Extended Source-Filter Model for Harmonic Instruments for Expressive Control of Sound Synthesis and Transformation

Henrik Hahn, Axel Röbel

EXTENDED SOURCE-FILTER MODEL FOR HARMONIC INSTRUMENTS FOR EXPRESSIVE CONTROL OF SOUND SYNTHESIS AND TRANSFORMATION

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 revised and improved version of a recently proposed extended source-filter model for sound synthesis, transformation and hybridization of harmonic instruments. This extension focuses mainly on the application for impulsively excited instruments like piano or guitar, but also improves synthesis results for continuously driven instruments including their hybrids. This technique comprises an extensive analysis of an instruments sound database, followed by the estimation of a generalized instrument model reflecting timbre variations according to selected control parameters. Such an instrument model allows for natural sounding transformations and expressive control of instrument sounds regarding its control parameters.

1. INTRODUCTION

Digital music synthesizers based on prerecorded sound samples of an instrument (Sampler) only represent a discretized version of the instruments possible timbre space. Creating such a data base of sound examples always includes making a compromise between memory usage, recording efforts and granularity of timbre space. This usually either leads to audible jumps during playback of notes with similar intensity or to the so called *Machine-Gun-Effect*, if one note is synthesized with equal intensity and pitch several times, always leading to the playback of the same sound sample and therefore making the sound synthesis becoming static.

There are contextual systems using small musical units [1] or phrases [2] that also try to establish a less static sample based sound synthesis. These systems however are not based on single note recordings of musical instruments, but more complex sound data sets and therefore cannot be established on existing libraries like RWC [3], Vienna Symphonic Library or IRCAM Solo Instruments. Current musical samplers therefore rely on sound transformations to somehow interpolate the discretized timbre space using the available sound samples. Such transformations are usually based on the source–filter model of sound production [4] [5] with an assumed white source and a coloring filter. This model is still being used for transformations of sounds [6] [7], even though the assumption of a white source and a single filter does not match physical phenomena. The glottal source in voice signals as well as the modes of excitation of musical instruments (strings, air pipes, etc.) do not exhibit white distributions of harmonics, but highly pronounced spectral envelopes. For voice processing, glottal source estimation became first addressed with the *Liljencrants-Fant model* [8] and non–white source assumptions are now widely being accepted for all kind of voice signal transformations [9] [10] [11] [12]. For instrument sounds, this is much more an emerging topic. Models exploiting the idea of non–white sources and control of gestural parameters like pitch and intensity have recently been addressed for instrumental sound synthesis [13], but also the use of several separate filters with different parametrizations have been proposed for instrument recognition [14] and source separation [15]. An extension by using an additional sinusoid plus noise model [16] [17] for sound morphing has been presented in [18].

We recently proposed a generalized instrument model [19] for driven instruments like bowed strings and blown instruments, which extends the basic source–filter approach by the separate treatment of sinusoidal and noise signal components with individually established filter modules. For the sinusoidal component, a white source with 2 filter functions with gestural control parameters has been proposed. The noise component has been introduced in equivalence to its sinusoidal counterpart, but with a single filter function with equal control parameters. In our approach, model parameters are learned from an instruments sound database using a gradient decent method. In this paper we will present a revised mathematical formalism and a generalized constraint framework for an improved modeling of driven instruments and enhancements dedicated to impulsively excited instruments to ameliorate synthesis results in general. We further introduce a method for gradient normalization to significantly decrease computational efforts for model adaptation and give a description how to use the proposed model for sound synthesis with expressive control parameters.

In section 2 we will give a summary of our proposed instrument model with a detailed discussion of issues arising from the different modes of excitations of musical instruments, followed by the introduction of our new constraint framework in section 3 and the description of the gradient normalization in 4. Section 5 depicts visual results of our instrument modeling, whereas section 6 describes our synthesis model and links to a number of sound examples available online. The conclusion is given within section 7.

2. GENERALIZED INSTRUMENT MODEL

2.1. System Overview

The initial version of our source-filter instrument model was described in [20] and extended in [19] for sound transformation and interpolation. The model is learned from an instruments data base of recorded sounds and represents timbre variations of an instrument according to certain selected control parameters. In [19] we proposed the use of the global note intensity, temporal intensity

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

envelope and pitch to be these parameters, as we assumed them to be the most influential ones for sound variations of a musical instrument. We established an instrument model, reflecting these variations separately for a signals harmonic and noise content. With a model reflecting instrument timbre variations due to certain gestural control parameters, filter envelopes can be generated to transform any selected sound sample according to manipulations of these control parameters. Therefore, transformations include changes of a notes intensity, pitch transpositions and changes to the temporal intensity envelope. And since the instrument model is learned from recorded sounds, these transformations are expected to generate synthesis results, which are indistinguishable from their natural recordings.

Hence, this approach comprises a sound analysis, a model adaptation and a source signal creation stage to generate the desired components for our purpose of an expressive sound synthesis with control of gestural parameters.

