Extended Source-Filter Model of Quasi-Harmonic Instruments for Sound Synthesis, Transformation and Interpolation

Henrik Hahn, Axel Röbel

Extended Source-Filter Model of Quasi-Harmonic Instruments for Sound Synthesis, Transformation and Interpolation

Henrik Hahn
IRCAM-CNRS UMR 9912-STMS
Paris, France
henrik.hahn@ircam.fr

Axel Röbel
IRCAM-CNRS UMR 9912-STMS
Paris, France
axel.roebel@ircam.fr

ABSTRACT

In this paper we present a new technique for sample-based sound synthesis. The approach comprises the analysis of sounds of an instruments sound database, a parameter estimation for an instrument model and a sound synthesis using this model together with the analyzed sounds. The analysis of the sounds is carried out by the segregation of each sound into a sinusoidal and noise component and extracting certain control parameters from both. The components will be modeled using an extended source-filter model, whereas the harmonic component will be represented by a non-white source and a resonator filter and the noise component by a single filter. Model parameters are represented by means of weights of tensor product B-splines (basic-splines) covering the instruments sound characteristics over its full pitch range, global intensities and the sounds temporal evolution. This structured sound representation will allow enhanced source filter based sound manipulations. The paper concludes with a subjective evaluation presented for comparison with state of the art sound transformations.

1. INTRODUCTION

The purpose of the present article is to establish a compact model of the source and filter characteristics of a complete musical instrument, including all the modifications of the spectral color that are related to changes of the intensity and/or pitch. The target application for this research is expressive sound synthesis in music samplers that provides control parameters for dynamic intensity and pitch changes that produce realistic sound changes that relate to the parameter transitions.

Source filter models are often used for physical modeling of musical instruments [1] and sound transformations [2]. These source filter models, however, are not learned from sound signals, and generally, due to their specific form, they cannot be used to control and improve results of sound transformations. The huge amount of instrument sample databases that are available today (RWC [3], Vienna Symphonic Library, IRCAM/Univers sons Solo Instruments) opens a new approach to sound representation that consists of training instrument sound signal models with physically meaningful parameters using the available sound databases as training data. Related approaches to instrument modeling have been used recently notably in the context of polyphonic instrument recognition and sound separation [4], [5], [6], [7]. Especially interesting in the present context is the approach described in [4] because there, similar to the joint estimation of glottic pulse and vocal tract for speech signals, a joint estimation of source and filter allows estimating excitation signals that are physically reasonable and not necessary white. These source filter models of musical instruments can be used to guide the representation and transformation of musical sounds of solo instruments. The separation of sinusoidal and noise signal components [8], [9] extends the source filter approach, because filters for sinusoidal and noise signal components can be separately established. In our work we followed a similar extension of the source filter model as proposed by [4] targeting however the synthesis of musical instrument sounds from control parameters that cover pitch and intensity contours as well as global note intensity. Initial results related to the instrument model representation using an extended source filter model including non white source and a fixed resonator filter have been presented in [10]. In the following section we describe a significantly refined version of the approach including a pitch and global as well as local intensity dependent source/filter model and we will describe first results obtained from signal synthesis from the models showing that the model allows to faithfully reproduce the sound modifications related to intensity and pitch changes. The article is organized as follow.

Section 2 will describe the general system to do sound transformations based on the source-filter approach using trained instrument models. In sections 3 we will present the analysis of a database of instrument sounds and in section 4 a detailed description will be given how to establish an instrument model using tensor product B-splines. The estimation of model parameters will be discussed in section 5 and the synthesis of the sounds will be shown in section 6. An evaluation of our approach will be presented in section 7 followed by a conclusion.

2. SYSTEM OVERVIEW

An overview of the system we created is given in Figure 1. Our approach first needs a database of recorded in-

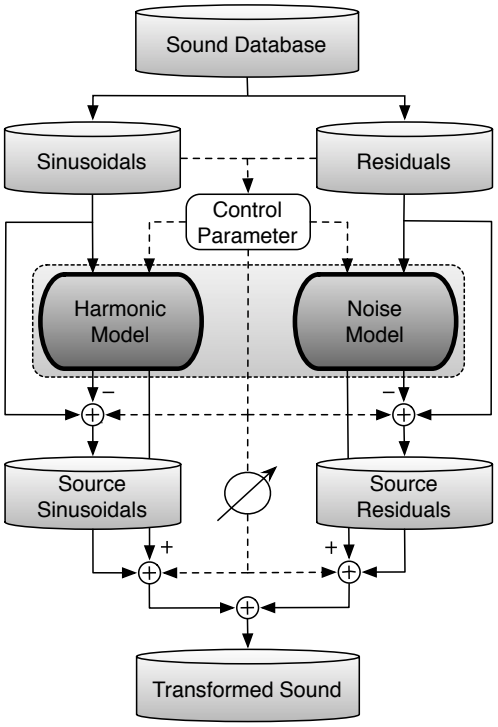


Figure 1. System overview.

