SOURCE-FILTER MODEL FOR QUASI-HARMONIC INSTRUMENTS

First author
Affiliation1
author1@ismir.edu

Second author
Retain these fake authors in
submission to preserve the formatting

Third author
Affiliation3
author3@ismir.edu

ABSTRACT

In this paper we propose a new method for a generalized model representing the time-varying spectral characteristics of quasi- harmonic instruments. This approach comprises a linear source-filter model, a parameter estimation method and a variance based model evaluation. The model is composed of an excitation source generating sinusoidal parameter trajectories and a modeling resonance filter. To estimate the model parameters we apply a gradient decent method to a training database. To finally evaluate our model we estimate the model variance on a test database. Such a model could later be used as a priori knowledge for polyphonic instrument recognition, polyphonic transcription and source separation algorithms.

1. INTRODUCTION

The purpose of our approach is to define a generalized, as well as an accurate and compact representation of the time-varying spectral characteristics of a single, quasi-harmonic instrument. While we assume the spectral envelope to be determined by the partial amplitude trajectories, our model is meant to predict the time-varying amplitude trajectories for unknown signals. Therefore, in order to prototype an instrument, we need to estimate the model parameters using a training database. To evaluate the performance of each prototype we finally measure the variance between the predicted partial amplitude trajectories of the prototype and those found in a test database.

Two approaches for a generalized representation of the spectral characteristics of quasi-harmonic instruments have been proposed recently. In [1] a representational model based on additive analysis and Principal Component Analysis (PCA) is presented in a first step, while in a subsequent stage, the spectral evolutions are modeled as Gaussian Processes (GP), i.e., as trajectories of varying mean and covariance in PCA space. Applied to musical instrument recognition, the model has been shown to significantly improve classification results compared to a Mel-Frequency-Cepstral-Coefficient (MFCC) based method. A source-filter-decay model is proposed in [4] and successfully applied to musical content analysis in [7] and [3]. In this approach the spectral envelope of an instrument sound is modelled by a source representing a vibrating object and a resonance filter related to the instrument's body which colors the generated sound. A decay filter is further used to model the time-varying characteristics.

In our approach we also adopt a linear source-filter model with similar interpretations of its components, but we extend this approach by taking the time variability into account for the complete amplitude envelope including attack and release regions. We will further introduce the use of B-Splines [5] for modeling the time varying spectral envelope as a smooth trajectory with respect to the different regions of the amplitude envelope. Finally, this yields a model parameterization determined only by the definition of the B-Splines.

In section 2 we will give a comprehensive description of our assumed signal model as well as our proposed generalized model, while section 3 describes how to estimate the model parameters from a given database. Section 4 will present our variance measure to evaluate the model and results for some selected prototypes including their evaluation are presented in section 5.

2. THE MODEL

Based on the general assumption that the spectral characteristics of an instrument sound are primarily determined by the partials amplitudes, we start with an overview of the signal model being used throughout this work. The subsequent paragraph will show how this signal model will be represented by our approach for a generalized model.

2.1 Signal Model

In additive analysis/synthesis it is assumed that a signal $x[n]$ can be approximated by a sum of stable sinusoids [6], so called partials. These partials may vary slowly in amplitude and frequency over time n.

$$x[n] \approx \tilde{x}[n] = \sum_{k=1}^{K} a[k,n] cos(\phi[k,n]) \qquad (1)$$

In eq. (1) k is the partial index, while K denotes the number of partials, a denotes the amplitude for partial k at time n, as ϕ its phase. Therefore the signal is modelled by its deterministic component only. We furthermore normalize the partial amplitude values by their maximum over time to analyse the signal characteristics independently of the actual energy. As a consequence, we denote $A[k,n]$ to be

the normalized energy level of a partial amplitude given in dB.

