EURASIP Journal on Applied Signal Processing 2004:7, 1036–1044

Warped Linear Prediction of Physical Model Excitations with Applications in Audio Compression and Instrument Synthesis

Alexis Glass

Department of Acoustic Design, Graduate School of Design, Kyushu University, 4-9-1 Shiobaru, Minami-ku, Fukuoka 815-8540, Japan
Email: alexis@andes.ad.design.kyushu-u.ac.jp

Kimitoshi Fukudome

Department of Acoustic Design, Faculty of Design, Kyushu University, 4-9-1 Shiobaru, Minami-ku, Fukuoka 815-8540, Japan
Email: fukudome@design.kyushu-u.ac.jp

Received 8 July 2003; Revised 13 December 2003

A sound recording of a plucked string instrument is encoded and resynthesized using two stages of prediction. In the first stage of prediction, a simple physical model of a plucked string is estimated and the instrument excitation is obtained. The second stage of prediction compensates for the simplicity of the model in the first stage by encoding either the instrument excitation or the model error using warped linear prediction. These two methods of compensation are compared with each other, and to the case of single-stage warped linear prediction, adjustments are introduced, and their applications to instrument synthesis and MPEG4's audio compression within the structured audio format are discussed.

Keywords and phrases: warped linear prediction, audio compression, structured audio, physical modelling, sound synthesis.

1. INTRODUCTION

Since the discovery of the Karplus-Strong algorithm [1] and its subsequent reformulation as a physical model of a string, a subset of the digital waveguide [2], physical modelling has seen the rapid development of increasingly accurate and disparate instrument models. Not limited to string model implementations of the digital waveguide, such as the kantele [3] and the clavichord [4], models for brass, woodwind, and percussive instruments have made physical modelling ubiquitous.

With the increasingly complex models, however, the task of parameter selection has become correspondingly difficult. Techniques for calculating the loop filter coefficients and excitation for basic plucked string models have been refined [5, 6] and can be quickly calculated. However, as the one-dimensional model gave way to models with weakly interacting transverse and vertical polarizations, research has looked to new ways of optimizing parameter selection. These new methods of optimizing parameter selection use neural networks or genetic algorithms [7, 8] to automate tasks which would otherwise take human operators an inordinate amount of time to adjust. This research has yielded more accurate instrument models, but for some applications it also leaves a few problems unaddressed.

The MPEG-4 structured audio codec allows for the implementation of any coding algorithm, from linear predictive coding to adaptive transform coding to, at its most efficient, the transmission of instrument models and performance data [9]. This coding flexibility means that MPEG-4 has the potential to implement any coding algorithm and to be within an order of magnitude of the most efficient codec for any given input data set [10]. Moreover, for sources that are synthetic in nature, or can be closely approximated by physical or other instrument models, structured audio promises levels of compression orders of magnitude better than what is currently possible using conventional pure signal-based codecs.

Current methods used to parameterize physical models from recordings require, however, a great deal of time for complex models [8]. They also often require very precise and comprehensive original recordings, such as recordings of the impulse response of the acoustic body [5, 11], in order to achieve reproductions that are indistinguishable from the original. Given current processor speeds, these limitations preclude the use of genetic algorithm parameter selection techniques for real-time coding. Real-time coding is also made exceedingly difficult in such cases where body impulse responses are not available or playing styles vary from model expectations.

This paper proposes a solution to this real-time parameterization and coding problem for string modelling in the marriage of two common techniques, the basic plucked string physical model and warped linear prediction (WLP) [12].

The justifications for this approach are as follows. Most string recordings can be analyzed using the techniques developed by Smith, Karjalainen et al. [2, 6] in order to parameterize a basic plucked string model, and a considerable prediction gain can be achieved using these techniques. The excitation signal for the plucked string model is constituted by an attack transient that represents the plucking of the string according to the player's style and plucking position [11], and is followed by a decay component. This decay component includes the body resonances of the instrument [11, 13], beating introduced by the string's three-dimensional movement and further excitation caused by the player's performance. Additional excitations from the player's performance include deliberate expression through vibrato or even unintentional influences, such as scratching of the string or the rattling caused by the string vibrating against the fret with weak fingering pressure. The body resonances and contributions from the three-dimensional movement of the string mean that the excitation signal is strongly correlated and therefore a good candidate for WLP coding. Furthermore, while residual quantization noise in a warped predictive codec is shaped so as to be masked by the signal's spectral peaks [12], in one of the proposed topologies, the noise in the physical model's excitation signal is likewise shaped into the modelled harmonics. This shaping of the noise by the physical model results in distortion that, if audible, is neither unnatural nor distracting, thereby allowing codec sound quality to degrade gracefully with decreasing bit rate. In the ideal case, we imagine that at the lowest bit rate, the guitar would be transmitted using only the physical model parameters and that with increasing excitation bit rate, the reproduced guitar timbre would become closer to the target original one.

This paper is composed of six sections. Following the introduction, the second section describes the plucked string model used in this experiment and the analysis methods used to parameterize it. The third section describes the recording of a classic guitar and an electric guitar for testing. The coding of the guitar tones using a combination of physical modelling and warped linear predictive coding is outlined in Section 4. Section 5 analyzes the results from simulated coding scenarios using the recorded samples from Section 3 and the topologies of Section 4, while investigating methods of further improving the quality of the codec. Section 6 concludes the paper.