struments sounds covering the whole pitch range of the instrument and also providing several intensity levels for each pitch. Every recording must have a fixed pitch and a certain intensity. As our approach assumes that harmonic and noise signal components need to be treated separately, all input sounds from the database will be segregated into their deterministic and stochastic components. Additionally, several control parameters will be extracted from the input sounds, representing the criteria for which varying sound characteristics may be captured by the models. The model adaption scheme therefore takes a database of analyzed input sounds as well as their respective control parameters. The adapted models will then be used to extract white source signals from the input signals by removing the captured instrument characteristics estimated by the model from each single sound. Accordingly, these source signals will only contain information which is not covered by the instrument model. Therefore the source residual signals will be almost white noise, hence the harmonic source signals will be almost white discrete signals containing pure sinusoids. The filter functions for all transformations can then be obtained from both models by using control parameters according to a desired sound and applying the estimated filters on interpolated versions of the source signals to get the target sound.

3. SOUND ANALYSIS

Assuming a sound can be fully represented as the sum of time-varying sinusoids and a residual signal as shown in (1) we utilize an analysis method based on [8] to obtain these signal components from an input signal.

$$y(n) = \sum_{k}^{K} a(k,n)cos(\phi(k,n)) + \epsilon(n) \qquad (1)$$

In equation (1), k reflects the index of the sinusoid and n denotes the frame index. Finally, all sinusoidal amplitudes are transformed to decibel domain denoted $A(k,n)$. Their slowly varying frequency values will be written $f(k,n)$.

The residual noise signal, obtained by subtracting the synthesized sinusoids from the input signal is further processed by means of estimation of the time-varying spectral envelope. Since we assume the residual signal to be stochastic we process the filtered cepstrum [11] with an order set to $L \leq \frac{f_s}{2f_0}$ and denote the cepstral coefficients $C(l,n)$ with $1 \geq l \leq L$ and n represents again the frame index. The filtered cepstrum can be considered to be a smoothed version of the amplitude spectrum, but as its following the mean of the amplitude spectrum it estimates the signals energy and while filtering all components above L the envelope becomes smooth enough to not follow the spectral gaps introduced by subtracting the sinusoids from the original input signal.

4. ANALYSIS OF INDEPENDENT PARAMETERS

We identified four control parameters being independent of each other and separately adjustable with a considerable impact on an instruments sound. These parameters therefore have to be taken into account explicitly to establish an instrument model covering a significant amount of the sound characteristics of a quasi-harmonic instrument.

4.1 Pitch m

The pitch of each single instrument recording given as MIDI value in our case can be read from meta tagged file names, but could also be obtained from the input signal by means of pitch estimation.

4.2 Global Intensity I_g

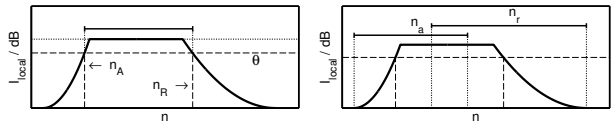
We denote the overall loudness of each signal as its global intensity. The database we use includes meta tags for each sound file indicating one of the three global intensity levels *pp*, *mf* and *ff*.

4.3 Local Intensity I_l

The evolution of the energy over time n is processed separately for the sinusoidal and noise components. These functions will be denoted as local intensity $I_l(n)$ expressed in decibel and normalized to a maximum of 0dB to describe each sound component in relation to its maximum energy.

4.4 Temporal Segmentation $s \in \{1,2\}, n_a, n_r$

The segmentation of an instrument sound into its attack, sustain and release states is performed by analyzing the local intensity function using an adaptive threshold θ shown in Figure 2a. The threshold will be set to some reasonable value below the average energy value of the signals



(a) Attack/Release determinination (b) Segmentation into 2 overlapping segments

Figure 2. Temporal segmentation scheme.

sustain region to determine the signals attack (n_A) and release point (n_R). Within the instrument model but also for sound transformation we do not need an exact estimate of the attack or release time frame since we are only modeling two segments as illustrated in Figure 2b. The segment $s = 1$ using frame indices n_a is covering the signals onset to some point within the signals sustain region while segment $s = 2$ obtaining frame indices n_r covers a region starting in the middle of the sustain and ending with the signals offset. The indices are obtained using eq. (2) and (3) considering N to be overall amount of frames within the signal.

$$1 \leq n_a \leq \frac{1}{3}\left(n_A + 2\, n_R\right) \tag{2}$$

$$N \geq n_r \geq \frac{1}{3}\left(2\, n_A + n_R\right) \tag{3}$$

5. INSTRUMENT MODEL

Following the assumption that instrument sound changes due to pitch or intensity modifications affect their harmonic and noise component differently, we establish two models to separately represent their sound characteristics. The harmonic model will be expressed by partial amplitudes represented in decibels, whereas the noise model will be established in cepstral domain modeling cepstral coefficients directly, but both models will incorporate the temporal segmentation following section 4.4.