Since the spectral envelope varies over time, particularly attack and release, an assumption has to be made, with regard to this variability. As time itself is an unfavourable unit due to varying durations of attack and release and by a multiple of arbitrary signal lengths, we assume the variation of the spectral envelope to be directly related to the relative energy level of the signal. Accordingly, we assume the spectral envelope to be constant for a specific relative energy level and consider different envelopes for different levels. This also includes the assumption of the spectral envelope to be independent of the actual volume. Thus, the relative energy level $L[n]$ over time is given by eq. (3).

$$E[n] = \sum_{k=1}^{K} \big(a[k,n]\big)^2 \tag{2}$$

$$L[n] = 10 \cdot \log_{10}\left(\frac{E[n]}{E_{max}[n]}\right) \tag{3}$$

Moreover, we have to take into consideration that levels below 0 dB may correspond to either the attack or release, but the spectral envelopes may differ for these regions. We therefore have to determine the signal's attack n_A and release n_R time frames and find some suitable partitioning n_a, respective n_r of an entire signal $\tilde{x}[n]$. In case of a continuously excited signal, we assume a sustain part to be present in the signal and therefore n_a denotes time frames, covering the attack to sustain region within the signal, whereas n_r denotes the sustain to release region. For an impulsive excited signal in contrast, n_a denotes the attack region only as n_r does for the release region, because a sustain part is assumed to be absent. To determine n_A and n_R we use a simple threshold method applied to the relative energy level over time and distinguish between the continuous and the impulsive case by applying different threshold values. The continuous case is shown in fig. 1 and a suitable partitioning using adjoint bounds to determine n_a and n_r is presented in the inequalities (4) and (5). Here, we use a threshold value below 0 dB, whereas in the impulsive case the threshold is set to 0 dB giving $n_A = n_R$ reflecting the absence of a sustain region. Such a partitioning using adjoint bounds has shown to be favourable, as other variants did not give better results in accuracy and variance of the trained models.

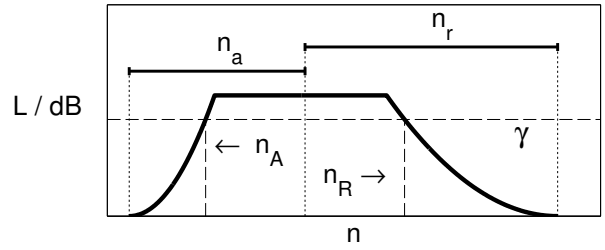

Figure 1. *Partitioning of a signal using a threshold method and adjoined bounds.*

$$n_a : n \leq \frac{1}{2}(n_A + n_R) \tag{4}$$

$$n_r : n > \frac{1}{2}(n_A + n_R) \tag{5}$$

Regarding our signal model the partitioned amplitudes of the partials can be denoted $A[k, n_a]$ and $A[k, n_r]$. The resulting amount of time frames for either the attack to sustain or sustain to release regions will later be referred to by N_A and N_R.

Additionally, as we are only considering quasi-harmonic instruments, the frequency values of the partials will be approximated as being in an integer ratio regarding its fundamental and being constant throughout an entire signal leading to eq. (6).

$$f(k) = f_0 \cdot k \quad , \quad k = 1 \ldots K \tag{6}$$

While f_0 denotes the fundamental frequency, $f(k)$ gives a sequence of frequency values of size K. As a result this approximation significantly simplifies our modeling approach.

2.2 Source-Filter Model

Our approach is based on the distinction of features being correlated to the fundamental frequency f_0 and features being independent of the fundamental. Features correlated to f_0 may refer to characteristics such as odd harmonics being stronger than even ones and therefore are better described as a function of the partial index k instead of actual frequencies. In contrast, formants or resonances refer to f_0 independent features and have to be described explicitly by their frequency values. In our source-filter model we refer to this distinction by expressing the f_0 correlated features within the source and the f_0 independent features within the filter. By this approach, the source will generate an envelope as a function of the partial index and without considering the fundamental, while the filter colors this envelope by taking the partials frequencies into account explicitly.

Source Model By assuming the source to include the f_0 correlated features we use an oscillator model to refleft this. Additionally, in contrast to [4], we assume the variation of the spectral envelope in time to be correlated with the fundamental frequency rather than independent to f_0. Thus, the temporal behaviour of the spectral envelope is assumed to be related to the partial index rather than to actual frequencies. This makes our oscillator dependent on the relative signal level L as well as on the partial index k. By taking into account that the progression of each partial over the relative energy is continuous, we model the partials trajectories using piecewise polynomials. As described in [5], the linear superposition of weighted basic-splines (B-Splines) gives maximally smooth trajectories and B-Spline functions are completely determined by the size of their segments and their order o, denoting the number of segments covered by a single B-Spline polynomial. Due to

linear superposition, the order of the piecewise polynomials follows $o - 1$, therefore the order also defines the degree of smoothness of the B-Spline function. As B-Spline polynomials are defined to converge to zero at their limits, zero size segments are used to model trajectory values at the limits differing from 0. Figure 2 shows a set of 7 B-Splines U_p as functions of level L. So, as we want to model

