Object-Based Sound Source Modeling for Musical Signals

Tero Tolonen
Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
Espoo, Finland
Tero.Tolonen@hut.fi
http://www.acoustics.hut.fi/˜ttolonen

Abstract

This study presents a framework for audio and music processing which consists of an analysis and a synthesis path that are connected at three representational levels. Auditory signal analysis techniques include a multi-pitch analysis model, an event-detector, and sinusoidal modeling that are combined in an iterative sound separation system. Techniques are presented for detection of perceptually relevant features, such as inharmonicity, vibrato, and decay characteristic, from polyphonic mixtures of harmonic sounds. The integration of the analysis and synthesis parts is demonstrated with examples where two-voice acoustic guitar signals are analyzed into an object-based representation and resynthesized using sound source models.

0 Introduction

Rapid development of communication technology imposes new requirements for creation, processing, and rendering of music and audio. Particularly, multimedia communication including interactive music and audio will be a major driver in development of new applications and solutions for fixed and wireless networks. Great challenges and opportunities lie ahead for developers of audio technology.

One of the promising methodologies for interactive audio solutions is object-based audio. Rather than to process the sound waveform or its frequency representation, the aim is to process sound through relevant objects and their attributes that depend on the application at hand. These can include perceptual objects for studying human audition, musical objects in music applications, and sound source objects for study of musical instruments and voice as well as for sound synthesis. For instance, the concept of auditory scene analysis (ASA) and creation of computational models for ASA is one of the hot topics of research. On the other hand, researchers in synthetic audio, musical acoustics, and model-based sound synthesis have developed object-based methods for generation of sound. The recent advances in object-based audio [1], and particularly the MPEG-4 multimedia standard [2, 3], enable interactive, low-bit-rate audio and music solutions that are attractive, e.g., for mobile multimedia applications. In addition, the object-based methodology can be applied in automatic transcription, musical information retrieval, identification of musical instruments, and object-based audio coding.

This study presents a framework for audio and music processing. The framework consists of an analysis and a synthesis path which are connected at three representational levels. The presented model-based analysis elaborates previous work on auditory analysis techniques and integrates it into a system that analyzes musical signals into an object-based representation. The synthesis part of the framework is exemplified using model-based sound synthesis methods for plucked string instruments. These include computationally efficient models that can be used to generate synthetic tones that are virtually indistinguishable from original tones.

Understanding of the human auditory system and its ability to process sound into content is often important for development of audio and music processing systems. However, the framework presented here attempts to be general in that it combines auditory-based techniques with techniques that find little motivation from the human audition but that are useful in audio and music applications. An example is signal-level sound source separation: while it is unlikely that such a separation takes place in human auditory processing, such an audio processing tool is useful in numerous applications. Furthermore, combining pure signal processing techniques with auditorily motivated analysis methods has proven successful results in sound source separation as well as in other applications such as audio coding [4].

The presented framework is not unique; many parts are similar to those in computational auditory scene analysis (CASA) and sound-source recognition systems [5, 6, 7, 8]. The framework is versatile in that its constituent parts can be used in a wide range of applications including sound-source recognition, extraction of expressive features, automatic transcription and musical information retrieval, audio coding, and sound synthesis.

The organization of the paper is the following. Section 1 describes the framework for audio and music processing. Section 2 discusses an analysis system that is employed in this study, and shows with examples how perceptually relevant features of sound sources can be obtained. It also presents how the parameters of a simple but powerful model of the acoustic guitar can be obtained from realistic musical signals. Analysis/synthesis examples of two-voice guitar music is demonstrated in Section 3. Finally, Section 4 concludes the paper and presents directions for future research. The test signals include natural and synthetic polyphonic mixtures of musical instruments. Sound demonstrations of these signals are available at [9].

1 Framework for Audio and Music Processing

Since digital audio and music processing consists of a multitude of technologies, it is useful to devise a framework which helps to organize and relate the techniques and methods. Figure 1 depicts such a general framework based on the virtual acoustics framework presented by Karjalainen [10]. The framework shows two vertical paths: an analysis path on the left and a

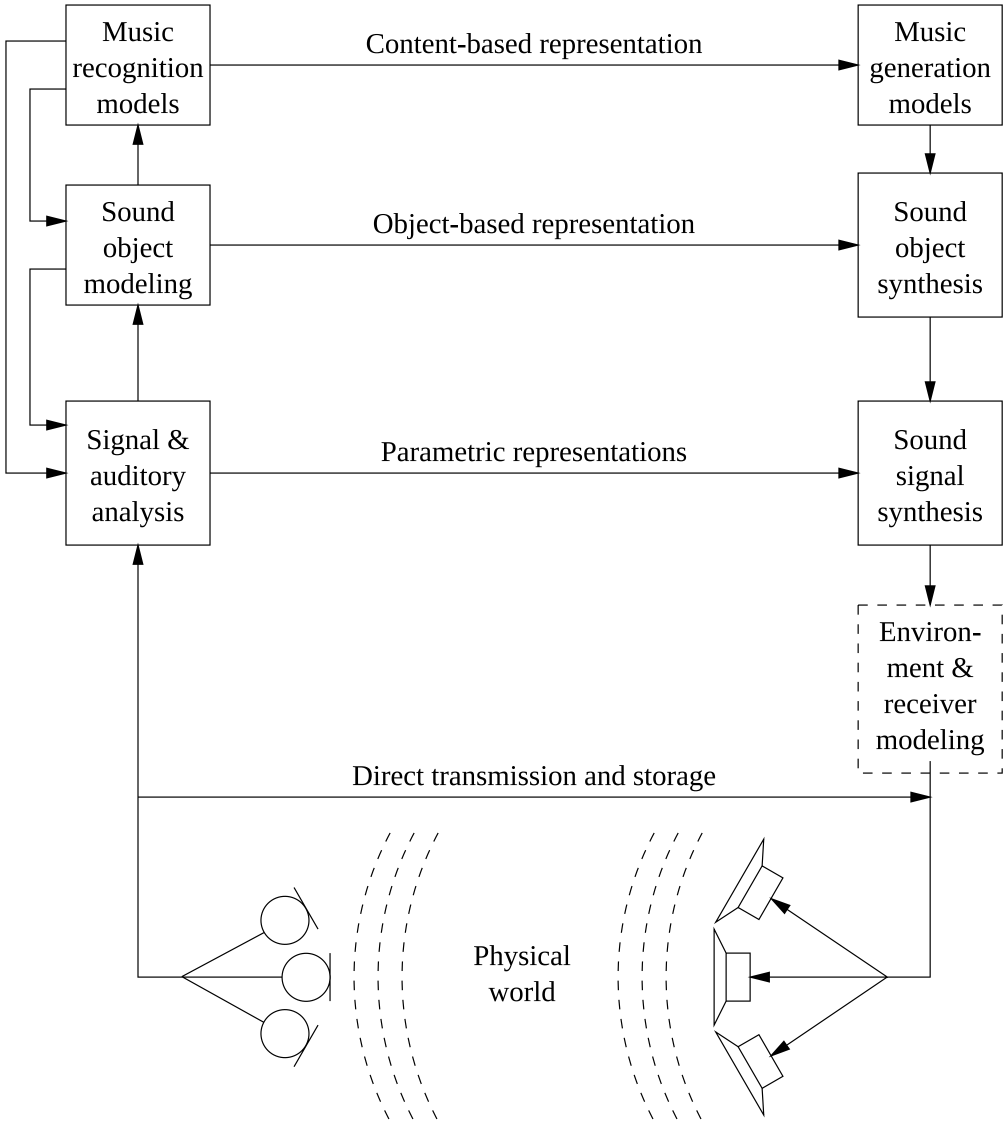
Music recognition models
Content-based representation
Music generation models
Sound object modeling
Object-based representation
Sound object synthesis
Signal & auditory analysis
Parametric representations
Sound signal synthesis
Environment & receiver modeling
Direct transmission and storage
Physical world

Figure 1: A framework for audio and music processing (after [10]).

synthesis path on the right. The analysis and synthesis sides are connected at three representation levels corresponding to different levels of abstraction between the signal and content representations. Naturally, this framework organization is only one of many alternatives, but it suites the work presented in this study and the methodologies on which the ideas presented here are built.

Most of the present digital audio applications consist of techniques at direct transmission/storage or parametric representation level. However, as the focus of research is steadily climbing up in both the analysis and synthesis path, we are likely to see new applications and solutions that operate on the object- and content-based levels. The following discussion briefly addresses the three representational levels, and gives examples of analysis and synthesis techniques in each.

1.1 Parametric Representations

If the “direct transmission and storage” path is discarded, the parametric representation provides the strongest connection between the analysis and synthesis sides of the framework. For instance, audio compression methods, such as those of the natural audio coding part of MPEG-4 [2, 11] belong to this level implementing the whole chain starting from and ending in a sound signal or an acoustic field. However, more interesting for the present discussion are methods that can be used to link to the two higher levels of the framework. On the analysis side, these consist of signal and auditory analysis tools that provide low-level parametric representations that are useful at the object-based level for, e.g., identification of perceptually relevant entities and objects and their perceptual features. On the synthesis side, such techniques include model-based sound synthesis methods (synthesis engines), i.e., computational algorithms that are controlled from the sound-object synthesis level.