2. MODEL STRUCTURE

A simple linear string model extended from the Karplus-Strong algorithm, by Jaffe and Smith [14], was used in this study, comprised of one delay line z^{-L} with a first-order allpass fractional delay filter $F(z)$ and a single pole low-pass loop filter $G(z)$ as shown in Figure 1, where,

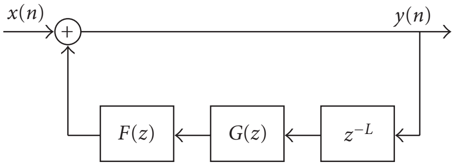


FIGURE 1: Topology of a basic plucked string physical model.

$$F(z) = \frac{a + z^{-1}}{1 + az^{-1}}, \tag{1}$$

$$G(z) = \frac{g(1 + a_1)}{1 + a_1 z^{-1}}, \tag{2}$$

and the overall transfer function of the system can be expressed as

$$H(z) = \frac{1}{1 - F(z)G(z)z^{-L}}. \tag{3}$$

This string model is very simple and much more accurate and versatile models have been developed since [6, 11, 15]. For the purposes of this study, however, it was required that the model could be quickly and accurately parameterized without the use of complex or time consuming algorithms and sufficient that it offers a reasonable first-stage coding gain. The algorithms used to parameterize the first-order model are described in detail in [15] and will only be outlined here as they were implemented for this study.

In the first stage of the model parameterization, the pitch of the target sound was detected from the target's autocorrelation function. The length of the delay line z^{-L} and the fractional delay filter $F(z)$ were determined by dividing the sampling frequency (44.1 kHz) by the pitch of the target. Next, the magnitude of up to the first 20 harmonics were tracked using short-term Fourier transforms (STFTs). The magnitude of each harmonic versus time was recorded on a logarithmic scale after the attack transient of the pluck was determined to have dissipated and until the harmonic had decayed 40 dB or disappeared into the noise floor.

A linear regression was performed on each harmonic's decay to determine its slope, β_k, as shown in Figure 2, and the measured loop gain for each harmonic, G_k, was calculated according to the following equation,

$$G_k = 10^{\beta_k L/20H}, \quad k = 1, 2, \ldots, N_h, \tag{4}$$

where L is the length of the delay line (including the fractional component), and H is the hop size (adjusted to account for hop overlap). The loop gain at DC, g, was estimated to equal the loop gain of the first harmonic, G_1, as in [15]. Because the target guitar sounds were arbitrary and nonideal, the harmonic envelop trajectories were quite noisy in some cases, so, additional measures had to be introduced to stop tracking harmonics when their decays became too

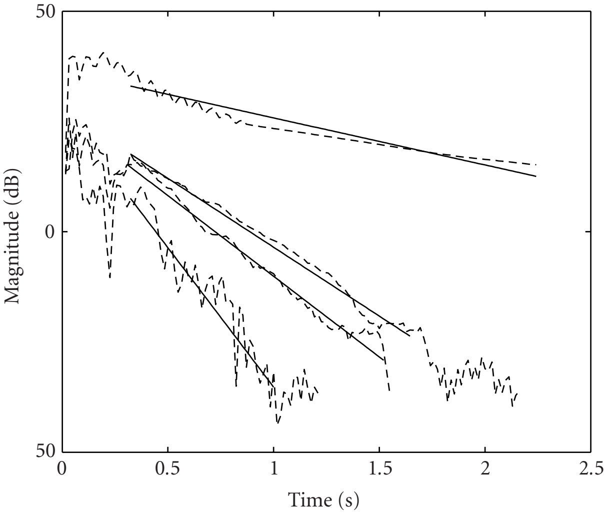


Figure 2: The temporal envelopes of the lowest four harmonics of a guitar pluck (dashed) and their estimated decays (solid).

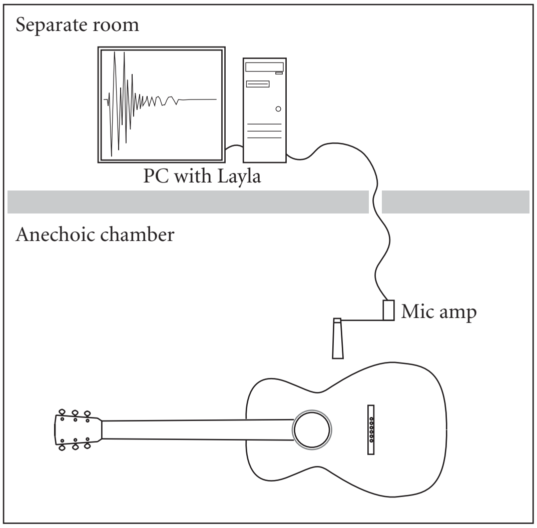


Figure 3: Schematic for classic guitar pluck recording.

erratic or, as in some cases, negative. In such cases as when the guitar fret was held with insufficient pressure, additional transients occurred after the first attack transient and this tended to raise the gain factor in the loop filter, resulting in a model that did not accurately reflect string losses. For the purposes of this study, such effects were generally ignored so long as a positive decay could be measured from the harmonics tracked.

The first-order loop filter coefficient a_1 was estimated by minimizing the weighted error between the target loop filter G_k, as calculated in (4), and candidate filters $G(z)$ from (2). A weighting function W_k, suggested by [15] and defined as

$$W_k = \frac{1}{(1 - G_k)}, \qquad (5)$$

was used such that the error could be calculated as follows:

$$E(a_1) = \sum_{k=1}^{N_h} W_k \left(G_k - \left|G\left(e^{j\omega_k}, a_1\right)\right|\right), \qquad (6)$$

where ω_k is the frequency at the harmonic being evaluated and $0 < a_1 < 1$. This error function is roughly quadratic in the vicinity of the minimum, and parabolic interpolation was found to yield accurate values for the minimum in less time than iterative methods.