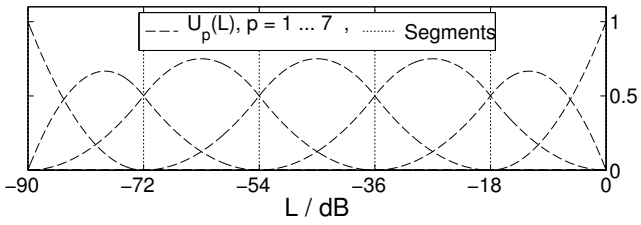

Figure 2. *B-Spline polynomials U_p of order 3 for 5 segments over a level range of -90 to 0dB. 2 zero size segments have been added to both extrema.*

the spectral envelope for a specific range of relative levels of an entire signal, we need to introduce two different oscillators to either express the signal's attack to sustain or sustain to release regions.

$$O(k, L)_A = \sum_{p}^{P} g^A_{k,p} U_p(L) \tag{7}$$

$$O(k, L)_R = \sum_{p}^{P} g^R_{k,p} U_p(L) \tag{8}$$

In equations (7) and (8) the source model for the attack-sustain and sustain-release oscillator is shown – indicated by the subscripts A and R, respectively. Each oscillator value is expressed by the weighted superposition of piecewise B-Spline polynomials for a single partial k at an arbitrary level L. Therefore, for both oscillators the weighting parameters g represent a sequence of coefficients for the piecewise polynomials for each single partial. The overal number of polynomials is denoted by P. Hence, the source generates an entire partial envelope for arbitrary relative signal levels, whereas the progression of a single partial over the relative level is expressed as a continuous trajectory. This perspective holds for the attack to sustain as well as the sustain to release oscillator.

Filter Model The filter covers every part of an instrument not directly associated with the source. This primarily relates to the instruments corpus. As we want the filter to model the features independent of f_0, the filter is assumed to lower or raise the partials amplitude values $O(k, L)$ excited by the source with respect to their actual frequency values $f(k)$. Furthermore, the filter is assumed to be time-invariant and therefore frequency dependent only. Since we are only using information regarding the partials for all instrument signals, all information regarding the filter's frequency response will only be obtained at the frequency positions of the partials. But as with the oscillators partial trajectories, the frequency response is assumed to be continuous, thus we also use B-Splines to model the filter. In contrast to the oscillator models, the B-Spline functions have to be defined in the frequency domain and as we expect the filter curve to exhibit prominent resonance peaks at lower levels and less prominent but dense peaks at higher frequencies, we propose frequency dependent segment sizes. Therefore the segments will be distinguished by multiples of octave bandwidths with factors less or equal to 1, starting with the lowest possible fundamental frequency of each instrument. Consequently, the filter may model resonances and formants at lower frequencies more accurately and averaged within certain bandwidths at higher frequencies. Figure 3 shows 12 B-Spline polynomials V_q as a function of frequency f. So we can finally express the filter as a

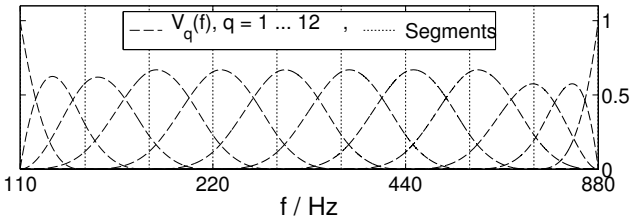

Figure 3. *B-Spline polynomials V_q of order 4 for 3 octaves with a segment size of $1/3$ octave. 3 zero size segments are added at extreme values.*

weighted, linear superposition of B-Splines for a complete set of polynomials denoted by Q.

$$F(f) = \sum_{q}^{Q} z_q V_q(f) \tag{9}$$

The parameters z_q therefore denote a sequence of size Q, weighting the appropriate B-Spline functions. Note, that the actual amount of B-Splines will later be determined by the bandwidth expressed as a fraction of an octave width that will fit into the frequency range beginning with the lowest possible pitch of each instrument up to some maximum frequency value.