Low Level Signal Components Sinusoidal modeling is a common example of a technique for presenting a signal with low level components, namely sinusoids, noise clouds, and transients [12, 13, 14, 15, 16, 17, 18]. The basic sinusoidal modeling is a pure signal processing method in which peaks corresponding to sinusoids are detected in a short-time Fourier transform (STFT) representation and associated in consecutive frames so that slowly time-varying parameter trajectories are obtained. Amplitude, frequency, and phase trajectories are typically used although in some cases the phase trajectory is discarded. Recently, auditorily motivated methods have been devised for sinusoidal modeling [19, 16]. In this case, the sinusoids are iteratively detected and sorted according to their masking properties.

Sinusoidal modeling is a useful tool in many audio applications. In principle, sinusoids with slowly time-varying parameters, noise with slowly varying frequency envelope, and transients appear an attractive combination of primitive signal components.

Periodicity Representations Periodicity or pitch detectors try to imitate auditory system’s amazing ability to segregate pitched sound objects in polyphonic mixtures. Although the current computational methods are far behind the performance of the auditory system, recent models have been applicable in simple polyphonies. One of the promising approach is based on peripheral perception and it uses a multi-channel filter bank with periodicity detection in each channel [20, 21, 22, 7]. Recently, a computationally efficient two-channel model has been presented that in many cases produces similar results to the multi-channel case [23].

The main attraction of periodicity detectors is that the auditory system often fuses periodic or nearly periodic signal components into a single percept. This is particularly important with musical instruments many of which produce harmonic tones. It is hard to imagine, e.g., a musical scene analysis system without pitch or periodicity detection at some stage. Periodicity representations are also useful when combined with sinusoidal modeling since these two together provide a means for separation of signal components corresponding to harmonic tones. Examples of this are presented in Section 3.

Onset/Offset Representations The auditory system is perceptive to rapid changes in the auditory scene [24]. Thus, an onset/offset detector is a typical component of an analysis system, particularly in an CASA system [5, 6, 7]. Such a representation is typically based on changes in the signal energy in different channels. Onset/offset representation is useful when combined with the periodicity representation so that the perceptual sound events can be identified more accurately. However, the system presented here does not currently employ an explicit onset/offset detector. As explained in Section 2, the onsets and offsets are handled through the multipitch representation.

Common Modulations Yet another means for identification of perceptual events and association of low-level components such as sinusoids is detection of common periodicity and amplitude modulations [25, 26, 27, 28]. Also this detector has a strong background in human audition [29]. Common modulation detection is useful when combined with periodicity and onset/offset detectors in identifying perceptual objects.

Model-Based Synthesis Model-based synthesis consists of numerous methods that simulate the sound production mechanisms of various musical instruments and of the human voice production. The synthesis models are computational algorithms that are executed at the sound signal generation block of the framework. Their control is typically obtained from the higher level and the control information is presented as events. Recent overviews of physical modeling and particularly digital waveguides, which are particularly popular for sound synthesis, are presented in [30, 31]. Section 3 shows a digital-waveguide-derived model applied in analysis/synthesis of acoustic guitar signals.

While the environment and receiver modeling is important for generation of high-quality virtual acoustics, it is beyond the scope of this work. More information on 3-D sound and virtual acoustics can be found, e.g., in [32, 33].

1.2 Object-Based Representations

The sound object modeling block of Figure 1 can be divided into three groups of techniques that are applicable in different applications, namely, perceptual modeling, musical object analysis, and sound source modeling. While the first corresponds to the perceptual aspects of the framework, the latter two reflect more the engineering side: solving practical audio and music applications with the help of signal and auditory analysis.

Perceptual Modeling Perceptual modeling is a set of techniques that analyze signal and auditory representations into perceptually relevant entities. Computational auditory scene analysis is a typical example of these applications [24, 25, 6, 7]. The goal of perceptual modeling is to computationally identify and describe auditory percepts, thus simulating the human auditory system. Note that perceptual objects do not always correspond to musical objects [8, 26, 27].

Musical Object Analysis In certain applications, the interest is in obtaining an object-based representation that reflects the musical content, e.g., in terms of common notation. These applications include automatic transcription (cf. a review in [34]) in which the aim is to obtain a score from a recording, and musical information retrieval [35, 36] where, e.g., a song or a melody is being identified. Another application in this category is tempo and beat tracking [37].

Musical objects do not always correspond to perceptual objects, although often the correlation is strong [8, 26, 27]. Thus, application such as automatic transcription differs from perceptual modeling since humans do not perceive common notation. However, such "engineering" problems can often benefit from the theory of perception, as proven in, e.g., audio coding [4]. Similarly, it is likely that musical object analysis will improve through inclusion of auditorily motivated processing principles.

Sound Source Modeling In sound source modeling the goal is to obtain a representation of a sound in terms of models of the sources that produced the sound. Depending on the application, sound source modeling techniques may be closely related to the synthesis side of the framework. For instance, the generalized audio coding paradigm allows to represent individual sound sources as they are presented with synthesis models in the synthesis path [38]. Another example is sound source identification where the aim is to identify a musical instrument from a recording of the sound of the instrument [8, 39].

Sound-Object Generation The object-based representation for synthesis typically consists of information about sound generation models and their dynamic control. This involves specification of properties of a sound generation model and event control by passing note-on and note-off events as well as other control data. A widely spread but limited protocol for sound-object control is MIDI [40]. Another example is the structured audio part of MPEG-4 that has been derived from the CSound language [1, 3]. This paradigm also hosts the generalized audio coding concept.

1.3 Content-Based Representation

At this time, content-based audio analysis is limited. In a primitive sense, applications such as automatic transcription and musical information retrieval can be interpreted as approaching content-based processing. On the synthesis side, the chain from a content-based representation into an audio signal is more developed. An application example is the expressive notation package (ENP) which allows a composer or performer to add expressive information related to the performance with a particular virtual instrument into an extended common music notation [41]. Laurson et al. have demonstrated how the ENP notation can be processed into an object-level control stream that is used to control a virtual guitar at the sound signal generation level (see [41] for examples, links also available at [9]). See also [42, Part V] for literature review of high-level control of music synthesis.

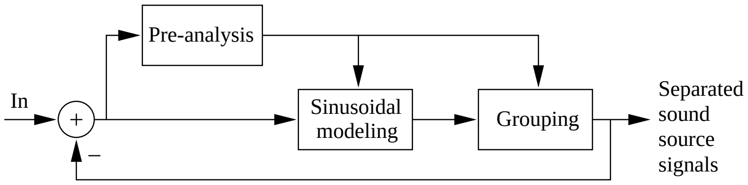

Figure 2: Block diagram of an analysis system based on iterative sound source identification and separation.

2 Analysis Techniques for Sound Source Modeling

Depending on the application, the signal and auditory analysis techniques and perceptual sound source modeling can consist of a multitude of methods, as discussed in Section 1. As an example of how the signal and auditory analysis methods can be combined with perceptual sound source modeling including the feedback paths (cf. Figure 1), an iterative sound source separation system is presented. Figure 2 shows a general block diagram of the system. The pre-analysis block extracts information related to pitched sound objects, such as fundamental frequency trajectories, onsets/offsets, and common modulations, as described in Section 1. These representations are used on one hand to help the extraction of low level components (sinusoids in this case) and to group the components into perceptually relevant entities. The feedback path is included so that a detected prominent components can be extracted either from the signal or from the low-level representations.

In the following we briefly summarize a previously presented system of iterative sound source separation based on multi-pitch analysis and identification of pitched sound components that follows the structure of Figure 2 [43]. A heuristic sound-source identification technique is described, and a previously reported technique for separation of colliding sinusoids is implemented in a computationally efficient way. A method for fine-tuning the estimate of fundamental frequency is described and it is applied in detection of inharmonicity and vibrato in mixtures of tones. Finally, techniques are described for obtaining a model-based representation using a plucked string model described in Section 2.4.

2.1 Iterative Identification and Separation of Harmonic Tones

The key component of the separation system is a computationally efficient multi-periodicity model [23]. The model has also been incorporated in an iterative multi-pitch analysis and prediction system for separation of speech signals [44]. Figure 3 shows a block diagram of the model. The input signal is first pre-whitened using the warped linear prediction (WLP) technique [45, 46]. The signal is divided into two channels below and above 1 kHz. The amplitude envelope of the high-channel signal is detected using half-wave rectification and low-pass filtering. A periodicity representation is computed on both channels by means of an autocorrelation with magnitude compression in the frequency domain. The periodicity representations are summed into a summary autocorrelation function (SACF). Finally, the SACF representation is post-processed into an enhanced SACF (ESACF) in which pitched sound components may be identified.