5.1 Harmonic Model

For a model of the harmonic content of an instrument sound we establish an extended source filter approach as in eq. 4, assuming the source S to be a function of the partial index k and temporal segment s depending on the control parameters $(\ldots)$ and a resonance filter R depending on the partials frequencies only.

$$\hat{A}^{k,s}(\ldots, f(k)) = S^{k,s}(\ldots) + R(f(k)) \tag{4}$$

In [10] we introduced a harmonic model employing a source and a resonator filter expressing features correlated with f_0 by a source and f_0 independent features by a resonance filter. Using this distinction the source filter can also be described as a non white excitation filter representing a vibrating string or air pipe, whereas the resonator will exhibit all signal components not directly related to the excitation. This primarily refers to the corpus of a musical instrument. According to this approach, the source will generate an envelope as a function of the partial index k for a certain temporal segment and without considering the fundamental, while the resonance filter colors this envelope taking the frequencies of the partials into account explicitly.

In the formerly published approach the time-varying characteristics of an instrument sound has been assumed to be directly related to the temporal evolution of the local intensity. We further considered the sounds temporal variations to be reflected by the source as they are assumed to be correlated with the fundamental. In our current approach we extend this model by facilitating the source filter to also incorporate variations according to the signals global intensity. This is straightforward since the source filter already expresses the signals excitation and variations of the global intensity of instrumental sounds are mainly achieved by varying its excitation. Moreover, we endow the source filter to slightly vary with the sounds fundamental frequency to reproduce sound variations due to pitch variations, not originating in the instruments resonator, but a certain change in the excitation of the signal. Considering our first distinction of expressing features correlated to or independent of f_0 separately, this can be contradictory and therefore it will be mandatory to enforce the model to capture only slow changes over pitch within the source filter while retaining the resonator to reflect rapid changes in frequency. This will be discussed in detail in section 6.1.

Given this requisitions for the source filter to model the f_0 correlated features they can be expressed by their partial indices as these features are in a logical relation to its fundamental neglecting their actual frequency. Taking into account our postulate of separate models for the attack to sustain and sustain to release phase of a signal, each partial therefore needs to get an individually modeled trajectory as a function of global intensity I_g, local intensity I_l and pitch m for both temporal segments. To smoothly model such multi-dimensional trajectories we propose to use one-dimensional B-splines extended to multiple dimensions by using a tensor product. The spline subspace is defined for each single dimension as usual while the tensor product function for our three-dimensional case can be expressed as in eq. (5).

$$\begin{aligned} S^{k,s}(I_g, I_l, m) &= \sum_{p,q,t}^{P,Q,T} B_p(I_g) B_q(I_l) B_t(m) \quad \cdot \gamma_{p,q,t}^{k,s} \\ S^{k,s}(I_g, I_l, m) &= \sum_{u}^{U} \mathcal{B}_u^S(I_g, I_l, m) \quad \cdot \gamma_u^{k,s} \end{aligned} \tag{5}$$

Such tensor products serve as a straightforward and simple generalization of one-dimensional basis functions to the n-dimensional case. As shown in eq. (5) a hyperplane to model the characteristic for a single partial k in either temporal segment s is constructed by P B-spline functions along the axis for global intensity I_g, Q B-spline functions along I_l and another T basis functions covering the pitch dimension. To simplify this expression we will make use of $\mathcal{B}_u^S$ with respective weights $\gamma_u^{k,s}$ to indicate the usage of a tensor product spline whereas ^{S} indicates *source*. An

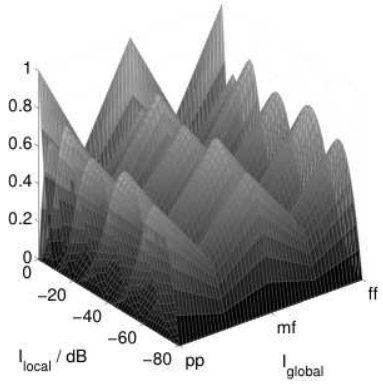


Figure 3. Two-dimensional tensor product B-spline model for I_g and I_l modeling a surface using a 2nd order B-spline model with 2 segments along I_g and a 3rd order along I_l with 5 segments.

example for the two-dimensional case using tensor product B-splines for I_g and I_l is illustrated in Figure 3.