3. PARAMETER ESTIMATION

Parameter estimation is applied to the weighting coefficients of the B-Spline functions only. This introduces a model parameterization that is restricted to the amount of segments over a predefined level range for the oscillator models as well as the size of the segments for the filter model given as a fraction of octave bandwidths. Additionally, the B-Spline order has to be set to some suitable value to completely define the B-Spline functions.

As can be seen from equations 7 and 8, U_p depends on the level L. Since we model the partial's amplitude trajectories according to the relative level of the signal, it is obvious to determine $U_p(L[n_a])$ for the attack to sustain and $U_p(L[n_r])$ for the sustain-release regions of each training sample. Similarly, V_q in equation 9 depends on the frequency f and as we have approximated the frequency

values $f(k)$ for all partials of every training sample, we can easily define $V_q(f(k))$. These three functions will remain constant while estimating the weighting parameters and can further be regarded as a projection of the training samples from input space to model space. For convenience and better readability we introduce a matrix/vector notation, shown in table 1. The parameters to estimate are shown at the top, while all data dependent variables are shown at the bottom. Since we make use of a gradient de-

$\mathbf{G}_A$	:	$g^A_{k,p}$	$\mathbf{G}_R$	:	$g^R_{k,p}$
$\mathbf{z}$	:	z_q			
$\mathbf{U}_A$	:	$U_p(L[n_a])$	$\mathbf{U}_R$	:	$U_p(L[n_r])$
$\mathbf{A}_A$	:	$A(k, n_a)$	$\mathbf{A}_R$	:	$A(k, n_r)$
$\mathbf{V}$	:	$V_q(f(k))$			

Table 1. *Matrix/Vector notation conventions*

cent method to estimate the models parameters, we need to define a cost function and its gradients. In this method, the paramaters $\mathbf{G}_A$, $\mathbf{G}_R$ and $\mathbf{z}$ will be adapted iteratively according to their negative gradient of the cost function until the function converges. Finally, after the cost function has converged, the fixed set of model parameters is called a prototype for the instrument, which has been used for estimation.

3.1 Cost Function

For estimation of the model parameters we introduce a squared cost function. As shown in eq. 10 we take the squared value of every single value within the matrices and average over all partials k and time frames n to resolve the equation to a scalar value. Finally, both costs for the attack to sustain and sustain to release sections are averaged.

$$c = \frac{1}{2K}\Big(\frac{1}{2N_A}\sum_{k,n_a}\big((\mathbf{G}_A\mathbf{U}_A + \mathbf{V}^T\mathbf{z}) - \mathbf{A}_A\big)^2 + \frac{1}{2N_R}\sum_{k,n_r}\big((\mathbf{G}_R\mathbf{U}_R + \mathbf{V}^T\mathbf{z}) - \mathbf{A}_R\big)^2\Big) \quad (10)$$

Since this equation gives the cost for a single training sample, we average over all sample costs to measure the cost for the training database.

3.2 Gradient Functions

To get the gradients, the first derivation of the cost function with respect to the parameters has to be solved. This can be done by applying the chain rule once.

$$\frac{\partial c}{\partial \mathbf{G}_A} = \frac{1}{N_A}\big((\mathbf{G}_A\mathbf{U}_A + \mathbf{V}^T\mathbf{z}) - \mathbf{A}_A\big)\mathbf{U}_A^T \quad (11)$$

$$\frac{\partial c}{\partial \mathbf{G}_R} = \frac{1}{N_R}\big((\mathbf{G}_R\mathbf{U}_R + \mathbf{V}^T\mathbf{z}) - \mathbf{A}_R\big)\mathbf{U}_R^T \quad (12)$$

Note, as we want to get the gradients for all K sequences of B-Spline coefficients for the oscillator models, neither averaging over k nor averaging over the two oscillator models has to be done. For the filter coefficients, on the other hand, averaging over both oscillator models remains necessary.

$$\frac{\partial c}{\partial \mathbf{z}} = \frac{1}{2K}\Big(\mathbf{V}\frac{1}{N_A}\sum_{n_a}\big((\mathbf{G}_A\mathbf{U}_A + \mathbf{V}^T\mathbf{z}) - \mathbf{A}_A\big) + \mathbf{V}\frac{1}{N_R}\sum_{n_r}\big((\mathbf{G}_R\mathbf{U}_R + \mathbf{V}^T\mathbf{z}) - \mathbf{A}_R\big)\Big) \quad (13)$$