Figure 4 shows an example of the ESACF representation. The analyzed signal is a violin

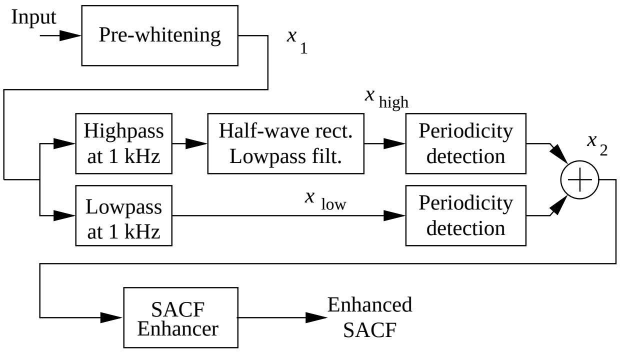

Figure 3: Block diagram of the multi-periodicity analysis model.

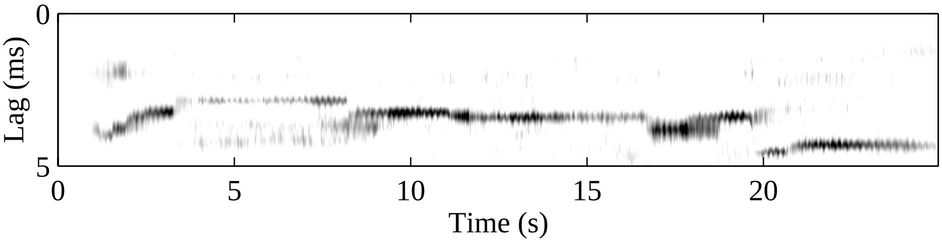

Figure 4: ESACF representation of a violin melody in an orchestral background.

melody in an orchestral background [9, Example 1]. As seen in the figure, the melody is clearly visually identifiable in the representation. However, the plot also shows spurious peaks that do not correspond to pitched sound objects. Thus, we present a heuristic event detector for identification of pitched sound objects which helps to automate the crucial identification step in the separation process. The proposed event detector also sorts the event candidates according to a relevance measure.

The technique for detecting pitched sound objects is based on trajectory tracking. A similar technique is frequently used in sinusoidal modeling [47]. Starting from the first frame, all local maxima are detected. After a maximum has been detected, the corresponding peak is deleted in the representation [48]. For each detected maximum, a corresponding maximum is sought in the consecutive frames in a vicinity of the location of the original maximum. This search is continued until all maxima are assigned to a trajectory. These trajectories are treated as pitched sound event candidates.

In order to sort the events, a relevancy measure is attached to each event. The measure is computed as follows. The sum of the ESACF values of each frame are computed for normalization. Each of the values of the maxima are normalized with the corresponding ESACF frame sum. The normalized maxima values are summed over each event. This kind of heuristic measure was selected so that the relative weight of each ESACF frame is constant (normalization) and the events that last long in time are weighted more than short ones (summing over events). A similar approach of favoring long events was taken in detecting sinusoids in consecutive frames [17]. Naturally, this is but one alternative for sorting the events but this method has been useful in separating the spurious short events from events corresponding to actual tones.

The solid lines in Figure 5 (a) show the events corresponding to the violin melody of Figure 4. The event detector was in this case able to follow the melody with only a few discontinuities

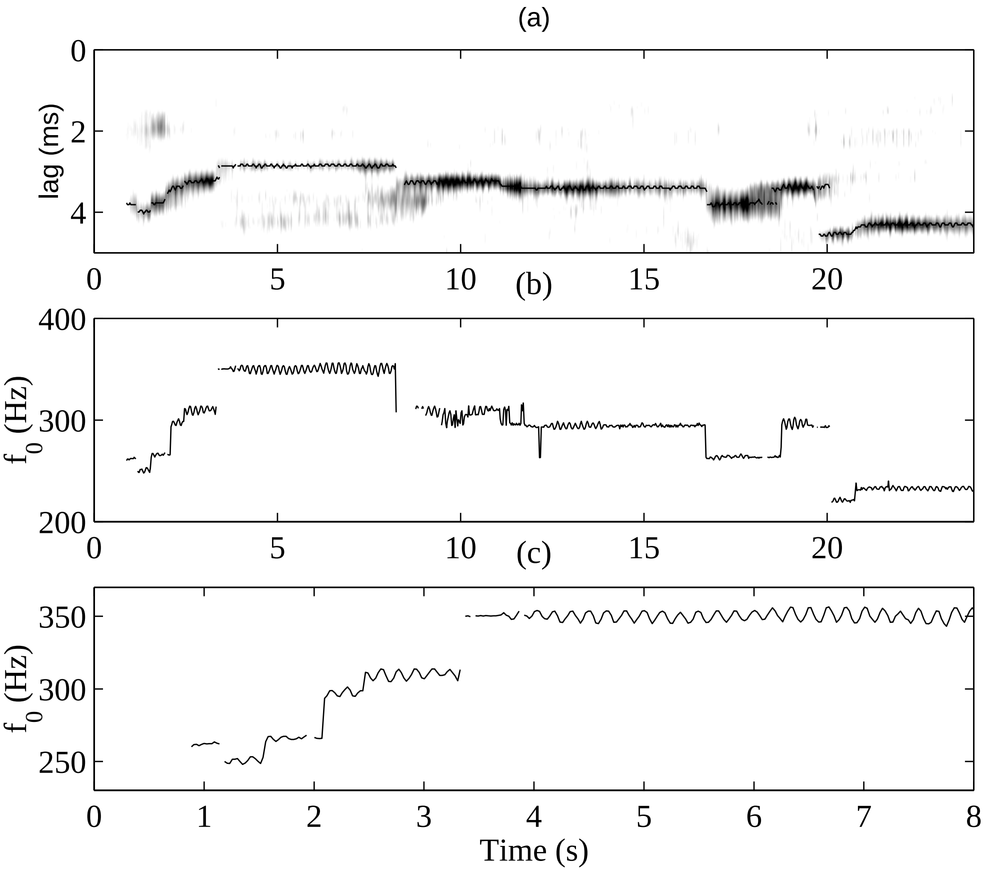

Figure 5: An example of detection of vibrato.

with events. These, however, may be problematic and a post-processing step for merging discontinuities that are likely to correspond to a single event may be useful. Similar merging techniques have been used in sinusoidal modeling [47].

In many cases, the separation of a complex sound signal into its all constituent components at the signal level is not possible or the result is of poor perceptual quality (cf. [9, Examples 4–6]). Note that this kind of functionality is very unlikely in the human auditory system. However, there are applications where the separation of prominent contributors, e.g. a melody, is sufficient. Such applications include musical information retrieval, e.g., identification of a song based on melody, and melody sound-source identification. Reasonable tasks at this time may include the following: Is there a prominent plucked string instrument in this segment? Name a likely prominent instrument in this segment. Even these simple applications still require extensive development. In addition to these examples, the sinusoidal separation approach is useful in obtaining a model-based representation from recordings, as described in Section 3.

The examples that follow show how the ESACF representation combined with a refining estimator of fundamental frequency can be used to detect vibrato and inharmonicity of a tone in a mixture of tones, and to obtain a model-based representation using a plucked-string model.

2.2 Computationally Efficient Implementation of NLS Sinusoidal Modeling

In a previous work, a technique for separation of colliding sinusoids was presented based on the *nonlinear least-squares* (NLS) method in a relatively small vicinity of the frequency space [43]. The NLS technique is the most accurate (minimum-variance) unbiased method for estimating sinusoids in additive Gaussian white noise [49, 50, 51, 52].

The basic idea is to apply the estimator locally in a vicinity that is pre-determined from analysis of the fundamental frequencies or from investigation of the significance measure. Global application of the estimator is infeasible since that would involve a highly nonlinear search over a high-dimensional parameter space. In this local application, the parameter space is essentially two-dimensional and the search space may be defined in advance for faster convergence. Furthermore, the estimates of the fundamental frequencies may provide initial values for the search algorithm.

As described in [43], the cost function of the local model given as

$$G(f,a,\phi) = \sum_{n=0}^{N-1} \left| y(n) - \sum_{k=1}^{2} a_k e^{i(2\pi f_k n + \phi_k)} \right|^2, \ f_1 \in [f_{1,\min}, f_{1,\max}],\ f_2 \in [f_{2,\min}, f_{2,\max}] \tag{1}$$

where the ranges $[f_{1,\min}, f_{1,\max}]$ and $[f_{2,\min}, f_{2,\max}]$ are pre-determined from the estimates of the corresponding fundamental frequencies or from the shape of the peak.

As shown in, e.g., [51], the cost function of Equation 1 is minimized with the following separated equations

$$\begin{aligned} \hat{f} &= \arg\max_f [Y^H B (B^H B)^{-1} B^H Y] \\ \hat{\beta} &= (B^H B)^{-1} B^H Y|_{f=\hat{f}} \end{aligned} \tag{2}$$

where

$$\begin{aligned} \beta_k &= a_k e^{i\phi_k} \\ \beta &= [\beta_1\ \beta_2]^T \\ Y &= [y(0) \cdots y(N-1)]^T \\ B &= \begin{bmatrix} 1 & 1 \\ e^{i2\pi f_1} & e^{i2\pi f_2} \\ \vdots & \vdots \\ e^{i2\pi (N-1) f_1} & e^{i2\pi (N-1) f_2} \end{bmatrix} \end{aligned}$$

The sinusoidal frequencies are obtained using the first equation of (2) after which the amplitudes and initial phases are computed using the second equation of (2).

