37. Badeau, R.; Boyer, R.; David, B. EDS Parametric Modeling And Tracking of Audio Signals. In Proceedings of the 5th International Conference on Digital Audio Effects (DAFx), Hambourg, Germany, 26–28 September 2002; pp. 26–28.
38. Jensen, J.; Heusdens, R. A comparison of sinusoidal model variants for speech and audio representation. In Proceedings of the 2002 11th European Signal Processing Conference (EUSIPCO), Toulouse, France, 3–6 September 2002; pp. 1–4.
39. Auger, F.; Flandrin, P. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Trans. Signal Process.* **1995**, *43*, 1068–1089.
40. Fulop, S.A.; Fitz, K. Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. *J. Acoust. Soc. Am.* **2006**, *119*, 360–371.
41. Li, X.; Bi, G. The reassigned local polynomial periodogram and its properties. *Signal Process.* **2009**, *89*, 206–217.
42. Girin, L.; Marchand, S.; Di Martino, J.; Röbel, A.; Peeters, G. Comparing the order of a polynomial phase model for the synthesis of quasi-harmonic audio signals. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 19–22 October 2003; pp. 193–196.
43. Kafentzis, G.P.; Pantazis, Y.; Rosec, O.; Stylianou, Y. An extension of the adaptive quasi-harmonic model. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing, Kyoto, Japan, 25–30 March 2012; pp. 4605–4608.
44. Kafentzis, G.P.; Rosec, O.; Stylianou, Y. On the modeling of voiceless stop sounds of speech using adaptive quasi-harmonic models. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Portland, OR, USA, 9–13 September 2012.
45. Pantazis, Y.; Rosec, O.; Stylianou, Y. Adaptive AM–FM signal decomposition with application to speech analysis. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 290–300.
46. Degottex, G.; Stylianou, Y. Analysis and synthesis of speech using an adaptive full-band harmonic model. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 2085–2095.
47. Caetano, M.; Kafentzis, G.P.; Mouchtaris, A.; Stylianou, Y. Adaptive sinusoidal modeling of percussive musical instrument sounds. In Proceedings of the European Signal Processing Conference (EUSIPCO), Marrakech, Morocco, 9–13 September 2013; pp. 1–5.
48. Pantazis, Y.; Rosec, O.; Stylianou, Y. On the Properties of a time-varying quasi-harmonic model of speech. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Brisbane, Australia, 22–26 September 2008; pp. 1044–1047.
49. Smyth, T.; Abel, J.S. Toward an estimation of the clarinet reed pulse from instrument performance. *J. Acoust. Soc. Am.* **2012**, *131*, 4799–4810.
50. Smyth, T.; Scott, F. Trombone synthesis by model and measurement. *EURASIP J. Adv. Signal Process.* **2011**, doi:10.1155/2011/151436.
51. Brown, J.C. Frequency ratios of spectral components of musical sounds. *J. Acoust. Soc. Am.* **1996**, *99*, 1210–1218.
52. Borss, C.; Martin, R. On the construction of window functions with constant overlap-add constraint for arbitrary window shifts. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 337–340.
53. Camacho, A.; Flory, H.Y. A sawtooth waveform inspired pitch estimator for speech and music. *J. Acoust. Soc. Am.* **2008**, *124*, 1638–1652.

54. Goto, M.; Hashiguchi, H.; Nishimura, T.; Oka, R. RWC Music Database: Music Genre Database and Musical Instrument Sound Database. In Proceedings of the International Conference on Music Information Retrieval (ISMIR), Baltimore, MD, USA, 26–30 October 2003; pp. 229–230. Available online: http://staff.aist.go.jp/m.goto/RWC-MDB/ (accessed on 26 April 2016).
55. Vienna Symphonic Library–GmbH. Available online: http://www.vsl.co.at/ (accessed on 26 April 2016).
56. Grey, J.M.; Gordon, J.W. Multidimensional perceptual scaling of musical timbre. *J. Acoust. Soc. Am.* **1977**, *61*, 1270–1277.
57. Krumhansl, C.L. Why is musical timbre so hard to understand? In *Structure and Perception of Electroacoustic Sound and Music*; Nielzén, S., Olsson, O., Eds.; Excerpta Medica: New York, NY, USA, 1989; pp. 43–54.
58. McAdams, S.; Giordano, B.L. The perception of musical timbre. In *The Oxford Handbook of Music Psychology*; Hallam, S., Cross, I., Thaut, M., Eds.; Oxford University Press: New York, NY, USA, 2009; pp. 72–80.
59. Listening Test. Webpage for the Listening Test. Available online: http://ixion.csd.uoc.gr/kafentz/listest/pmwiki.php?n=Main.JMusLT (accessed on 26 April 2016).
60. AdaptiveSinMus. Companion webpage with sound examples. Available online: http://www.csd.uoc.gr/kafentz/listest/pmwiki.php?n=Main.AdaptiveSinMus (accessed on 26 April 2016).