4. MODEL VARIANCE

To evaluate the performance of a prototype we define a variance measure with respect to some arbitrary relative level sequence L^σ for the attack to sustain as well as sustain to release regions. Since we have modelled the oscillators as being dependent on the relative level of a signal, we are free to define arbitrary level values to measure the related amount of variance. These variances shown in 14 and 15 will be estimated on the test database that has not been used to train the model. Both equations make use of $\mathbf{U}$, which is defined to be the B-Spline values of the oscillator models for the arbitrary level sequence L^σ, therefore $\mathbf{U} : U_w(L^\sigma)$. Note, as the squared term refers to the dimensions of k over n_a or rather n_r, $\mathbf{U}_A^T$ as well as $\mathbf{U}_R^T$ apply the transformation from this data specific input space to the oscillator model space of $[kxP]$, while $\mathbf{U}$ finally transforms the variance to the space spanned by our arbitrary level sequence.

$$\sigma^2(L^\sigma)_A = \frac{1}{KN_A}\sum_k\big(\mathbf{A}_A - (\mathbf{G}_A\mathbf{U}_A + \mathbf{V}^T\mathbf{z})\big)^2\mathbf{U}_A^T\mathbf{U} \quad (14)$$

$$\sigma^2(L^\sigma)_R = \frac{1}{KN_R}\sum_k\big(\mathbf{A}_R - (\mathbf{G}_R\mathbf{U}_R + \mathbf{V}^T\mathbf{z})\big)^2\mathbf{U}_R^T\mathbf{U} \quad (15)$$

Since these equations give the variance between a prototype and a single test sample, measuring the model's variance onto a whole test database requires averaging over all test samples.

5. RESULTS

To evaluate our approach for a generalized model of quasi-harmonic instruments we have used various model parameterizations regarding the definition of the B-Splines for the oscillator models as well as the filter model. More precisely we have estimated the model performance for 5, 10 and 20 segments for the B-Splines functions of the oscillator models with a fixed order of 3 and also segment widths of 1/12, 1/4, 1/2, 3/4 and 1 multiples of an octave for the B-Spline functions of the filter model with a fixed order of 4. As the B-Spline order defines the degree of smoothness of the final superposition of the B-Spline functions, we use an order of 3 for the oscillator model to tolerate highly variable spectral envelopes with still smooth transitions and an

order of 4 for the filter model to achieve a resulting resonance curve with some higher degree of smoothness. This creates a total of 15 parameter configurations and for each a 10-fold cross-validation method has been used. The instruments sample database used for training as well as a variance estimation has been taken from the RWC musical instrument database [2]. This database contains three variants for a single instrument, each played by a different instrumentalist, giving us the possibility to achieve a high degree of generalization for their expected spectral envelope. We have further implemented an on- and offline method for parameter estimation and evaluated various initialization values for the B-Spline coefficients.

Finally, it has been shown, that online estimation performs up to 5 times faster than its offline counterpart, making it the favourable estimation method. Online estimation tended to converge in cost already after 10 to 25 epochs over the training database, while the offline method needed more than a 100 epochs. The number of epochs being for convergence can significantly be reduced by incorporating a priori knowledge about the characteristics of the filter as well as the oscillator while initializing the coefficients prior to estimation, but even a random initialization gives comparable results apart from the number of epochs needed. For visualization as well as fpr measuring the variance of the prototypes, we use an arbitrary sequence of relative level values shown in eq. (16).

$$L^\sigma = \{-30, -27.5, \ldots, -2.5, 0\} \tag{16}$$

This sequence is used to process the respective partial envelopes $O(k, L^\sigma)$ generated by the oscillator models. Therefore, each single level value gives a partial envelope for either the attack to sustain or sustain to release oscillator as well as an associated variance value. Figures 4, 6 and 8 present prototypes for a clarinet, a grand piano, and a violin. At the top, the partial envelopes generated by the two oscillators are shown with respect to the partial index k. The various dotted lines refer to different values of the level sequence L^σ, whereas for the sustain to release region the sequence is resolved in reversed order. The resonance curve of the filter is shown at the bottom of each figure together with the B-Spline functions V_q and their coefficients z_q used to generate the final curve. Moreover, figures 5, 7 and 9 show the variances measured for the selected level values of the sequence L^σ for the attack to sustain as well as sustain to release part.