Direct application of Equation 2 is unattractive due to the computational demands. Particularly, computation of $[Y^H B (B^H B)^{-1} B^H Y]$ is expensive. However, examination of the term $(B^H B)^{-1}$ reveals that it can be expressed as a function of the difference of the frequencies

f_1 and f_2:

$$
\begin{aligned}
(B^H B)^{-1} &= \begin{bmatrix} (e^{i2\pi f_1 \mathbf{n}})^H e^{i2\pi f_1 \mathbf{n}} & (e^{i2\pi f_1 \mathbf{n}})^H e^{i2\pi f_2 \mathbf{n}} \\ (e^{i2\pi f_2 \mathbf{n}})^H e^{i2\pi f_1 \mathbf{n}} & (e^{i2\pi f_2 \mathbf{n}})^H e^{i2\pi f_2 \mathbf{n}} \end{bmatrix}^{-1} && (3) \\
&= \begin{bmatrix} N & \sum_{n=0}^{N-1} e^{i2\pi(f_2-f_1)n} \\ \sum_{n=0}^{N-1} e^{-i2\pi(f_2-f_1)n} & N \end{bmatrix}^{-1} && (4)
\end{aligned}
$$

The values of the term $(B^H B)^{-1}$ can be stored in memory with appropriate resolution and range of the difference $f_2 - f_1$. In addition, the values of $Y^H B$ can be stored during the computation so that each vector product is only computed once. These procedures have resulted in significant computational savings making the NLS technique feasible in practical analysis applications.

2.3 NLS-Based Estimation of Fundamental Frequency

The NLS estimator is readily employed in refining an estimate of the fundamental frequency of a harmonic tone. In this case the cost function to be minimized is

$$
G(f_0) = \sum_{n=0}^{N-1} \left| y(n) - \sum_{k=1}^{N_{\text{harm}}} \beta_k e^{i(2\pi k f_0 n)} \right|^2 \tag{5}
$$

over a meaningful range of the fundamental frequency f_0. Parameters N and N_{harm} are the window length and the number of harmonics, respectively.

Equation 5 can be solved using Equation 2 where β now consist of N_{harm} complex amplitudes and the B matrix of N_{harm} complex sinusoids. Notice that computational complexity of the method can be considerably larger compared to detection of two complex sinusoids depending on the number of harmonics used in detection. Particularly with low tones with numerous harmonics in the signal band, it is advisable to limit the number of harmonics used in estimation. If computational efficiency is of the essence, the Goertzel algorithm [53, 54] can be used to reduce the computational requirements. In addition, the values of $(B^H B)^{-1}$ can be precomputed and stored in a memory with a sufficient f_0 range and resolution.

This method has been applied in separation of speech signals for refining the estimate of the fundamental frequency [44]. The NLS fundamental frequency estimator may be readily included in the pre-analysis block of Figure 2.

Application of the fundamental frequency estimator in detection of vibrato and tremolo and of inharmonicity parameters is described next.

2.3.1 Detection of Vibrato

Vibrato is a perceptually important expressive feature related to variation of the fundamental frequency. It is controlled by the singer or the player of the instrument. Vibrato and its characteristics have also been found useful for instrument identification [8].

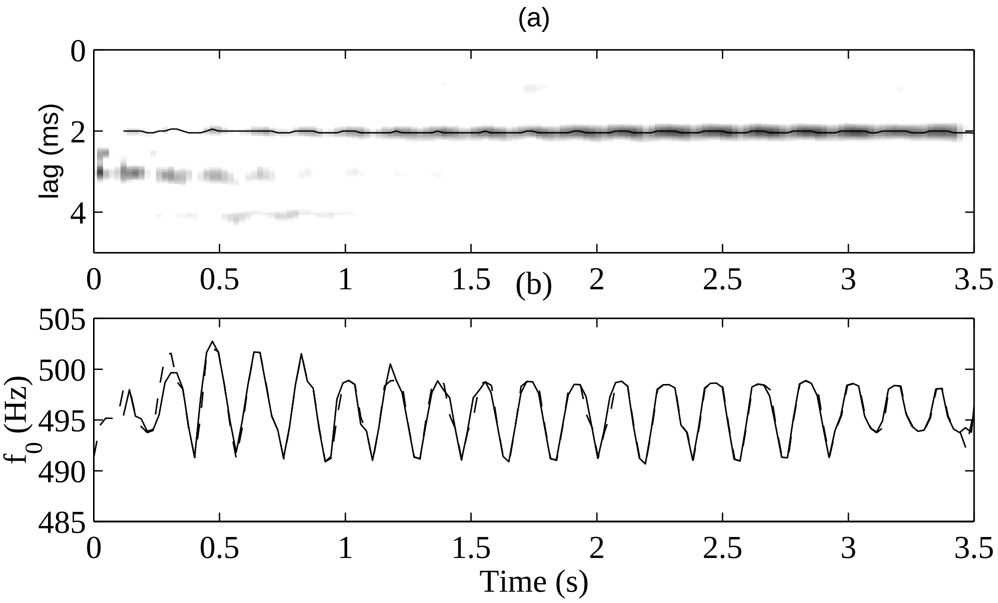

Figure 6: (a): ESACF representation of a mixture of piano and flute tones. (b): detected fundamental frequency trajectories of the mixture (solid) and the isolated tone (dashed).

The NLS fundamental frequency analysis model provides an f_0 trajectory on which the vibrato can be estimated. An example of detection of vibrato is demonstrated in Figure 5. The test signal is an excerpt of classical music with a violin solo and an orchestral background. Figure (a) shows the ESACF representation of the excerpt. The solid lines mark the detected events that correspond to the violin melody (cf. the sound signal at [9, Example 1]). Plot (b) shows the f_0 trajectory obtained using Equation 5 using the fundamental periods of the events in (a) as initial f_0 estimates. Figure (c) shows a more detailed view of the first eight seconds of fundamental frequency trajectory. The modulation of fundamental frequency is clearly observable in the plot.

As noted, e.g., in [8, 55], in many cases amplitude modulation of partials is related to the fundamental frequency modulation. Common period and amplitude modulations of partial components also appear to provide a means for segregation of individual sources [24]. This property has been applied for a computational model for segregation of sound sources components [26, 27]. As Figures 6 and 7 show, the NLS-based fundamental frequency estimator provides an alternative way of detection of common period and amplitude modulations. In this case the test signal has been composed of tones from the McGill University Master Samples set [56] (cf. [9, Example 2]). A C major seventh chord is obtained by summing three piano tones (C_3, E_4, and G_4), and a flute tone B_4 with vibrato. The period modulation and the amplitude modulation of the first five harmonics of the flute tone was analyzed. Figure 6 (a) shows the ESACF representation of the signal. The detected event corresponding to the flute tone is depicted with a solid line. The solid and dashed plots in (b) show the fundamental frequency trajectories detected in the mixture of tones and in the isolated tone, respectively. The plots suggest that the fundamental frequency variation can be detected in a mixture of tones, as long as the desired event can be identified in an ESACF or a similar representation so that the initial estimate used in the NLS-method is sufficiently accurate.

The NLS estimator also tracks the amplitudes of individual harmonics, as demonstrated in

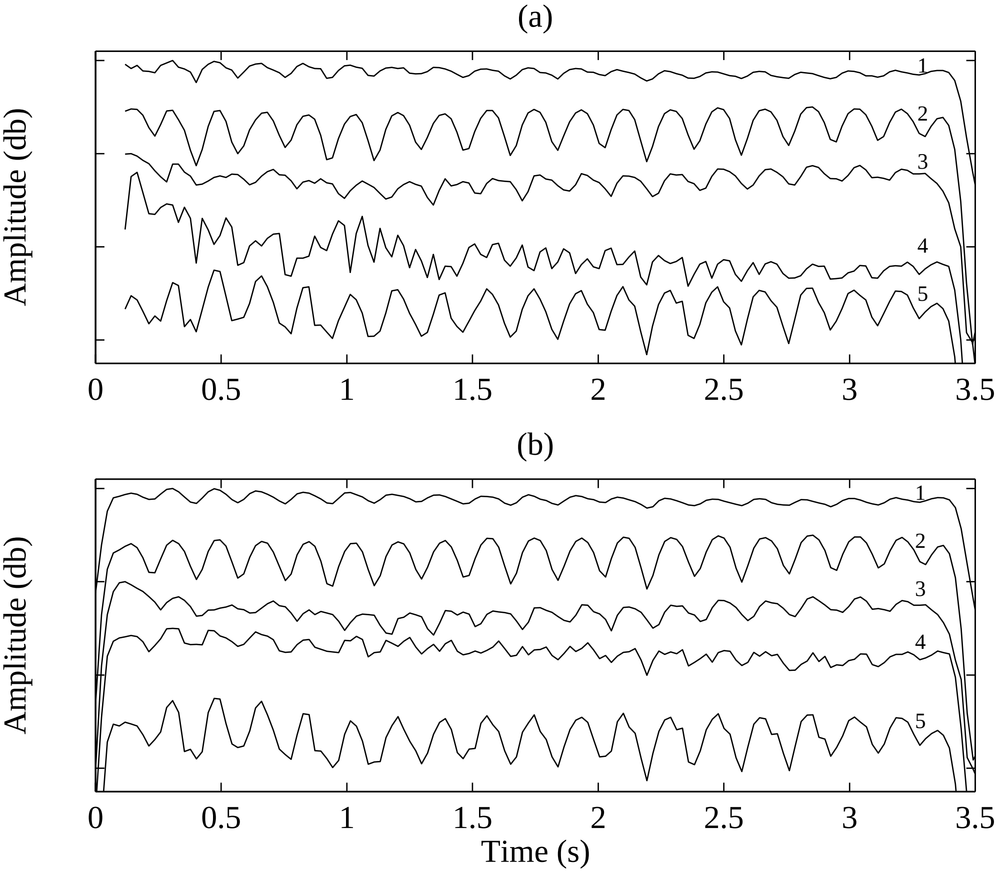

Figure 7: Amplitude envelopes of the five lowest partials of the flute tone in mixture (a) and in isolation (b).