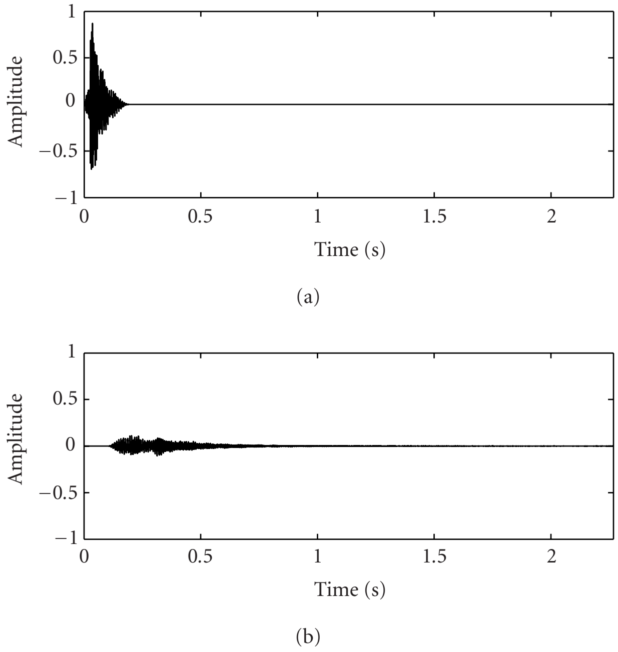
For controlled calibration of the loop filter extraction algorithm, synthesized plucked string samples were created using the extended Karplus-Strong algorithm and the model as described by Välimäki [11], with two string polarizations and a weak sympathetic coupling between the strings.

3. DATA ACQUISITION

The purpose of the algorithms explored in this research was to resynthesize real, nontrivial plucked string sounds using the combination of the basic plucked string model and WLP coding. No special care was taken, therefore, in the selection of the instruments to be used or the nature of the guitar tones to be analyzed and resynthesized beyond that they were monophonic, recorded in an anechoic chamber and each pluck was preceded by silence to facilitate the analysis process. A schematic of the recording environment and signal flow for the classic guitar is pictured in Figure 3.

Two guitars were recorded. The first, a classic guitar, was recorded in an anechoic chamber with the guitar held approximately 50 cm from a Bruel & Kjaer type 4191 free field 1/2′′ microphone, the output of which was amplified by a Falcon Range 1/2′′ type 2669 microphone preamp with a Bruel & Kjaer type 5935 power supply and fed into a PC through a Layla 24/96 multitrack recording system. The electric guitar was recorded through its line out and a Yamaha O3D mixer into the Layla. A variety of plucking styles were recorded in both cases, along with the application of vibrato, string scratching, and several cases where insufficient finger pressure on the frets lead to further string excitation (i.e., a rattling of the string) after the initial pluck.

After capturing approximately 8 minutes of playing with each guitar, suitable candidates for the study were selected on the basis of their unique timbres, durations, and potential difficulty for accurate resynthesis using existing plucked string models. More explicitly, in the case of the classic guitar, bright plucks of E1 (82 Hz) were recorded along with several recordings of B1 (124 Hz), where weak finger pressure lead to a rattling of the string. Another sample selected involved this weak finger pressure leading to an early damping of the string by the fret hand, though without the nearly instantaneous subsequent decay that a fully damped string would yield. A third, higher pitch was recorded with an open string at E3 (335 Hz). In the case of the electric guitar, two samples were used—one of slapped E1 (82 Hz) with almost no decay and another of E2 (165 Hz) with some vibrato applied.



(a)

(b)

Figure 4: The decomposition of an excitation into (a) attack and (b) decay. The attack window is 200 milliseconds long. In this case, decay refers to the portion of the pluck where the greatest attenuation is a result of string losses. Because the string is not otherwise damped, it may also be considered to be the sustain segment of the envelope.

4. ANALYSIS/RESYNTHESIS ALGORITHMS

4.1. *Warped linear prediction*

Frequency warping methods [16] can be used with linear prediction coding so that the prediction resolution closely matches the human auditory system's nonuniform frequency resolution. Härmä found that WLP realizes a basic psychoacoustic model [12]. As a control for the study, the target signal was therefore first processed using a twentieth-order WLP coder of lattice structure.

The lattice filter's reflection coefficients were not quantized, and after inverse filtering, the residual was split into two sections, attack and decay, which were quantized using a mid-riser algorithm. The step size in the mid-riser quantizer was set such that the square error of the residual was minimized. The number of bits per sample in the attack residual (BITSA) was set to each of BITSA $= \{16, 8, 4\}$ for each of the bits per sample in the decay residual BITSD $= \{2, 1\}$. The frame size for the coding was set to equal two periods of the guitar pluck being coded, and the reflection coefficients were linearly interpolated between frames. The bit allocation method was used in order to match the case of the topologies that use a first-stage physical model predictor, where more bits were allocated to the attack excitation than the decay excitation. Härmä found in [12] that near transparent quality could be achieved with 3 bits per sample using a WLP codec. It is therefore reasonable to suggest that the WLP used here could have been optimized by distributing the high number of bits used in the attack throughout the length of the sound to be coded. However, since similar optimizations could also be made in the two-stage algorithms, only the simplest method was investigated in this study.