2.2. Sound datasets

To establish our synthesis model we use several instruments from the IRCAM Solo Instruments library. Each data set consists of single note recordings with constant pitch and intensity. They are always covering the whole instruments pitch range in chromatic steps whereas 3 intensity levels (*pp*, *mf*, *ff*) have been recorded for the driven instruments and up to 8 different intensity values for the piano. All sounds have been recorded in 24Bit and 44.1kHz.

2.3. Sound Analysis

The target of our sound analysis is to transform the given input signals into a domain, which can be used for modeling as well as for sound synthesis and since our aim is to treat the harmonic and noise components of instrument signals separately, the first step in sound analysis is their segregation. We therefore employ a sinusoidal analysis by means of partial tracking [16]. Following eq. (1) an input signal $x^{(\alpha)}(t)$ can be expressed as a sum of K sinusoids with time-varying amplitude $a_k^{(\alpha)}(t)$ and phase $\phi_k^{(\alpha)}(t)$ and some residual noise $\nu^{(\alpha)}(t)$, with (α) denoting a certain input signal from the data set.

$$x^{(\alpha)}(t) = \sum_{k}^{K} a_k^{(\alpha)}(t) cos(\phi_k^{(\alpha)}(t)) + \nu^{(\alpha)}(t) \qquad (1)$$

Since all samples exhibit a constant pitch, the time-varying amplitude and frequency of the harmonic partials k can be estimated for each time frame n of a signal, using a filterbank with center frequencies according to $f_k = k \cdot f_0$. However, for the piano as well as for all impulsively excited string based instruments, we need to take their inharmonicity into account, since the frequencies of the harmonics are not exact multiplies of its fundamental but at increased positions. We are using a novel algorithm [21] to reliably estimate the positions of the harmonics, which we use for the partial tracking of the piano. These estimated sinusoids with time-varying amplitude and frequency therefore, represents our harmonic signal component. Considering the use of single note recordings, we assume the sounds to be stationary in pitch and harmonic contour and we therefore approximate the frequencies of the harmonics by their assumed optimal position according to the rule in eq. (2) with $\beta = 0$ for driven and an estimated $\beta > 0$ for piano instruments.

$$f_k = k f_0 \sqrt{1 + k^2 \beta} \qquad (2)$$

As our instrument model shall reflect timbre variations of a musical instrument according to certain gestural control parameters, we define these parameters to be the global note intensity I, the pitch P and the time-varying temporal intensity envelope $E(n)$. The global intensity reflects the notes overall loudness and is denoted in musical terms by *pp* (pianissimo) for a very quiet and by *ff* (fortissimo) for a very loud playing style, which usually comes with highly different spectral shapes. The pitch encodes the position of the fundamental frequency and therefore of all higher harmonics making it a decisive parameter as well. All global intensity as well as pitch values will be organized for our model in equidistant steps. This means, the pitch values will be scaled on a logarithmic frequency axis using its MIDI values. Both, the intensity as well as the pitch values are being taken from meta information belonging to each sound data set.

These parameters however do not account for temporal timbre changes due to the attack or release of a signal, but as we target for an instrument model reflecting these temporal changes as well, we assume them to be associated with the temporal intensity envelope. Therefore, we will represent timbre changes due to the attack, sustain and release phase of a signal by means of level changes of the temporal intensity envelope.

With this assumption, timbre variations during the attack or release will be reflected by lower values of the intensity envelope, whereas an approximately steady timbre is assumed during the signals sustain phase showing a more or less constant temporal signal intensity over time with maximum value. The temporal intensity envelope for the sinusoidal signal is processed by the summation of the energy of all detected harmonic sinusoids and represented on a decibel scale normalized to 0dB. As the attack and release phase share values below the maximum, we employ a threshold based temporal segmentation of the intensity envelope to differ between the attack, sustain and release signal components as shown in fig. 1. Since impulsively excited signals do not contain a steady state, the end of the attack n_A and the beginning of the release n_R become equal. The segmentation is being done by creating two sets $s = \{n_1, n_2\}$ of frames associated with the attack to sustain and sustain to release region respectively. These regions are overlapping for driven instrument signals, but distinct for impulsive ones.