Figure 7 which shows the amplitude envelopes of the five lowest partials of the flute tone in the mixture (a) and in isolation (b). The envelopes are positioned on the figure manually so that they can be easily visually examined; the figure does not reflect the relative levels of amplitudes of partials. Partials 1, 2, 3, and 5 exhibit amplitude modulation that is related to the period modulation (cf. Fig. 6). Only the envelope of the fourth partial differs considerably in the two cases.

Vibrato is an expressive feature with characteristics that are controlled by the player, including the frequency and the amplitude of the modulation. In [41, 57], the vibrato characteristic of isolated classical acoustic guitar tones were analyzed and the data was used to obtain parameters for model-based synthesis. The use of the NLS-based fundamental frequency estimator allows detection of the vibrato in real musical performances. It also appears an attractive front-end tool for investigation of fine-structure of vibrato in actual performances and provides a feature extractor for music recognition models presented in Figure 1.

A recent study of sound-source recognition proposes to use several features related to spectral structure, pitch, vibrato, tremolo (over-all amplitude modulation), and inharmonicity [8]. Although application of the NLS-based techniques to sound-source recognition is beyond the scope of this work, it appears promising for detection of such features in mixtures of simple polyphonic signals or in signals with a prominent source, e.g., an instrument playing a melody.

Another phenomenon that causes variation of the fundamental frequency is the tension modulation: a freely vibrating string exhibits a nonlinearity that causes the pitch of the perceived tone to vary during vibration. The nonlinearity is generated by tension modulation that is related to elongation of the string during vibration. For a review of analytic and experimental studies on tension modulation as well as methods for simulating the phenomenon using computational models, see [58].

The fundamental frequency variation takes place in every vibrating string but is typically perceivable only in plucked strings where the overall displacement distribution variation is larger compared to, e.g., struck strings. Moreover, the effect is large if the initial tension of the string is relatively small allowing large-amplitude vibration. Instruments with perceptually relevant tension modulation include guitars, the tanbur, and the kantele, which are traditional Turkish and Finnish string instruments, respectively. Examples of fundamental frequency variation in these instruments are presented in [58, 59, 60].

The NLS fundamental frequency estimator can be used to detect the pitch drift caused by tension modulation. Since this pitch drift is related to the attenuation of the tone, it may be a useful feature, e.g., in recognition of musical instruments. Note, however, that the phenomenon is only observable in tones that are plucked relatively hard and at a position with a considerable distance from the string termination for a sufficiently large initial displacement distribution.

2.3.2 Model for Inharmonicity

The NLS fundamental frequency estimator can also be applied for detection of inharmonicity parameters in tones. More precisely, a model of inharmonicity can be incorporated in the estimation. For instance, the frequencies of partials of a freely vibrating stiff string are given as [61]

$$f_k = kf_0(1+Bk^2)^{1/2} \tag{6}$$

where f_0 is the frequency of the first partial, and B is the inharmonicity coefficient given by

$$B = \frac{\pi^3 ES^4}{64\ell^2 F_\mathrm{t}},$$

where E is the Young's modulus, S is the string diameter, ℓ is the string length, and F_t is the tension of the string.

Using Equations 5 and 6, the cost function for detection of inharmonicity is given by

$$G(f_0) = \sum_{n=0}^{N-1} \left| y(n) - \sum_{k=1}^{N_\mathrm{harm}} a_k e^{i(2\pi k f_0 (1+Bk^2) n)} \right|^2 \tag{7}$$

The following example demonstrates how the inharmonicity coefficient of a piano tone can be estimated in a mixture of three tones. The tones are from the McGill University Master Samples set [56] (cf. [9, Example 3]). The instrument was 9' Hamburg Steinway, that was played loud (volume 3, track 2). The selected tones were C$_3$, E$_4$, and G$_5$. The tones were selected so that there were several partials of the tones colliding in narrow frequency bands. The tones were manually set to start approximately simultaneously, and the tones were summed without altering the amplitude ratios.

The inharmonicity coefficient B of the tone C$_3$ was estimated using Equation 7 both from the mixture and the C$_3$ tone without mixing. The number of harmonics used in analysis was limited to 30 for computational reasons. This limit has also a perceptual motivation based on a study where the bandwidth of perceived inharmonicity of piano tones was studied [62, 63]. That study suggested that in synthesis of the piano tone C$_3$, it suffices to capture the inharmonicity of the first 30 partials accurately.

A 740-ms segment was selected after the attack of both signals. An exhaustive search was performed over the two parameters f_0 and B. An initial estimate of f_0 was obtained through investigation of the DFT. The f_0 range was set around this value. The grid of the B parameter values was set on a logarithmic scale between $B = 10^-6$, $B = 10^-3$. Also a value $B = 0$ was included in the estimation corresponding to a perfectly harmonic tone.

Table 1 shows the results of the estimation. It is seen that the estimate of B in the mixed case is quite close to that estimated in the C$_3$ tone directly. Notice also that there is a small deviation in the value of f_0 estimated using the model of Equation 7 and the value that was obtained initially using the DFT. From a computational efficiency viewpoint, it would be attractive to use the initial estimate as fixed f_0 and perform only a one-dimensional search over the B parameter. Although the deviation in f_0 estimates is small, it is large enough to degrade the estimation procedure if only the B parameter value is estimated. In fact, in the case of the present example, no local maxima was found in the the one-dimensional cost function when the f_0 was fixed to the initial estimate.

signal	B	f_0	$f_{0,\mathrm{in}}$
C$_3$	$1.04 * 10^{-4}$	130.6	130.4
mixed	$1.05 * 10^{-4}$	130.6	130.4

Table 1: Inharmonicity coefficient estimation example.

Figure 8 shows the spectra of the mixture of piano tones (a) and of the isolated tone C$_3$. The dotted vertical lines show the frequencies of the 30 lowest partials according to the detected values of the model in Equation 6.

The presented example and other similar test cases performed during this study suggest that the NLS-based inharmonicity estimator is a useful tool for analysis of musical sound sources. However, experiments with a wider variety of test signals need to performed for gaining insight into the limitations of the method. One clear drawback is the computational complexity of the method particularly when an exhaustive search is applied. Further analysis of the estimator performance and cost functions computed on realistic musical signals may justify the use of more efficient numerical methods.

An application where an inharmonicity estimator, such as the one presented here, is useful is determination of the perceptual relevance or audibility of inharmonicity of tones. Recently, listening experiments have been conducted for determining the audibility thresholds of inharmonicity of plucked string tones [64]. The listening experiment results and the inharmonicity estimator can be integrated into an analysis tool for this perceptual feature.