4.2. *Windowed excitation*

As the most basic implementation of the physical model, the residual from the string model's inverse filter can be windowed and used as the excitation for the model. In this study, the excitation was first coded using a warped linear predictive coder of order 20 and with BITSA bits of quantization for each sample of the residual. In many cases, the first 100 milliseconds of the excitation contains enough information about the pluck and the guitar's body resonances for accurate resynthesis [13, 15]. The beating caused by the slight three-dimension movement of the string and the rattling caused by the energetic plucks used in the study, however, were significant enough that a longer excitation was used.

Specifically, the window used was thus unity for the first 100 milliseconds of the excitation and then decayed as the second half of a Hanning window for the following 100 milliseconds. An example of this windowed excitation can be seen in the top of Figure 4. This windowed excitation, considered as the attack component, was input to the string model for comparison to the WLP case and used in the modified extended Karplus-Strong algorithm which will now be described.

4.3. *Two-stage coding topologies*

As described in [9], structured audio allows for the parameterization and transmission of audio using arbitrary codecs. These codecs may be comprised of instrument models, effect models, psychoacoustic models, or combinations thereof. The most common methods used for the psychoacoustic compression of audio are transform codecs, such as MP3 [17] and ATRAC [18] and time-domain approaches such as WLP [12]. Because the specific application being considered here is that of the guitar, the first stage of our codec is the simple string model described in Section 2. The second stage of coding was then approached using one of two methods:

(1) the model's output signal error (referred to as model error) could be immediately coded using WLP, or

(2) the model's excitation could be coded using WLP, with the attack segment of the excitation receiving more bits as in the WLP case of Section 4.2.

The topologies of these two strategies are illustrated in Figure 5.

Both topologies require the inverse filtering of the target pluck sound in order to extract the excitation. The decomposition of the excitation into attack and decay components for the first topology, as formerly proposed by Smith [19] and implemented by Välimäki and Tolonen in [13], reflects the wideband and high amplitude portion which marks the beginning of the excitation signal and the decay which typically contains lower frequency components from body resonances

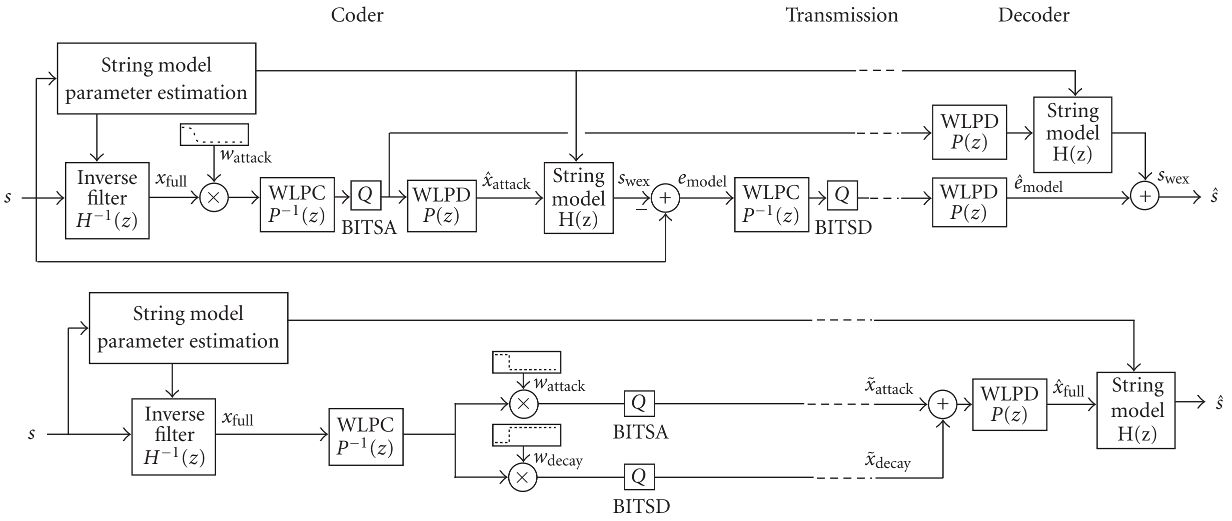


FIGURE 5: The WLP coding of model error (WLPCME) topology (top) and WLP coding of model excitation (WLPCMX) topology (bottom). Here, s represents the plucked string recording to be coded and $\hat{s}$ the reconstructed signal. In this diagram, WLPC indicates the WLP coder, or inverse filter, and WLPD indicates the WLP decoder. Q is the quantizer, with BITSA and BITSD being the number of bits with which the respective signals are quantized.

or from the three-dimensional movement of the string. However, whereas the authors of [13] synthesized the decay excitation at a lower sampling rate, justified by its predominantly lower frequency components, the excitations in our study often contained wideband excitations following the initial attack and no such multirate synthesis was therefore used. Typical attack and decay decomposition of an excitation is shown in Figure 4. The high frequency decay components are a result of the mismatch between the string model and the source recording.

4.4. Warped linear prediction coding of model error

The WLPCME topology from Figure 5 was implemented such that WLP was applied to the model error as follows

$$\begin{aligned} s_{\text{wex}} &= h * \hat{x}_{\text{attack}}, \\ e_{\text{model}} &= s - s_{\text{wex}}, \\ \hat{s} &= s_{\text{wex}} + \hat{e}_{\text{model}}, \end{aligned} \tag{7}$$

where s is the recorded plucked string input, h is the impulse response of the derived pluck string model from (3), $\hat{x}_{\text{attack}}$ is the WLP-coded windowed excitation introduced in Section 4.2, s_{wex} is the pluck resynthesized using only the windowed excitation, and e_{model} is the model error. $\hat{e}_{\text{model}}$ is thus the model error coded using WLP and BITSD bits per sample and $\hat{s}$ is the reconstructed pluck.