As shown in the figures for the prototypes of the selected instruments, various f_0 independent resonances have been estimated by the filter as well as f_0-correlated features by the oscillator models. Moreover, the time-varying spectral envelope has also been reflected by the partial envelopes of the oscillator models. However, as our proposed model has been designed to cover all characteristics of the spectral envelopes of an instrument in a generalized manner, all features estimated also refer to general characteristics of such an instrument.

In case of the clarinet shown in fig. 4 a very salient characteristic of the partial envelope has been estimated, a

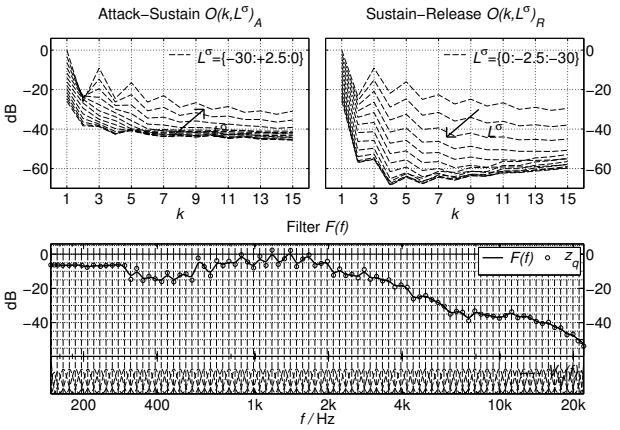

Figure 4. *Prototype of a clarinet using 5 B-Spline Segments for the oscillator and a segment size of 1/12 octave for the B-Spline modelling of the filter.*

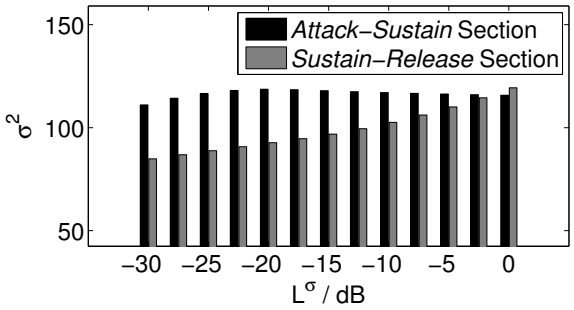

Figure 5. *Variances for the clarinet prototype.*

strong damping of around 10 dB between 300 and 500 Hz as well as a f_0 independent decay of 10 dB/octave above 2 kHz. The appropriate variances indicate a constant model performance over the full level range of -30 to 0 dB and even an improving accuracy with a decreasing level during the release.

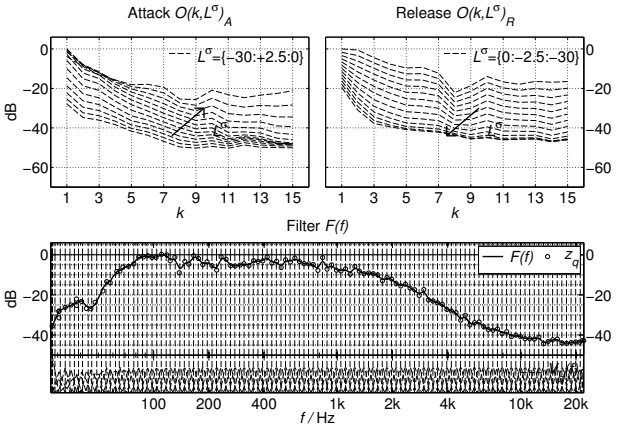

Figure 6. *Prototype of a grand piano using 5 B-Spline segments for the oscillator and a segment size of 1/12 octave for the B-Spline modelling of the filter.*

The resonance characteristic of the grand piano shows a distinctive formant structure and a pronounced damping

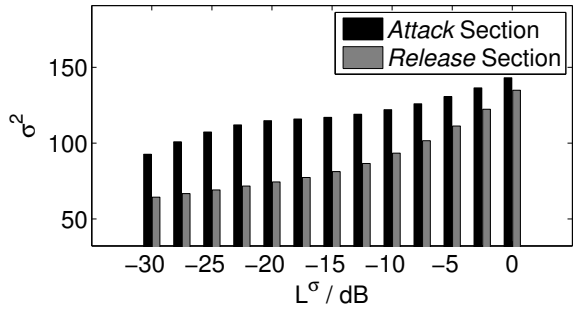

Figure 7. *Variances for the grand piano prototype.*

behaviour at low and high frequencies. As depicted in 7 the prototypes performance increases while the signals energy decreases during its release. This decrease in energy actually covers most of a piano sound, due to its impulsive excitation, which usually has a short attack and no sustain region. Therefore, the prototype performs best within the prevailing signal part. The violin prototype performs