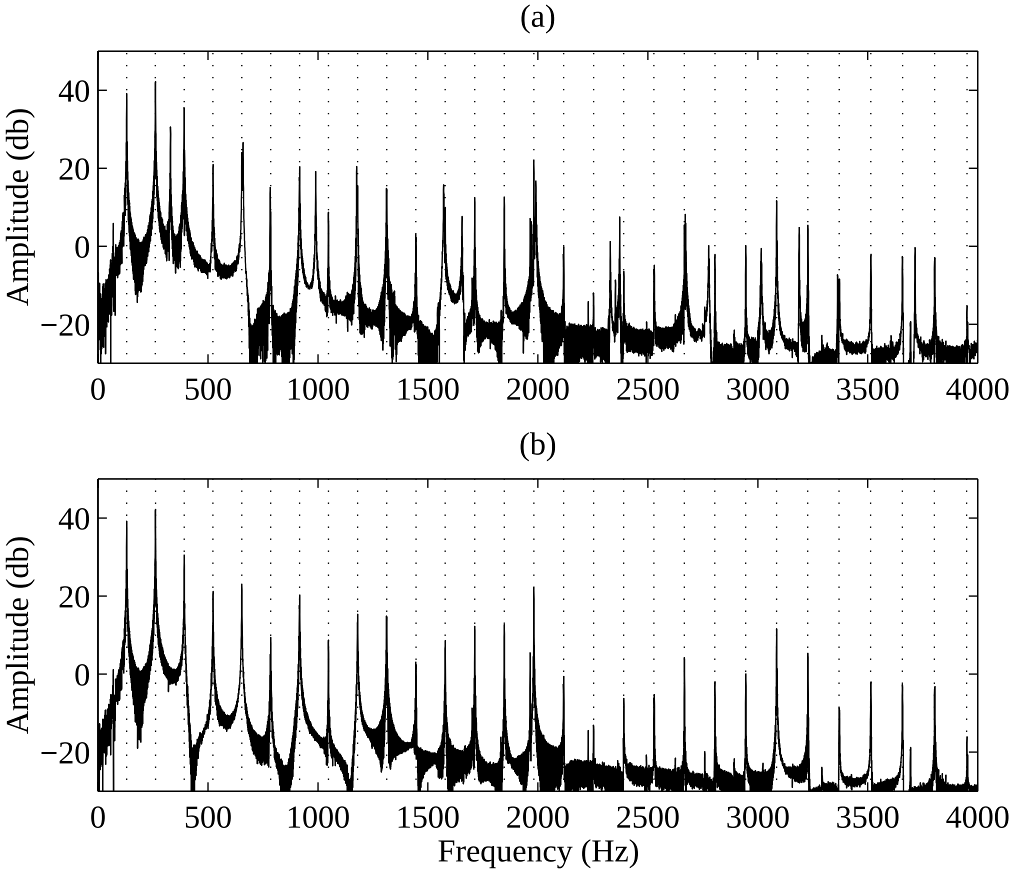

Figure 8: Spectra of the mixture of piano tones (a) and the isolated tone C_3 (b). The dotted lines indicate the estimated frequencies of the inharmonic partials.

As of now, no results were found in the literature for just noticeable difference of inharmonicity related to a model similar to that of Equation 6. Such results would be useful for setting the resolution of the B parameter in a perceptually motivated way, similarly to the decay characteristic of plucked-string tones described in a companion paper [65]. Another future direction is to incorporate the results of a listening experiment on the effect of inharmonicity to the pitch of string instrument sounds [66] into the perceptual analysis tool.

2.4 Model-Based Representation

Besides separation of sound source signals and estimation of perceptual features, the presented analysis methods can also be applied in obtaining a model-based representation of a sound signal. Next, we briefly describe a physics-based synthesis model of a plucked-string instrument and its parameter estimation. In the following section, we apply the model in an analysis/synthesis task.

2.4.1 Model-Based Plucked-String Synthesis

A block diagram of the string model is presented in Figure 9. The model is derived from a bi-directional digital waveguide [67, 68, 69], and it uses the method of commuted waveguide

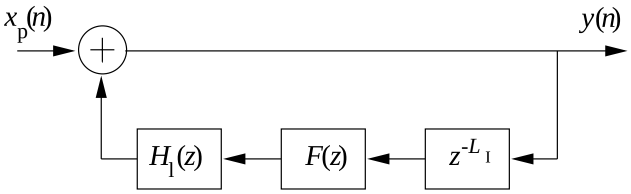

Figure 9: A block diagram of the string model [73].

synthesis (CWS) [70, 71]. Derivation of the model of Figure 9 from a digital waveguide model is presented in [72].

The transfer function for the string is

$$S(z) = \frac{1}{1 - z^{L_\mathrm{I}} F(z) H_\mathrm{l}(z)}, \tag{8}$$

where L_I is the length of the delay line,

$$H_\mathrm{l}(z) = \frac{g(1-a)}{1 - az^{-1}} \tag{9}$$

is the one-pole lowpass loop filter which determines the decay of the tone, and $F(z)$ a fractional delay filter implementing the non-integer part of the string length [74, 75]. The string transfer function $S(z)$ is fully described by the string length L in samples, the loop gain g and the loop filter cutoff parameter a.

The model of Equation 8 can be used for synthesis of high-quality tones when the CWS technique is employed. In commuted synthesis, the string model parameters are calibrated based on analysis of recorded tones [73, 76, 57]. After parameter calibration, the inverse of the model in Equation 8 is used to inverse-filter the recorded tones. If the calibration is done properly, the residual of the inverse-filtering is a relatively short signal that consists of the contributions of the pluck and the body response. When this excitation is used in synthesis, an identical copy to the original is obtained. The excitation signals are typically windowed into a length of approximately several hundreds of milliseconds in order to save memory. Other methods of reducing the length of the excitation signal include modeling of the signal with a digital filter and the use of separate parametric models for the most prominent body resonances [77, 76, 60]. Sound examples of synthetic guitar tones are available at [9].

2.4.2 Estimation of Model Parameters

The methods presented above can also be applied in obtaining an object-based representation of a polyphonic guitar signal based on the model of Figure 9. Next, we describe how its parameters can be obtained in the case of polyphonic signals.

The ESACF representation combined with the NLS fundamental-frequency refiner may be directly applied to obtain a fundamental-frequency trajectory from which the fundamental frequency can be estimated. If desired, the NLS-refiner may be applied in a longer time window to obtain a long-term estimate of f_0. Thus, the main problem is in detection of parameters g and a.

In previous studies, the decay parameters have been obtained from recordings of isolated tones using a sinusoidal modeling approach [73, 76, 57]. The sinusoidal models provide amplitude envelopes of the decaying partials from which the decay time constants are estimated. A digital filter is designed so that the synthetic tone produces an optimal decay. A similar approach can be applied directly also in the polyphonic case. However, now the sinusoidal modeling step is more demanding requiring identification of the components corresponding to the desired tone. Another problem is reverberation that is typically present in actual recordings. Estimating a decay characteristic of a tone with reverberation is prone to more errors than the anechoic case. The following section includes examples with synthetic signals with and without reverberation and a recorded anechoic signal. These examples give an idea how the reverberation may affect the synthesis results.

A companion paper describes listening experiments conducted to investigate the perception of changes in decay of synthetic plucked string tones [65]. These results are valuable also here since they give perceptual thresholds for the decay parameters g and a. The paper also describes how the listening experiment results can be used in an iterative parameter estimation method proposed in [57].

The companion paper also suggests an alternative method for parameterizing the decay characteristic of the string model in terms of overall decay time constant and the cut-off frequency of the decay. This parameterization is more general and descriptive than using directly the values of the parameters. In addition, as described in [78], this approach is justified from the perceptual viewpoint since the perceptual thresholds can be easily expressed in this parameterization.

The plucked-string model used here as a case example is a simple linear model with straightforward parameter estimation methods. However, many model-based synthesis techniques are nonlinear which makes their parameter estimation more elaborate. Drioli and Rocchesso describe a nonlinear predictor and synthesis model where the nonlinear part is realized with a neural network [79]. The model learns its parameters with nonlinear identification procedures based on waveform or spectral matching. Although this work is a promising step, in order for the nonlinear models to be applicable in analysis/synthesis systems, techniques for parameter estimation need to developed further.

3 Model-Based Analysis and Synthesis of the Acoustic Guitar

Model-based synthesis of acoustics guitar tones has been an active area of research for the last two decades. Combined with techniques for obtaining model parameters from recordings, this approach has resulted in realistic high-quality synthetic guitar pieces [80, 41, links to demos: [9]].

Many synthesis examples have been based on the CWS technique [70, 71], and the model parameters have been obtained from anechoic recordings of individual guitar tones [73, 76, 57]. With CWS, the synthetic tones can be copies of the original within the limits of numerical accuracy. However, typically the parameters need to be fine-tuned according to musical

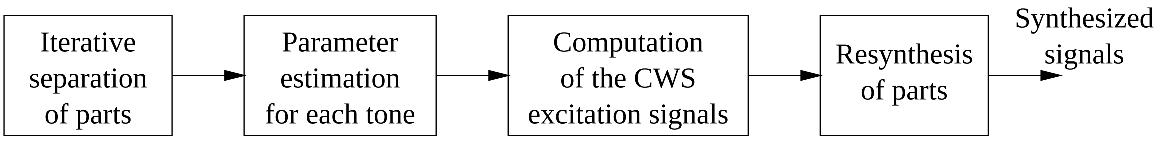
Iterative separation of parts
Parameter estimation for each tone
Computation of the CWS excitation signals
Resynthesis of parts
Synthesized signals

Figure 10: Block diagram of analysis/synthesis of guitar tones.

context so that naturalness of the virtual performance is increased. Another problem is that the models require through analysis of a particular instrument with preferably an anechoic chamber recordings.

In the following, we apply the sound source separation system discussed in Section 2 to obtain a model-based representation of simple two-voice guitar music and resynthesize the excerpt using the CWS-approach and the model of Equation 8. Three versions of the two-voice excerpt are analyzed and resynthesized: dry synthetic signal, synthetic signal with reverberation, and real recording in an anechoic chamber.

The selected examples are simple two-voice polyphonies with a long tone with low fundamental frequency and a six-tone melody at higher frequencies. We are not aiming to provide a comprehensive treatment of the model-based sound processing; rather, the objective here is to show what kind of results can be expected with the model-based coding approach, and describe the shortcomings that require further development. While this particular example is limited to classical guitar tones, it should at least qualitatively give insight into how this approach is suited to other plucked string instruments.