Contrary to the source, the resonator filter will be constant throughout the temporal evolution of an instrument sound and only be dependent on the partials frequency, but since the partials will only reveal a sampled version of the real instruments resonance characteristic we propose to use a one-dimensional B-spline to model a continuous filter function as shown in eq. (6) with spline functions B^R indicating resonance filter splines. However the maximum bandwidth of the resonance filter cannot exceed the range between the lowest and highest partial frequency existing in the instruments sound database.

$$R(f(k)) = \sum_{v}^{V} B_v^R(f(k)) \cdot \lambda_v \tag{6}$$

Assuming a musical instrument having pronounced and distinct resonances in the lower or middle registers and less prominent but dense in the upper, we utilize the well-known Mel-scale to define the bandwidth of the B-spline segments along the frequency axis.

5.2 Noise Model

The noise component of an instrument sound will contain all the non-harmonic content encompassing various wind sounds for brass and woodwind instruments or bow noise for string instruments for example. In contrast to the harmonic model we use a classical source filter approach to model the noise component of an instrument sound, assuming the noise to be produced by a source and colored using a single filter. This filter will however be modeled similar to the source filter for the harmonic model taking into account varying characteristics dependent on the global and local intensity as well as the pitch. Therefore it will utilize the same multi-dimensional B-spline model but with different weights δ, though in contrast we are not modeling an explicit noise envelope, but the cesptral coefficients directly.

$$\hat{C}^{l,s}(I_g, I_l, m) = \sum_{u}^{U} \mathcal{B}_u^S(I_g, I_l, m) \cdot \delta_u^{l,s} \tag{7}$$

As illustrated by eq. (7) each cesptral coefficient l is modeled separately for the attack to sustain and sustain to release segment of a signal according to the intensity parameters and the signals pitch.

6. PARAMETER ESTIMATION

For the instrument models to capture the timbral characteristics of a certain musical instrument, their B-spline weighting parameters γ, λ and δ need to be estimated using a database of sounds of this instrument. This trainings data consequently will be a set of $A(k, n)$ for the harmonic as well as a set of $C(l, n)$ for the noise model, but as our model reflects the characteristics of a sound according to the control parameters I_g, I_l, m and s, they will also be needed by the adaption scheme for each input sound. The estimation of the parameters for the harmonic and noise model can be solved independently, since both signal components have their own timbral characteristics.

Estimating the parameters means conducting a regression analysis and as both models are linear, a global optimum will exist. In case of the noise model the global optimum is unique, but for the harmonic model the sum of the two filters introduces a manifold of optimal solutions, because every solution for the two filters can also be expressed with a constant added to either one of them and subtracted from the other. A method to resolve this ambiguity will be shown in 6.1.

Finding the solution for a linear optimization problem usually is based on creating a set of linear equations as $\mathbf{A} = \mathbf{Mx}$ and somehow solving $\mathbf{M^{-1}}$ to compute $\mathbf{x}$. In case of the harmonic model this it hardly possible to solve, as all model parameters γ and λ have to be estimated jointly due to their interconnection within the resonator filter and therefore the size of the transformation matrix will become extraordinarily large. For small databases with 100 sounds of intermediate length and around 100 partials to model, the matrix will already be around 10GB large using only a few B-spline components for each independent parameter, but for larger databases and more complex models M can easily exceed several TB. In case of the noise model, the transformation matrix M can easily be constructed for each single cepstral coefficient of either temporal segment independently with much less memory demand and inverted to obtain $\delta^{l,s}$, but this has been shown to be not numerically robust. Therefore we use the conjugate gradient (CG) method [12] to estimate the model parameters in an iterative manner for both models. Utilizing any gradient decent method implies constructing an objective function and minimizing it regarding the models parameters. To establish such, eq. (8) denotes, how to use the instrument model, to estimate a partials amplitude given the input parameters regarding a certain instrument sound. Equation (8) can also easily be rewritten for the noise model.

$$\hat{A}(k, n_s) = \hat{A}^{k,s}(I_g, I_l(n_s), m, f(k, n_s)) \tag{8}$$

Using eq. (8) the objective functions (9) and (10) can be established in a least squares sense for a single instrument sound, taking into account the temporal segmentation s and the evolution of all partials k or cepstral coefficients l respectively along time frames n_s.

$$\mathcal{O}_h = \frac{1}{2}\sum_{s=1}^{2}\sum_{k,n_s}^{K,N_s}\Big(A(k,n_s) - \hat{A}(k,n_s)\Big)^2 \quad (9)$$

$$\mathcal{O}_n = \frac{1}{2}\sum_{s=1}^{2}\sum_{l,n_s}^{L,N_s}\Big(C(l,n_s) - \hat{C}(l,n_s)\Big)^2 \quad (10)$$

Calculation of the gradients of (9) and (10) according to the model parameters $\gamma_u^{k,s}$, λ_v and $\delta_u^{l,s}$ is simple and straightforward.