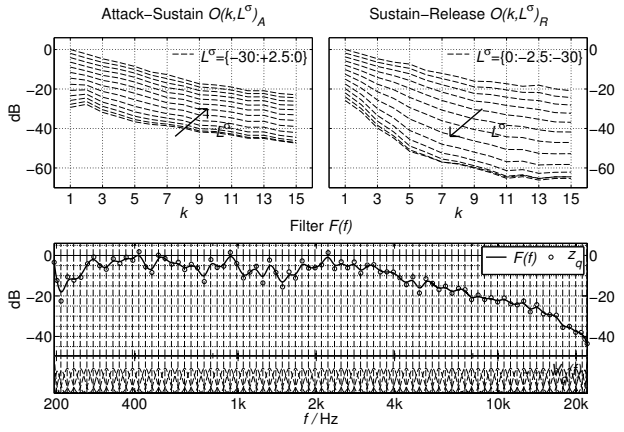

Figure 8. *Prototype of a violin using 5 B-Spline segments for the oscillator and a segment size of 1/12 octave for the B-Spline modelling of the filter.*

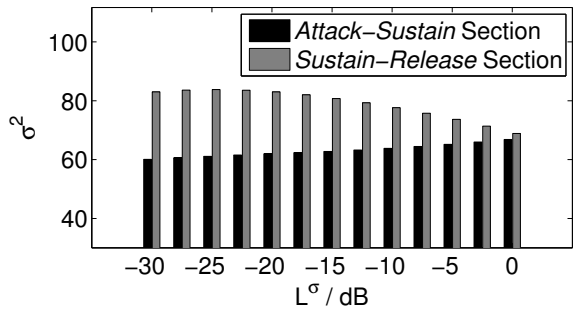

Figure 9. *Variances for the violin prototype.*

best in terms of absolute variance values, though it shows slightly less variance during the attack to sustain region. With such a low variance value we can assume the highly variable resonance structure of the filter to be accurate and close the general characteristic of a violin.

6. CONCLUSION

In this paper we have shown a new approach for representing quasi-harmonic instruments by a generalized source-filter model with the possibility to predict the time-varying spectral characteristics of an instrument with respect to the partial index and relative level of the signal. We have further given the mathematical basis to estimate the models parameters from a given training database and to evaluate the model using a test database. As shown by our results, the selected prototypes have estimated f_0-correlated as well as f_0 independent features with high accuracy and can presumeably be well generalized. Hence, we believe our results are promising and we will apply our modelling approach in future research to various applications regarding instrument recognition and source separation.

7. ACKNOWLEDGEMENTS

Many thanks to all reviewers.

8. REFERENCES

[1] Juan José Burred, Axel Röbel, and Xavier Rodet. An accurate timbre model for musical instruments and its application to classification. In *First Workshop on Learning the Semantics of Audio Signals*, Athens, Greek, December 2006.

[2] Masataka Goto and Takuichi Nishimura. Rwc music database: Music genre database and musical instrument sound database. In *4th International Society for Music Information Retrieval Conference (ISMIR)*, pages 229–230, 2003.

[3] Toni Heittola, Anssi Klapuri, and Tuomas Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 327–332, 2009.

[4] Anssi Klapuri. Analysis of musical instrument sounds by source-filter-decay model. *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 1:I–53–I–56, 2007.

[5] Axel Röbel. Adaptive additive modeling with continuous parameter trajectories. *IEEE Transactions on Speech and Audio Processing*, 14-4:1440–1453, 2006.

[6] X. Serra. *Musical Signal Processing*, chapter Musical Sound Modeling with Sinusoids plus Noise, pages 91–122. G. D. Poli, A. Picialli, S. T. Pope, and C. Roads Eds. Swets & Zeitlinger, Lisse, Switzerland, 1996.

[7] Tuomas Virtanen and Anssi Klapuri. Analysis of polyphonic audio using source-filter model and non-negative matrix factorization. In *Advances in Models for Acoustic Processing, Neural Information Processing Systems Workshop (AMAC)*, 2006.