Figure 10 shows a block diagram of the analysis/synthesis process. The sound source components are first iteratively separated using the iterative technique described in Section 2. The model parameters are estimated from the sinusoidal representation. The loop filter is matched to the decay of amplitude envelopes and the delay line length is selected so that the desired fundamental frequency is obtained. In the third stage, the CWS excitation signals are obtained using the estimated model parameters. The excitation signals are windowed into short 100 ms signals. So the representation of each tone consists of the attack and damp locations, the excitation signal and the model parameters. Finally, the synthetic parts are computed from the representations and summed into a synthetic version of the original signal.

The original, intermediate, and synthetic signals of all three cases are available at [9, Examples 4–6]. The synthesis results show that main features of the parts are captured in all cases but fine details related to the attack and time-variation of the timbre are either missing or degraded.

The analysis/synthesis approach works best on the synthetic signals. This was expected since the original signal is produced with a similar model. The reverberation caused problems in that the melody tones had a quite large variation in timbre from tone to tone. In the non-reverberant case the variation was not so pronounced. The main problems with the recorded excerpt was related to the attacks of the synthetic melody tones which exhibited clearly audible artifacts. This may be due both to inaccuracies in sinusoidal modeling in separation and to inaccurate estimation of the string model parameters. All the examples retained the identity of the instrument, although the quality was degraded.

In some applications a more attractive approach may be to not transmit the excitation signals but to use a completely different set of signals that are made available at the synthesis stage. This naturally reduces the bandwidth required for transmission considerably as all the tones are represented by merely attack and damp locations, overall amplitude, and the model parameters. Note that this approach does not typically produce exactly similar synthetic signal, but in many applications it can still be useful.

4 Conclusions and Future Work

A framework for audio and music processing has been presented. The framework essentially consists of two vertical paths, one for analysis and another for synthesis, and three representational levels for parametric, object-based, and conceptual representations. Analysis methods were described for identification and separation of harmonic tones and for detecting perceptual features of tones, such as vibrato and inharmonicity. Finally, analysis/synthesis examples were presented using a model-based synthesis of acoustic guitar signals.

The presented analysis techniques combine signal and auditory analysis methods with perceptual sound source modeling in an iterative way. Although the presented work has but scratched the surface, the results support that further research with this approach is attractive. There are numerous ways the improve the presented system. Improvements may be expected with development of the multi-periodicity analysis model and integration of that with onset/offset and common modulation representations, improving the separation stage, and elaborating the models and their parameter estimation.

It will take a long time before the analysis and synthesis paths of the framework are generally connected at all levels. In the mean time, research and development within the areas of the framework will yield novel and useful techniques that enable new solutions, services, and applications in digital audio and multimedia communication.

Acknowledgments

The author wishes to thank Prof. Matti Karjalainen for many fruitful discussions and support throughout this work, Mikael Laurson for the synthetic test signals, and Cumhur Erkut for guitar model parameter estimation tools. The financial support of the GETA graduate school, Nokia Research Center, Jenny ja Antti Wihurin rahasto (Jenny and Antti Wihuri Foundation), Tekniikan edistämissäätiö, and Nokia Foundation is gratefully acknowledged.

References

[1] B. L. Vercoe, W. G. Gardner, and E. D. Scheirer, “Structured audio: creation, transmission, and rendering of parametric sound representations,” *Proceedings of the IEEE*, vol. 86, no. 5, 1998.

[2] “ISO/IEC IS 14496-3 Information Technology—Coding of Audiovisual Objects, Part 3: Audio,” 1999.

[3] E. D. Scheirer, Y. Lee, and J.-W. Yang, “Synthetic and SNHC audio in MPEG-4,” *Signal Processing: Image Communication*, vol. 15, pp. 445–461, 2000.

[4] N. Jayant, J. Johnston, and R. Safranek, “Signal compression based on models of human perception,” *Proc. IEEE*, vol. 81, pp. 1385–1422, Oct. 1993.

[5] D. K. Mellinger, *Event formation and separation in musical sound*. PhD thesis, CCRMA, Stanford University, Stanford, California, USA, 1991.

[6] G. J. Brown, *Computational auditory scene analysis: a representational approach*. PhD thesis, University of Sheffield, Sheffield, UK., 1992.

[7] D. P. W. Ellis, *Prediction-driven computational auditory scene analysis*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, jun 1996.

[8] K. D. Martin, *Sound-Source Recognition: A Theory and Computational Model*. PhD thesis, Massachusetts Institute of Technology, 1999.

[9] T. Tolonen, “Sound demonstrations for ‘Object-Based Sound Source Modeling for Musical Signals’, presented at AES 109th Convention.” `http://www.acoustics.hut.fi/~ttolonen/aes109/SSM/`.

[10] M. Karjalainen, “Immersion and content—a framework for audio research,” in *Proceedings of the IEEE Workshop of Applications of Signal Processing to Audio and Acoustics*, Oct. 1999.

[11] K. Brandenburg, O. Kunz, and A. Sugiyama, “MPEG-4 Natural Audio Coding,” *Signal Processing: Image Communication*, vol. 15, Jan. 2000.

[12] X. Serra, *A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic plus Stochastic Decomposition*. PhD thesis, Stanford University, California, USA, 1989.

[13] X. Serra and J. O. Smith, “Spectral modeling synthesis: a sound analysis/synthesis system based on a deterministic plus stochastic decomposition,” *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[14] T. F. Quatieri and R. J. McAulay, “Audio signal processing based on sinusoidal analysis/synthesis,” in *Applications of Digital Signal Processing to Audio and Acoustics* (M. Kahrs and K. Brandenburg, eds.), pp. 343–416, Boston, USA: Kluwer Academic Publishers, 1998.

[15] T. S. Verma, S. N. Levine, and T. H. Y. Meng, “Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals,” in *Proceedings of the International Computer Music Conference*, (Thessaloniki, Greece), pp. 164–167, Sept. 1997.

[16] T. Verma, *A Perceptually Based Audio Signal Model With Application to Scalable Audio Compression*. PhD thesis, Stanford University, 2000.

[17] S. Levine, *Audio Representations for Data Compression and Compressed Domain Processing*. PhD thesis, Stanford University, CCRMA, Stanford, CA, 1998.

[18] S. Levine and J. O. Smith, “A sines+transient+noise audio representation for data compression and time/pitch-scale modifications,” in *Proceedings of the 105th Convention of the Audio Engineering Society*, (New York), 1998. Preprint 4781.

[19] H. Purnhagen, B. Edler, and C. Ferekidis, “Object-based analysis/synthesis audio codec for very low bit rates,” in *Proceedings of the 104nd Convention of the Audio Engineering Society, Preprint 4747*, 1998. http://www.tnt.uni-hannover.de/project/coding/audio/asac/aes_104.html.

[20] R. Meddis and L. O'Mard, "A unitary model for pitch perception," *Journal of the Acoustical Society of America*, vol. 102, pp. 1811–1820, Sept. 1997.

[21] R. Meddis and M. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i: pitch identification," *Journal of the Acoustical Society of America*, vol. 89, pp. 2866–2882, June 1991.

[22] M. Slaney and R. F. Lyon, "A perceptual pitch detector," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 357–360, 1990.

[23] T. Tolonen and M. Karjalainen, "A computationally efficient multi-pitch analysis model," *IEEE Transactions on Speech and Audio Processing*, 1999. Accepted for publication.

[24] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, Massachusetts, USA: The MIT Press, 1990.

[25] M. P. Cooke, *Modeling auditory processing and organization*. PhD thesis, University of Sheffield, Sheffield, UK., 1991.

[26] E. D. Scheirer, "Towards music understanding without separation: segmenting music with correlogram comodulation," in *Proceedings of the IEEE Workshop of Applications of Signal Processing to Audio and Acoustics*, (New Paltz, New York), Oct. 1999.

[27] E. D. Scheirer, "Sound scene segmentation by dynamic detection of correlogram comodulation," Tech. Rep. 491, M.I.T Media Laboratory Perceptual Computing Section, Apr. 1999.

[28] E. D. Scheirer, *Music-Listening Systems*. PhD thesis, Massachusetts Institute of Technology, 2000.

[29] C. J. Darwin and R. P. Carlyon, "Auditory grouping," in *Hearing* (B. C. J. Moore, ed.), pp. 387–424, Academic Press, 1995.

[30] J. O. Smith, "Physical modeling synthesis update," *Computer Music Journal*, vol. 20, no. 2, pp. 44–56, 1996.

[31] J. O. Smith, "Principles of digital waveguide models of musical instruments," in *Applications of Digital Signal Processing to Audio and Acoustics* (M. Kahrs and K. Brandenburg, eds.), pp. 417–466, Boston, USA: Kluwer Academic Publishers, 1998.

[32] D. Begault, *3-D Sound for Virtual Reality and Multimedia*. Boston, USA: Academic Press, 1994.

[33] J. Huopaniemi, *Virtual Acoustics and 3-D Sound in Multimedia Signal Processing*. PhD thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, 1999.

[34] A. Klapuri, "Automatic transcription of music," Master's thesis, Tampere University of Technology, Tampere, Finland, Mar. 1998. Available at http://www.cs.tut.fi/˜klap/iiro/index.html.