6.1 Regularization

The inherent ambiguity of the two filters of the harmonic model being processed at once can be solved by introducing a regularization term fixing either filter function around an arbitrary value. Due to the multi-dimensionality of the source filter, its preferable to bound the resonance filter. Eq. (11) fixes the resonance filter around 0dB while estimating the model parameters by penalizing filter configurations not centered around the desired value.

$$\mathcal{R}_1 = \epsilon_1 \sum_{f}^{F} \big(R(f)\big)^2 \quad (11)$$

Here, f defines a frequency grid at which the resonance filter is being evaluated and ϵ_1 denotes a regularization factor and again, its derivative is simple and straightforward.

A second ambiguity had been introduced by modeling the source filter in dependance of the pitch, resulting in an ambiguity with the resonance filter. To resolve we utilize eq. (12) to penalizes strong differences between neighboring B-spline coefficients along the pitch dimension and therefore favor only slight amplitude changes over pitch for each partial.

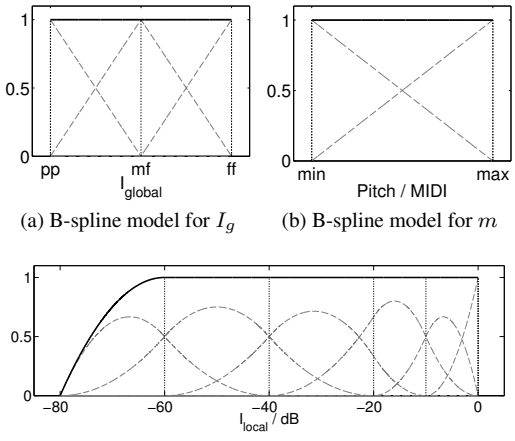
$$\mathcal{R}_2 = \epsilon_2 \sum_{k,s}^{K,2}\sum_{p,q}^{P,Q}\sum_{t=1}^{T-1}\Big(\gamma_{p,q,t+1}^{k,s} - \gamma_{p,q,t}^{k,s}\Big)^2 \quad (12)$$

This regularization is as long as being used for a linear function a correct approximation of its derivative, since it only takes its coefficients into account, but as only slight variations for the pitch model are favored anyway, this restriction is acceptable.

6.2 Model Selection

In this section we describe a potential setting for the instrument models to capture an instruments timbral characteristic given a specific set of control parameters.

In our case, the instrument database had been labeled with three different values for the global intensity of each single instrument sound, namely *pp*, *mf* and *ff*. Three discrete points along the I_g-axis therefore will restrict us from using more than three B-spline functions for the system to be well-defined. Moreover, due to this restriction, the system can not be more complex than linear and therefore we propose to use a B-spline model as illustrated in Figure 4a. In section 6.1 we showed, due the ambiguity between our source pitch model and the resonance filter, the need for a linear model along m, hence Figure 4b demonstrates a B-spline model representing just a linear function over the varying pitch.



(a) B-spline model for I_g (b) B-spline model for m

(c) B-spline model for I_l

Figure 4. All three B-spline models for the source filter. Spline functions are shown dashed, position of knots is depicted with dotted lines and summed splines are displayed as solid line.

Establishing the B-spline models for the local intensity as well as the resonator filter requires a deeper understanding of our instrument model. It is important to realize that the source filter will create a hyperplane for each single partial over all global as well as local intensities, even though they might never appear for certain values. This will surely happen at higher order partials, who will never appear in *pp* sounds or at lower values of the local intensity. This makes the source filter inevitably underdetermined for higher order partials and they therefore would converge to 0 at regions where the model never have seen data, denoting very high amplitude values. In order to avoid this, they need to get faded to very low amplitude values already within the model, thus we introduce a general offset for the instrument model together with a B-spline model for the local intensity without its lowest spline, forcing all partial amplitude values to fade to the offset value, when the local intensity reaches its minimum depicted in Figure 4c.

For the harmonic model, the offset can easily be added to the two filters (eq. (13)) and as the resonance filter is bounded around 0dB, its reasonable to set the offset for all partials to some value reflecting the maximum dynamic across all partials.

$$\hat{A}^{k,s} = S^{k,s} + R + \Theta_h \quad (13)$$

Since the noise model on the contrary models cepstral coefficients and not amplitudes, only the first coefficients will have a significant impact on the envelopes amplitude, whereas higher order coefficients only denote the envelopes

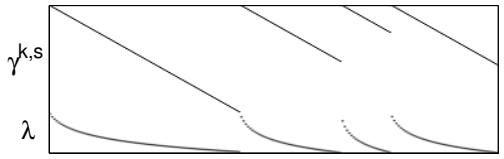


Figure 5. Approximative transformation matrix M showing 4 sound examples.