4.5. Warped linear prediction coding of model excitation

In this case, the model excitation was coded instead of the model error. Following the string model inverse filtering, the excitation is whitened using a twentieth-order WLP inverse filter. Next, the signal is quantized with BITSA bits per sample allotted to the residual in the attack, and BITSD bits per sample for the decay residual. This process can be expressed in the following terms:

$$\begin{aligned} x_{\text{full}} &= h^{-1} * s, \\ \tilde{x}_{\text{attack}} &= q_{\text{BITSA}}\left(p^{-1} * x_{\text{full}} \cdot w_{\text{attack}}\right), \\ \tilde{x}_{\text{decay}} &= q_{\text{BITSD}}\left(p^{-1} * x_{\text{full}} \cdot w_{\text{decay}}\right), \\ \hat{x}_{\text{full}} &= p * \left(\tilde{x}_{\text{attack}} + \tilde{x}_{\text{decay}}\right), \\ \hat{s} &= h * \hat{x}_{\text{full}}, \end{aligned} \tag{8}$$

where s is the original instrument recording being modelled, h is the string model's inverse filter, and x_{full} is thus the model excitation. $\tilde{x}_{\text{attack}}$ is therefore the string model excitation whitened by the WLP, p^{-1}, and quantized to BITSA, while $\tilde{x}_{\text{decay}}$ is likewise whitened and quantized to BITSD. The sum of the attack and decay is then resynthesized by the WLP decoder, p. The resulting $\hat{x}_{\text{full}}$ is subsequently considered as excitation to the string model, h, to form the resynthesized plucked string sound $\hat{s}$.

5. SIMULATION RESULTS AND DISCUSSION

In order to evaluate the effectiveness of the two proposed topologies, a measure of the sound quality was required. Informal listening tests suggested that the WLPCMX topology offered slightly improved sound quality and a more musical coding at lower bit rates, although it came at the cost of a much brighter timbre. At very low bit rates, WLPCMX introduced considerable distortion especially for sound sources that were poorly matched by the string model. WLPCME, on the other hand, was equivalent in sound quality to WLPC and sometimes worse. Resynthesis using windowed excitation yielded passable guitar-like timbres, but in none of the test cases came close to reproducing the nuance or fullness of the original target sounds.

For a more formal evaluation of the simulated codecs' sound quality, an objective measure of sound quality was calculated by measuring the spectral distance between the frequency warped STFTs, S_k, of the original pluck recording and the resynthesized output, $\hat{S}_k$, created using the codecs. The frequency-warped STFT sequences were created by first warping each successive frame of each signal using cascaded all-pass filters [16], followed by a Hanning window and a fast Fourier transform (FFT). The method by which the bark spectral distance (BSD) was measured is as follows:

$$\mathrm{BSD}_k = \sqrt{\left(\frac{1}{N}\sum_{n=0}^{N-1}\left(20\log_{10}\left|S_k(n)\right| - 20\log_{10}\left|\hat{S}_k(n)\right|\right)^2\right)}, \tag{9}$$

with the mean BSD for the whole sample being the unweighted mean of all frames k. A typical profile of BSD versus time is shown in Figure 6 for the three cases WLPC, WLPCMX, and WLPCME.

In the first round of simulations, all six input samples as described in Section 3 were processed using each of the algorithms described in Section 4. The resulting mean BSDs were then calculated to be as shown in Figure 7.

Subjective evaluation of the simulated coding revealed that as bit rate decreased, the WLPCMX topology maintained a timbre that, while brighter than the target, was recognizably as a guitar. In contrast, the other methods became noisy and synthetic. Objective evaluation of these same results reveals that both topologies using a first-stage physical model predictor have greater spectral distortion than the case of WLPC, particularly in the case of the recordings with very slow decays (i.e., with a high DC loop gain g). In identifying the cause of this distortion, we must first consider the model prediction. The degradation occurs for the following reason in each of the two topologies.

(A) In the case of the WLPCME, the beating that is caused by the three-dimensional vibration of the string causes considerable phase deviation from the phase of the modelled pluck, and the model error often becomes greater in magnitude than the original signal itself. This leads to a noisier reconstruction by the resynthesizer. Additionally, small model parameterization errors in pitch and the lack of vibrato in the model result in phase deviations.

(B) In the case of the WLPCMX, with a low bit rate in the residual quantization stage of the linear predictor, a small error in coding of the excitation is magnified by the resynthesis filter (string model). In addition to this, as noted in [15], the inverse filter may not have been of sufficiently high order to cancel all harmonics, and high frequency noise, magnified by the WLP coding, may have been further shaped by the plucked string synthesizer into bright higher harmonics.

The distortion caused by the topology in (A) seems impossible to improve significantly without using a more complex model that considers the three-dimensional vibration of the string, such as the model proposed by Välimäki et al. [11] and previously raised in Section 2. Performance control, such as vibrato, would also have to be extracted from the input for a locked phase to be achieved in the resynthesized pluck. The topology of (B), however, allows for some improvement in the reconstructed signal quality by compromising between the prediction gain of the first stage and the WLP coding of the second stage. More explicitly, if the loop filter gain was to be decreased, then the cumulative error being introduced by the quantization in the WLP stage would be correspondingly decreased.