[35] K. Lemström and P. Laine, "Musical information retrieval using musical parameters," in *Proceedings of the International Computer Music Conference*, (Ann Arbor, USA), pp. 341–348, Oct. 1998.

[36] K. Lemström, P. Laine, and S. Perttu, "Using relative interval slope in music information retrieval," in *Proceedings of the International Computer Music Conference*, (Beijing, China), pp. 317–320, Oct. 1999.

[37] E. D. Scheirer, “Tempo and beat analysis of acoustic musical signals,” *Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[38] E. D. Scheirer and Y. E. Kim, “Generalized audio coding with MPEG-4 structured audio,” in *Proceedings of the AES 17th International Conference on High-Quality Audio Coding*, (Florence, Italy), Sept. 1999.

[39] J. C. Brown, “Computer identification of musical instruments using pattern recognition with cepstral coefficients as features,” *Journal of the Acoustical Society of America*, vol. 105, pp. 1933–1941, Mar. 1999.

[40] IMA, “MIDI musical instrument digital interface specification 1.0,” 1983. Los Angeles: International MIDI Association.

[41] M. Laurson, J. Hiipakka, C. Erkut, M. Karjalainen, V. Välimäki, and M. Kuuskankare, “From expressive notation to model-based sound synthesis: a case study of the acoustic guitar,” in *Proceedings of the International Computer Music Conference*, (Beijing, China), pp. 1–4, Oct. 1999.

[42] C. Roads, *The Computer Music Tutorial*. Cambridge, Massachusetts, USA: The MIT Press, 1995.

[43] T. Tolonen, “Methods for separation of harmonic sound sources using sinusoidal modeling,” in *AES 106th Convention*, (Munich, Germany), May 1999.

[44] M. Karjalainen and T. Tolonen, “Separation of speech signals using iterative multi-pitch analysis and prediction,” in *Proceedings of the EuroSpeech’99*, vol. 5, Sept. 1999.

[45] U. K. Laine, M. Karjalainen, and T. Altosaar, “Warped linear prediction (WLP) in speech and audio processing,” *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. III.349 –III.352, 1994.

[46] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U. K. Laine, and J. Huopaniemi, “Frequency-warped signal processing for audio applications,” in *Proceedings of the 108th AES Convention*, (Paris, France), Feb. 2000. Preprint 5171.

[47] R. J. McAulay and T. F. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 744–754, Aug. 1986.

[48] J. O. Smith and X. Serra, “Parshl: an analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation,” in *Proceedings of the International Computer Music Conference*, (Urbana-Champaign, Illinois, USA), pp. 290–297, 1987.

[49] D. C. Rife and R. R. Boorstyn, “Single-tone parameter estimation from discrete-time observations,” *IEEE Transactions on Information Theory*, vol. 20, pp. 591–598, Sept. 1974.

[50] D. C. Rife and R. R. Boorstyn, “Multiple tone parameter estimation from discrete-time observations,” *The Bell System Technical Journal*, vol. 55, pp. 1389–1410, Nov. 1976.

[51] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Upper Saddle River, New Jersey: Prentice Hall, 1997.

[52] S. M. Kay, *Modern Spectral Estimation: Theory and Application*. Englewood Cliffs, New Jersey: Prentice Hall, 1988.

[53] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*. New York, USA: Macmillan, 2 ed., 1992.

[54] G. Goertzel, “An algorithm for the evaluation of finite trigonometric series,” *American Mathematical Monthly*, vol. 65, pp. 34–35, 1968.

[55] M. Mellody and G. H. Wakefield, “The time-frequency characteristic of violin vibrato: modal distribution analysis and synthesis,” *Journal of the Acoustical Society of America*, vol. 107, pp. 598–611, Jan. 2000.

[56] F. Opolko and J. Wapnick, “McGill University Master Samples.” CD Collection, 1988.

[57] C. Erkut, V. Välimäki, M. Karjalainen, and M. Laurson, “Extraction of physical and expressive parameters for model-based sound synthesis of the classical guitar,” in *Proceedings of the 108th AES Convention*, (Paris, France), 1999. Preprint 5114.

[58] T. Tolonen, V. Välimäki, and M. Karjalainen, “Modeling of tension modulation nonlinearity in plucked strings,” *IEEE Transactions on Speech and Audio Processing*, vol. 8, May 2000.

[59] C. Erkut, T. Tolonen, M. Karjalainen, and V. Välimäki, “Acoustical analysis of the tanbur, a turkish long-necked lute,” in *Proceedings of the 6th International Congress on Sound and Vibration*, vol. 1, (Lyngby, Denmark), pp. 345–352, July 1999.

[60] V. Välimäki, M. Karjalainen, T. Tolonen, and C. Erkut, “Nonlinear modeling and synthesis of the Kantele—a traditional Finnish string instrument,” in *Proceedings of the International Computer Music Conference*, (Beijing, China), pp. 220–223, Oct. 1999.

[61] H. Fletcher, E. D. Blackham, and R. Stratton, “Quality of piano tones,” *Journal of the Acoustical Society of America*, vol. 77, no. 6, pp. 749–761, 1962.

[62] D. Rocchesso and F. Scalcon, “Bandwidth of perceived inharmonicity for musical modeling of dispersive strings,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 597–601, Sept. 1999.

[63] F. Scalcon, D. Rocchesso, and G. Borin, “Subjective evaluation of the inharmonicity of synthetic piano tones,” in *Proceedings of the International Computer Music Conference*, pp. 53–56, 1998.

[64] H. Järveläinen, V. Välimäki, and M. Karjalainen, “Audibility of inharmonicity in string instrument sounds, and implications to digital sound synthesis,” in *Proceedings of the International Computer Music Conference*, (Beijing, China), pp. 359–362, Oct. 1999.

[65] T. Tolonen and H. Järveläinen, “Perceptual study of decay parameters in plucked string synthesis,” in *Proceedings of the 109th AES Convention*, (Los Angeles), Sept. 2000. Accepted for publication.

[66] H. Järveläinen, T. Verma, and V. Välimäki, “The effect of inharmonicity on pitch in string instrument sounds,” in *Proceedings of the International Computer Music Conference*, (Berlin, Germany), Sept. 2000. Submitted for publication.

[67] J. O. Smith, “Music applications of digital waveguides,” Tech. Rep. STAN-M-39, CCRMA, Dept. of Music, Stanford University, California, USA, May 1987.

[68] J. O. Smith, “Physical modeling using digital waveguides,” *Computer Music Journal*, vol. 16, no. 4, pp. 74–91, 1992.

[69] J. O. Smith, “Acoustic modeling using digital waveguides,” in *Musical Signal Processing* (C. Roads, S. T. Pope, A. Piccialli, and G. De Poli, eds.), ch. 7, pp. 221–264, Lisse, the Netherlands: Swets & Zeitlinger, 1997.

[70] J. O. Smith, “Efficient synthesis of stringed musical instruments,” in *Proceedings of the International Computer Music Conference*, (Tokyo, Japan), pp. 64–71, Sept. 1993.

[71] M. Karjalainen, V. Välimäki, and Z. Jánosy, “Towards high-quality sound synthesis of the guitar and string instruments,” in *Proceedings of the International Computer Music Conference*, (Tokyo, Japan), pp. 56–63, Sept. 1993.

[72] M. Karjalainen, V. Välimäki, and T. Tolonen, “Plucked-string models: from Karplus-Strong algorithm to digital waveguides and beyond,” *Computer Music Journal*, vol. 22, no. 3, pp. 17–32, 1998.

[73] V. Välimäki, J. Huopaniemi, M. Karjalainen, and Z. Jánosy, “Physical modeling of plucked string instruments with application to real-time sound synthesis,” *Journal of the Audio Engineering Society*, vol. 44, pp. 331–353, May 1996.

[74] V. Välimäki, *Discrete-Time Modeling of Acoustic Tubes Using Fractional Delay Filters*. PhD thesis, Helsinki University of Technology, Espoo, Finland, 1995.

[75] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, “Splitting the unit delay—tools for fractional delay filter design,” *IEEE Signal Processing Magazine*, vol. 13, pp. 30–60, Jan. 1996.

[76] T. Tolonen, “Model-based analysis and resynthesis of acoustic guitar tones,” Master’s thesis, Helsinki University of Technology, Espoo, Finland, Jan. 1998. Report 46, Laboratory of Acoustics and Audio Signal Processing.

[77] M. Karjalainen and J. O. Smith, “Body modeling techniques for string instrument synthesis,” in *Proceedings of the International Computer Music Conference*, (Hong Kong), pp. 232–239, Aug. 1996.

[78] T. Tolonen, V. Välimäki, and M. Karjalainen, “Modeling of tension modulation nonlinearity in plucked strings,” *IEEE Transactions on Speech and Audio Processing*, 1999. Accepted for Publication.

[79] C. Drioli and D. Rocchesso, “A generalized musical-tone generator with application to sound compression and synthesis,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 431–434, 1997.

[80] M. Karjalainen and V. Välimäki, “Model-based analysis/synthesis of the acoustic guitar,” in *Proceedings of the Stockholm Music Acoustic Conference*, (Stockholm, Sweden), pp. 443–447, 1993.