variation over frequency. Using the same B-spline model for the local intensity of the noise model, we only need to add an offset to the first cepstral coefficient indicating a minimum noise level and use a 0 offset for all other coefficients (eq. (14)). This will finally enforce the noise model to converge to a constant envelope over frequency at the minimum of the local intensity.

$$\hat{C}^{l,s} = S^{l,s} + \Theta_n^l \tag{14}$$

The B-spline model for the resonator filter will define its resolution over frequency. The more complex the model, the more accurately resonances can be modeled, but as the partials frequency values only reveal a sampled version of the instruments resonance characteristics, its resolution will be limited. To find a maximum resolution we use an approximation of the transformation matrix $\mathbf{M}$ to measure if it is determined. To approximate, we first reduce the source filter to a constant model, neglecting varying values for I_g, I_l, m and s, for all partials, because we already know that the source filter will be underdetermined. Second, all partials frequency values will assumed to be constant over time, therefore all time frames within the matrix will collapse to a single value. This makes the transformation matrix rather small and we can compute the rank of the matrix. In fig. 5 a transformation matrix M is shown.

Each point in the linear lines in the upper part of the matrix represent the source value for a single partial. Several lines indicate several sound files. The lower part illustrates the frequency position of each partial, by means of the spline functions v hit by the frequency value $f(k)$ of the current partial k. Computing the rank of this matrix will at maximum give a value $m-1$ (m being the amount of rows) reflecting the mentioned ambiguity between the two filters but also indicating a well defined resonator, as the sampled resonance function can be uniquely expressed by the spline functions. Consequently, the highest amount of spline functions for the resonator filter with a matrix rank of $m-1$, guarantees a unique solution for the resonance filter.

In Figure 6 a model for a resonator filter is shown using a knot sequence of only 25 elements, but increasing distance in between them according to the Mel-Scale.

7. SOUND SYNTHESIS

The approach for sound synthesis using the extended source-filter model is based on 2 consecutive steps. The first step utilizes the model to process the source signals while in the second step the model will be used to accomplish the final sound transformation.

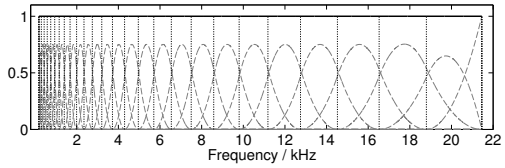


Figure 6. Spline model for the resonance filter. Spline functions are shown dashed, position of knots is depicted with dotted lines and summed splines are displayed as solid line. Segment sizes increase with frequency as Mel-scale filterbank bandwidth increases.

To create the source signal we in fact remove the timbral characteristics captured by the instrument model creating almost white source signals. In case of the harmonic component this refers to discrete harmonics while for the noise component this sound will be actually almost white noise. These source signals therefore will only contain information deviating from the estimated instruments sound characteristics and hence carry information which later makes the synthesis sound natural.

For the creation of the source signals, the input sinusoidal as well as the residual signals need to be filtered with the inverse of the estimated signals by the instrument models. For the harmonic signals this can be done in the sinusoidal domain by subtracting the estimates as shown in eq. 15.

$$\bar{A}^{k,s}(I_g, I_l, m, f(k)) = A^{k,s} - \hat{A}^{k,s}(I_g, I_l, m, f(k)) \tag{15}$$

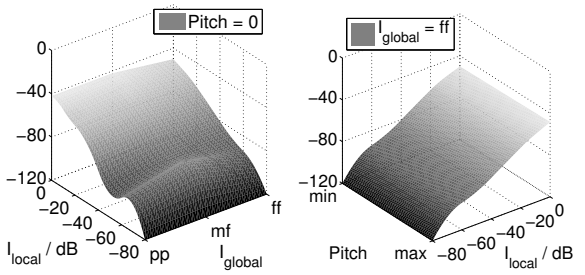
The noise signals however need to get filtered in time or spectral domain. We therefore generate the inverse spectral envelopes from the estimated cepstral coefficients and apply spectral domain filtering using the IRCAM software *superVP*.