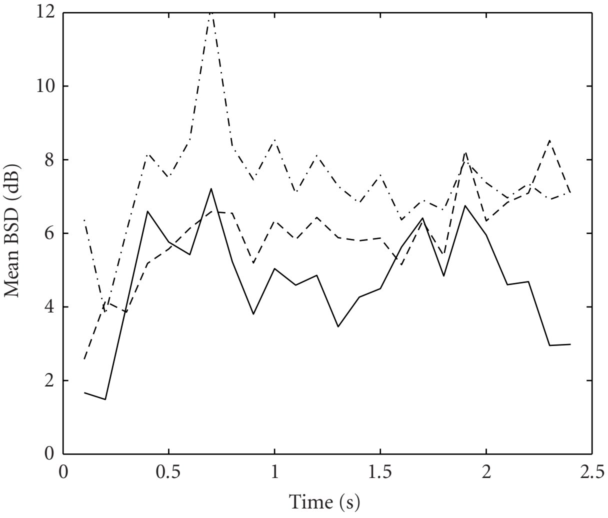


Figure 6: Bark scale spectral distortion (dB) versus time (seconds). WLPC is solid, WLPCMX is dashed-dotted, and WLPCME is the dashed line.

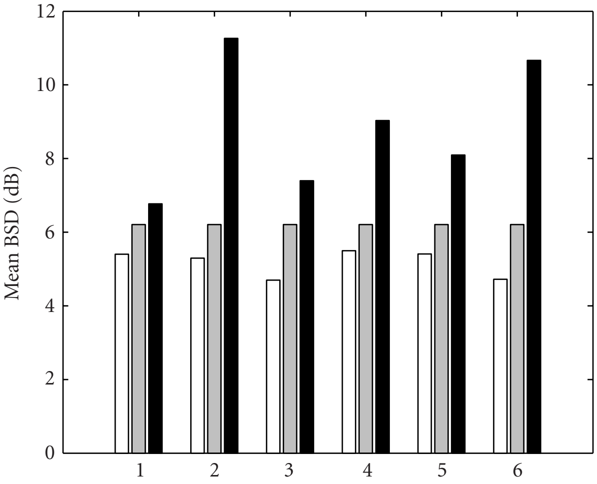


Figure 7: Mean Bark scale spectral distortion (dB) using each of WLPC, WLPCME, and WLPCMX (left to right) for (1) E3 classic, (2) E1 classic, (3) B1 classic (rattle 1), (4) B1 classic (rattle 2), (5) E1 electric, and (6) E2 electric. Simulation parameters were BITSA = 4 and BITSD = 1.

Such a downwards adjustment of the loop filter gain in order to minimize coding noise results in a physical model that represents a plucked string with an exaggerated decay. This almost makes the physical model prediction stage appear more like the long-term pitch predictor in a more conventional linear prediction (LP) codec targeted at speech. However, there is still the critical difference in that the physical model contains the low-pass component of the loop filter and can still be thought of as modelling the behaviour of a (highly damped) guitar string.

To obtain an appropriate value for the loop gain, multiplier tests were run on all six target samples. The electric guitar recordings and the recordings of the classical guitar at E3 represented "ideal" cases; there were no rattles subsequent to the initial pluck, in addition to negligible changes in pitch throughout their lengths. Amongst the remaining recordings, the two rattling guitar recordings represented two timbres very difficult to model without a lengthy excitation or a much more complex model of the guitar string. The mean BSD measure for the electric guitar at E1 is shown in Figure 8.

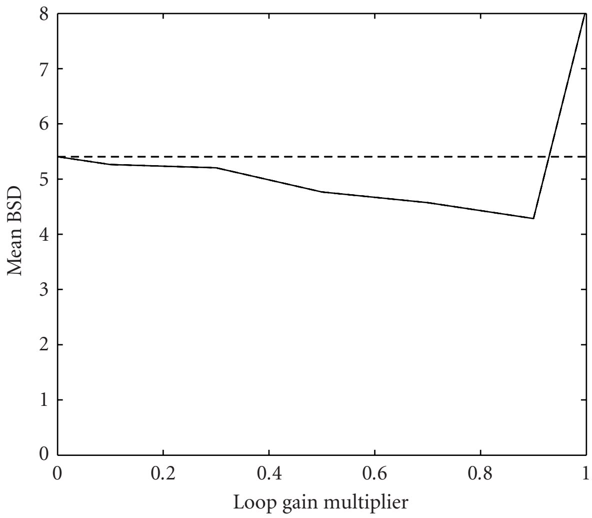


Figure 8: Mean Bark scale spectral distortion versus loop gain multiplier. WLPCMX is solid and WLPC is the dashed-dotted line.

As can be seen from Figure 8, reducing the loop gain of the physical model predictor increased the performance of the codec and yielded superior BSD scores for loop gain multipliers between 0.1 and 0.9. The greater the model mismatch, as in the case of the recordings with rattling strings, the less the string model predictor lowered the mean BSD. Models which did not closely match also featured minimal mean BSDs at lower loop gains (e.g., 0.5 to 0.7). The simulation used to produce Figure 7 was performed again using a single, approximately optimal, loop gain multiplier of 0.7. The results from this simulation are pictured in Figure 9.

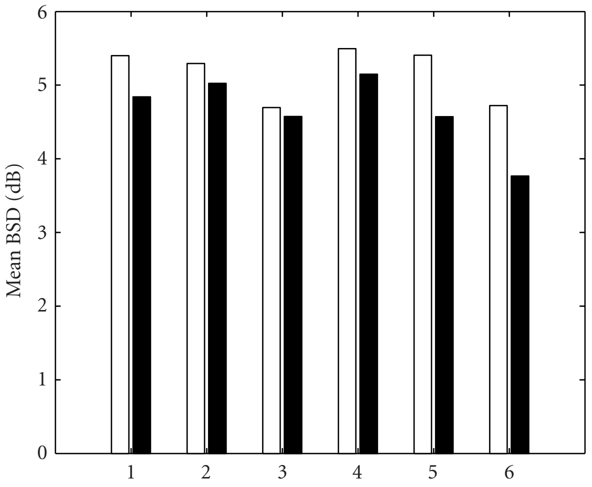


Figure 9: Mean Bark scale spectral distortion (dB) using each of WLPC, WLPCMX (left to right) for (1) E3 classic, (2) E1 classic, (3) B1 classic (rattle 1), (4) B1 classic (rattle 2), (5) E1 electric, and (6) E2 electric. Simulation parameters were BITSA = 4 and BITSD = 1.

The decreased BSD for all the samples in Figure 9 confirms the efficacy of the two-stage codec. Informal subjective listening tests described briefly at the beginning of this section also confirmed that decreasing the bit rate reduced the similarity of the reproduced timbre to the original timbre, without obscuring the fact that it was a guitar pluck and without the "thickening" of the mix that occurs due to the shaped noise in the WLPC codec. This improvement offered by the two-stage codec becomes even more noticeable at lower bit rates, such as with a constant 1 bit per sample quantization of WLP residual over both attack and decay.

To evaluate the utility of the proposed WLPCMX, it is important to compare it to the alternatives. Existing purely signal-based approaches such as MP3 and WLPC have proven their usefulness for encoding arbitrary wideband audio signals at low bit rates while preserving transparent quality. As an example, Härmä found that wideband audio could be coded using WLPC at 3 bits per sample (= 132.3 kbps @44.1 kHz) for good quality [12]. These models can be implemented in real-time with minimal computational overhead, but like sample-based synthesis, do not represent the transmitted signal parametrically in a form that is related to the original instrument. Pure signal-based approaches, using psychoacoustic models, are thus limited to the extent which they can remove psychoacoustically redundant data from an audio stream.

On the other hand, increasingly complex physical models can now reproduce many classes of instruments with excellent quality. Assuming a good calibration or, in the best case, a performance made using known physical modelling algorithms, transmission of model parameters and continuous controllers would result in a bit rate at least an order of magnitude lower than the case of pure signal-based methods. As an example, if we consider an average score file from a modern sequencing program using only virtual instruments and software effects, the file size (including simple instrument and effect model algorithms) is on the order of 500 kB. For an average song length of approximately 4 minutes, this leads to a bit rate of approximately 17 kbps. For optimized scores and simple instrument models, the bit rate could be lower than 1 kbps. Calibration of these complex instrument models to resynthesize acoustic instruments remains an obstacle for real-time use in coding, however. Likewise, parametric models are flexible within the class for which they are designed, but an arbitrary performance may contain elements not supported by the model. Such a performance cannot be reproduced by the pure physical model and may, indeed, result in poor model calibration for the performance as a whole.

This preliminary study of the WLPCMX topology offers a compromise between the pure physical-model-based approaches and the pure signal-based approaches. For the case of the monophonic plucked string considered in this study, a lower spectral distortion was realized using the model-based predictor. Because more bits were assigned to the attack portion of the string recording, the actual long-term bit rate of the codec is related to the frequency of plucks, but at its worst case it is limited by the rate of the WLP stage (assuming a loop gain multiplier of 0) and its best case, given a close match between model and recording, approaches the physical model case. For recordings that were well modelled by the string model, such as the electric guitar at E1 and E2 and the E3 classic guitar sample, subjective tests suggested that equivalent quality could be achieved with 1 bit per sample less than the WLPC case. Limitations of the string model prevent it from capturing all the nuances of the recording, such as the rattling of the classical guitar's string, but these unmodelled features are successfully encoded by the WLP stage. Because the predictor reflects the acoustics of a plucked string, degradation in quality with lower bit rates sounds more natural.

6. CONCLUSIONS

The implementation of a two-stage audio codec using a physical model predictor followed by WLP was simulated and the subjective and objective sound quality analyzed. Two codec topologies were investigated. In the first topology, the instrument response was estimated by windowing the first 200 milliseconds of the excitation, and this estimate was subtracted from the target sample, with the difference being coded using WLP coding. In the second topology, the excitation to the plucked string physical model was coded using WLP before being reconstructed by reapplying the coded excitation to the string model shown in Figure 1. Tests revealed that the limitations of the physical model resulted in model error in the first topology to be of greater amplitude than the target sound, and the codec therefore operated with inferior quality to the WLPC control case.

The second topology, however, showed promise in subjective tests whereby a decrease in the bits allocated to the coding of the decay segment of the excitation reduced the similarity of the timbre without changing its essential likeness to a plucked string. A further simulation was performed wherein the loop gain of the physical model was reduced in order to limit the propagation of the excitation's quantization error due to the physical model's long-time constant. This improved objective measures of the sound quality beyond those achieved by the similar WLPC design while maintaining the codec's advantages exposed by the subjective tests. Whereas the target plucks became noisy when coded at 1 bit per sample using WLPC, the allocation of quantization noise to higher harmonics in the second topology meant that the same plucks took on a drier, brighter timbre when coded at the same bit rate.