To interpolate between two source signals a mixing factor a needs to be defined first, denoting the amount of either source signal accounting to the mix. The factor is first used to time align the sounds to interpolate by processing the target length of the resulting signal as in eq. 16, T_t denoting the target length of the interpolated sound and T_1, T_2 denoting the lengths of the sounds to interpolate respectively.

$$T_t = (T_1 * (1-a) + T_2 * a) \tag{16}$$

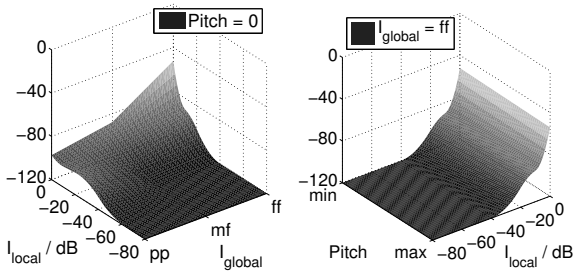
We again make use of *superVP* to apply the time-stretching for the noise signal, but as they do contain only few spectral information and as the stretch factors hardly become large this is uncritical, hence time-stretching can be applied without envelope preservation. The sinusoidal signals instead can be time-aligned in sinusoidal domain by means of stretching the frame positions $n \rightarrow n'$ of $\bar{A}$ without altering the actual partial amplitudes or frequencies.

Mixing the time-aligned source signals again has to be done in either sinusoidal- or time-domain. For the noise signals this can be done in time-domain adding time-domain data, but as they are stochastic processes, their energy needs to be retained. The harmonic signals instead are mixed in sinusoidal domain using again eq. 16 but substituting



(a) $k = 1$

(b) $k = 1$



(c) $k = 40$

(d) $k = 40$

Figure 7. Source filter functions of two partials for the attack to sustain segment of the trumpet. The z-axis depict the amplitude values for the specified partial.

T by partial amplitudes $\bar{A}(k, n_s)$ and normalized partial frequency $f(k, n_s)/f_0$. Partials not present in one of the source signals but in the other will treated as having 0 amplitude in this signal and a frequency value according to $f_0 \cdot k$. This will result in the partial appearing right when the mix factor starts favoring the signal containing this partial. The mixed source signals now need to get the the estimated envelopes applied on using altered control parameters I'_g, I'_l, m' and $f(k)'$ as in eq. 17.

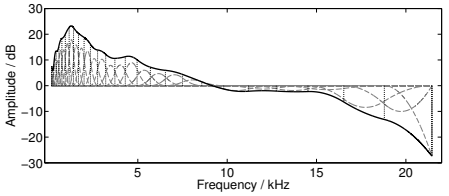
$$\begin{aligned}\tilde{A}^{k,s}(I'_g, I'_l, m', f(k)') = \bar{A}^{k,s}(I_g, I_l, m, f(k)) \\ + \hat{A}^{k,s}(I'_g, I'_l, m', f(k)')\end{aligned} \tag{17}$$

To finally get the desired sound, the harmonic signal needs to be synthesized using additive synthesis and linearly interpolated in the overlapping region of attack to sustain and sustain to release segments and added to the mixed and colored noise signal.

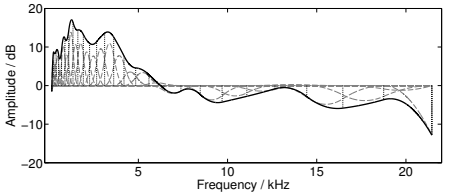
8. EVALUATION

To evaluate our approach we trained instrument models for a trumpet and a clarinet using the specifications for the model given in section 6.2, as these instruments do not need further investigations during analysis or synthesis.

Some estimated source filter functions are shown in Figure 7. Figures 7a and 7b illustrate the estimated function of the fundamental and Figures 7c and 7d for the 40th partial of the trumpet for the instruments whole pitch range denoted by *Pitch*, its range of possible global intensities *pp* to *ff* as well as for a signals temporal evolution expressed by the evolution of the local intensity I_g.



(a) Trumpet



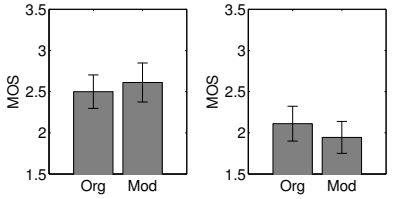
(b) Clarinet

Figure 8. Estimated Resonance Filter.

Both models have been used to create pitch transpositions as well as changes of the global intensity to apply them to the most common use cases within a music sampler. For the transposition we have chosen distances of 12 and 24 semitones to interpolate between them, whereas for the intensity interpolation one interpolation layer between each intensity value had been chosen.

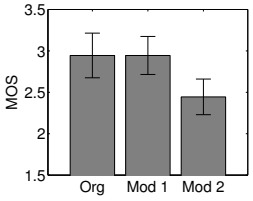
The test has been set up by presenting sequences of instrument sounds to the participants. The sequences for the pitch transpositions consisted of 3 consecutive pitches starting with an original lower pitch and going up to the highest pitch according to the transposition to be achieved. Each sequence has been presented twice, once containing an original sound sample from the database in the middle and once containing an interpolated version using the instrument model substituting the sound in the middle. This has been done for the trumpet as well as for the clarinet once for each transposition distance, resulting in two sequences for each pitch transposition and each instrument. The participants have been asked for the audibility of artifacts and if the sequence sounds convincing in general. Finally, they had to choose from a selection of 5 gradually decreasing values for each sequence ranging from 1 (*perfectly natural*) to 5 (*inacceptable*). The results are denoted *Org* in case of the original sounds and *Mod* for the results using our model.