WLP can easily be performed in real-time, and it could thus be applied to coding model excitations in both audio coders and in real-time instrument synthesizers. Analysis of polyphonic scenes is still beyond the scope of the model, however, and the realization of highly polyphonic instruments would entail a corresponding increase in computational demands from the WLP in the decoding of the excitation.

Future exploration of the two-stage physical model/WLP coding schemes should be investigated using more accurate physical models, such as the vertical/transverse string model mentioned in Section 1, which might allow the first topology investigated in this paper to realize coding gains. Implementation of more complicated models reintroduces, however, the difficulties of accurately parameterizing them—though this increased complexity is partially offset by the increased tolerance for error that the excitation coding allows.

ACKNOWLEDGMENTS

The authors would like to thank the Japanese Ministry of Education, Culture, Sports, Science and Technology for funding this research. They are also grateful to Professor Yoshikawa for his guidance throughout, and the students of the Signal Processing Lab for their assistance, particularly in making the guitar recordings.

REFERENCES

[1] K. Karplus and A. Strong, "Digital synthesis of plucked-string and drum timbres," *Computer Music Journal*, vol. 7, no. 2, pp. 43–55, 1983.

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

[3] C. Erkut, M. Karjalainen, P. Huang, and V. Välimäki, "Acoustical analysis and model-based sound synthesis of the kantele," *Journal of the Acoustical Society of America*, vol. 112, no. 4, pp. 1681–1691, 2002.

[4] V. Välimäki, M. Laurson, C. Erkut, and T. Tolonen, "Model-based synthesis of the clavichord," in *Proc. International Computer Music Convention*, pp. 50–53, Berlin, Germany, August–September 2000.

[5] V. Välimäki and T. Tolonen, "Development and calibration of a guitar synthesizer," *Journal of the Audio Engineering Society*, vol. 46, no. 9, pp. 766–778, 1998.

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

[7] A. Cemgil and C. Erkut, "Calibration of physical models using artificial neural networks with application to plucked string instruments," in *Proc. International Symposium on Musical Acoustics*, Edinburgh, UK, August 1997.

[8] J. Riionheimo and V. Välimäki, "Parameter estimation of a plucked string synthesis model using a genetic algorithm with perceptual fitness calculation," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 8, pp. 791–805, 2003.

[9] 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, pp. 922–940, 1998.

[10] E. D. Scheirer, "Structured audio, Kolmogorov complexity, and generalized audio coding," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 914–931, 2001.

[11] M. Karjalainen, V. Välimäki, and Z. Janosy, "Towards high-quality sound synthesis of the guitar and string instruments," in *Proc. International Computer Music Conference*, pp. 56–63, Tokyo, Japan, 1993.

[12] A. Härmä, *Audio coding with warped predictive methods*, Licentiate thesis, Helsinki University of Technology, Espoo, Finland, 1998.

[13] V. Välimäki and T. Tolonen, "Multirate extensions for model-based synthesis of plucked string instruments," in *Proc. International Computer Music Conference*, pp. 244–247, Thessaloniki, Greece, September 1997.

[14] D. Jaffe and J. O. Smith, "Extensions of the Karplus-Strong plucked-string algorithm," *Computer Music Journal*, vol. 7, no. 2, pp. 56–69, 1983.

[15] 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, no. 5, pp. 331–353, 1996.

[16] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U. K. Laine, and J. Huopaniemi, "Frequency-warped signal processing for audio applications," *Journal of the Audio Engineering Society*, vol. 48, no. 11, pp. 1011–1031, 2000.

[17] K. Brandenburg and G. Stoll, "ISO/MPEG-audio codec: a generic standard for coding of high quality digital audio," *Journal of the Audio Engineering Society*, vol. 42, no. 10, pp. 780–791, 1994.

[18] K. Tsutsui, H. Suzuki, O. Shimoyoshi, M. Sonohara, K. Akagiri, and R. M. Heddle, *ATRAC: Adaptive transform acoustic coding for MiniDisc*, reprinted from the 93rd Audio Engineering Society Convention, San Francisco, Calif, USA, 1992.

[19] J. O. Smith, "Efficient synthesis of stringed musical instruments," in *Proc. International Computer Music Conference*, pp. 64–71, Tokyo, Japan, September 1993.

Alexis Glass received his B.S.E.E. from Queen's University, Kingston, Ontario, Canada in 1998. During his bachelor's degree, he interned for nine months at Toshiba Semiconductor in Kawasaki, Japan. After graduating, he worked for a defense firm in Kanata, Ontario and a videogame developer in Montreal, Quebec before winning a Monbusho Scholarship from the Japanese government to pursue graduate studies at Kyushu Institute of Design (KID, now Kyushu University, Graduate School of Design). In 2002, he received his Master's of Design from KID and is currently a doctoral candidate there. His interests include sound, music signal processing, instrument modelling, and electronic music.

Kimitoshi Fukudome was born in Kagoshima, Japan in 1943. He received his B.E., M.E., and Dr.E. degrees from Kyushu University in 1966, 1968, and 1988, respectively. He joined Kyushu Institute of Design's Department of Acoustic Design as a Research Associate in 1971 and has been an Associate Professor there since 1990. With the October 1, 2003 integration of Kyushu Institute of Design into Kyushu University, his affiliation has changed to the Department of Acoustic Design, Faculty of Design, Kyushu University. His research interests include digital signal processing for 3D sound systems, binaural stereophony, engineering acoustics, and direction of arrival (DOA) estimation with sphere-baffled microphone arrays.