The sequences to evaluate the intensity changes consisted of 5 consecutive sounds. Starting with an original *pp* sound and going up to an original *ff* sound. For the sequence containing only original sounds an *mf* sound has been put at the 3rd position whereas for the 2nd and 4th position the respective upper one had been put and simply lowered in amplitude to represent a state of the approach. For evaluating the model 2 different sequences have been presented. In the first interpolated sounds between *pp* and *ff* have been presented using mix factors of .25, .5 and .75, whereas for the second sequence an original *mf* has been put in the middle and the sounds between *pp* and *mf* as well as *mf* and *ff* have been substituted their respective interpolations. Again, the participants were asked to judge artifacts and if it sounds convincing to them on the same scale. The two



(a) 12st Interpolation

(b) 24st Interpolation



(c) Intensity Interpolation

Figure 9. Mean opinion scores for the subjective evaluation.

different sequences using our model are denoted *Mod 1* in the former and *Mod 2* in the latter. For the test 9 expert listeners participated and they all were using headphones.

Unfortunately, as can be seen from the unexpected low values for all 3 evaluations, the participants gave bad *Mean Opinion Score* (MOS) even to the original sounds. This is most likely caused by a not suitable test setup which might have mislead the participants in judging aspects in the recordings not part of the investigation. Nevertheless, as can be seen in Figure 9a and 9b our model achieved a MOS comparable to the original sounds. For the intensity interpolation the model using the larger distance interpolation only showed state of art rates whereas for smaller distances, the MOP significantly outperforms the state of the art approach. This indicates the model captured the significant instrument sound characteristics and successfully transformed the interpolated sounds.

9. CONCLUSIONS

In this paper we have shown a substantially new approach of an instrument model applicable for sound transformation and synthesis. We have presented methods to establish an instrument model based on the source-filter approach using tensor product B-splines and a technique to estimate its parameters using a database of instrument sounds. We have further given a deep insight into some specific characteristics of this approach and how regularization can be used to capture the meaningful sound properties of a musical instrument. We have further shown how to apply sound synthesis for interpolation between sounds of the database and we have presented an evaluation of our model demonstrating the accuracy of the model in terms of pitch transposition and the significant improvement for interpolating intensity layers compared to a state of the art approach.

10. ACKNOWLEDGEMENTS

This research has been financed by the french ANR project Sample Orchestrator 2.

11. REFERENCES

[1] V. Välimäki, J. Pakarinen, C. Erkut, and M. Karjalainen, "Discrete-time modelling of musical instruments," *Reports on Progress in Physics*, vol. 69, no. 1, pp. 1 – 78, Jan. 2006.

[2] D. Arfib, F. Keiler, U. Zölzer, and V. Verfaille, *Digital Audio Effects (eds. U. Zölzer)*, 2nd ed. John Wiley & Sons, 2011, ch. 8 - Source-Filter Processing, pp. 279 – 320.

[3] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Music Genre Database and Musical Instrument Sound Database," in *4th International Society for Music Information Retrieval Conference*, October 2003, pp. 229 – 230.

[4] A. Klapuri, "Analysis of musical instrument sounds by source-filter-decay model," in *2007 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, April 2007, pp. I–53 – I–56.

[5] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *10th International Society for Music Information Retrieval Conference*, October 2009, pp. 327 – 332.

[6] J. Burred, A. Röbel, and T. Sikora, "Polyphonic musical instrument recognition based on a dynamic model of the spectral envelope," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Taipei, Taiwan*, April 2009.

[7] J. Burred and A. Röbel, "A segmental spectro-temporal model of musical timbre," in *13th International Conference on Digital Audio Effects*, September 2010.

[8] X. Serra, "A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition," Ph.D. dissertation, Stanford University, 1989. [Online]. Available: files/publications/PhD-Thesis-1989-xserra.pdf

[9] X. Amatriain, J. Bonada, A. Loscos, and X. Serra, *Digital Audio Effects (eds. U. Zölzer)*, 2nd ed. John Wiley & Sons, 2011, ch. 10 - Spectral Processing, pp. 393 – 446.

[10] H. Hahn, A. Röbel, J. J. Burred, and S. Weinzierl, "Source-filter model for quasi-harmonic instruments," in *13th International Conference on Digital Audio Effects*, September 2010.

[11] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *8th International Conference on Digital Audio Effects*, Madrid, Spain, Septembre 2005, pp. 30 – 35.

[12] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *NEURAL NETWORKS*, vol. 6, no. 4, pp. 525 – 533, 1993.