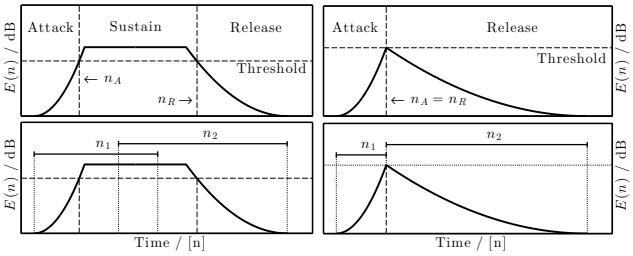

Figure 1: *Temporal segmentation scheme for sustained (left) and impulsive (right) instrument signals using a threshold method. Signal regions Attack, Sustain and Release are indicated (top) as well as temporal segmentation (bottom) using overlapping segments for driven instruments and distinct segments for impulsively excited signals.*

Shown in def. (3), the harmonic signal component $X_h^{(\alpha)}$ is being represented by the partial amplitudes $A_k^{(\alpha)}$ in dB scale, their approximated partial frequencies $f_k^{(\alpha)}$, the triplet of gestural control parameters denoted $\Theta^{(\alpha)}(n)$ and both temporal segments $\{n_1, n_2\}$. Note that the gestural parameters are a function of time since the temporal intensity envelope $E_h^{(\alpha)}(n)$ is time-varying, whereas $f_k^{(\alpha)}$ is not, since we use the ideal frequency values of the partials for modeling.

$$X_h^{(\alpha)} : \begin{cases} A_k^{(\alpha)}(n) \ ; & k = 1 \dots K, n = 1 \dots N \\ \Theta^{(\alpha)}(n) = & \{I, E_h(n), P\}^{(\alpha)} \\ f_k^{(\alpha)} \ ; & k = 1 \dots K \\ n_s = & \{n_s\}_{s \in \{1,2\}} \end{cases} \tag{3}$$

The residual noise signal $x_n^{(\alpha)}(t)$ is generated by the subtraction of the synthesized harmonic signal $x_h^{(\alpha)}(t)$ from the original input signal $x^{(\alpha)}(t)$ using the estimated partials. The noise signal is then analyzed by means of time-varying cepstral envelopes with a small number L of cepstral coefficients l for a smooth modeling of noise envelopes. These time-varying cepstral coefficients are written as $C_l^{(\alpha)}(n)$ and again we establish a triplet of gestural control parameters $\Theta^{(\alpha)}(n)$ with its own temporal intensity envelope $E_r^{(\alpha)}(n)$ and according temporal segmentation $\{n_1, n_2\}$ to obtain the noise signal representation shown in def. (4) similar to its harmonic counterpart. It technically only differs in the missing partial frequencies.

$$X_r^{(\alpha)} : \begin{cases} C_l^{(\alpha)}(n) & \\ \Theta^{(\alpha)}(n) = & \{I, E_r(n), P\}^{(\alpha)} \\ n_s = & \{n_s\}_{s \in \{1,2\}} \end{cases} \tag{4}$$

These representations of the harmonic and a noise signal component of an input signal $x^{(\alpha)}(t)$ are being used to create the desired models and since they share a similar design, both models can be established in a quite analogous manner.

2.4. Harmonic Model

Since we represent the harmonic signal component by means of an additive model, we are able to formulate a harmonic instrument model with respect to the partial index as well as the partials frequency to represent timbre variations according to the gestural control parameters. This allows to incorporate the idea, that the harmonics of an instrument sound exhibit certain features, which are related to the signals fundamental frequency and other features, who are not. In other words, there exist features which are being best described by the partial index and features being best described by the partials frequency. The former may refer to known effects like the missing even partials in clarinet sounds and therefore mainly represents the characteristics of the modes of excitation of an instrument signal, whereas the latter features characterize an instruments resonance behavior at certain fixed frequencies and may accordingly been interpreted as the instruments corpus. Both interpretations are just roughly consistent with real physical phenomena, but as we try to establish a generalized instrument model, certain specific features and characteristics need to be neglected for universality.

Following the distinction above, we introduce 2 separate filter functions. The first filter is being established to reflect instrumental sound features by partial index and is created by separate amplitude functions for each possible partial k of that instrument and both temporal segments s. We further assume the amplitude of a partial k being mainly characterized by our gestural control parameters: global note intensity I, temporal intensity envelope $E(n)$ and the signals pitch P. This allows to model the characteristics of each partial independently from its frequency and neighbors and also independently for its attack and release phase. We denote it as the source filter $S_{k,s}(\Theta)$. The second filter shall reflect the instruments features by frequency and will consequently be a function of the partials frequencies. We assume it to be independent of the gestural control parameters intensity, temporal intensity envelope and pitch, hence denoting it the resonance filter $F(f_k)$. This leads to eq. (5), where the log-amplitude $\hat{A}$ of a partial k within a certain temporal signal segment s depends on the parameters $\{\Theta, f_k\}$ and is processed by the summation of the two filters.

$$\hat{A}_{k,s}(\Theta, f_k) = S_{k,s}(\Theta) + F(f_k) \tag{5}$$

In [19] we have introduced the use of B-splines [22] to model both filters as continuous trajectories according to their parameters. While for the second filter a univariate B-spline B^v with its weights δ^v is used, we proposed the use of a tensor-product B-spline B^u [23] to model the partials source function with respect to the gestural control parameters using the weights $\gamma_{k,s}^u$. We also added an additional offset parameter Φ_h set to a desired amplitude minimum to obtain eq. (6). In fig. 2 three exemplarily chosen B-spline models are shown which are being used to generate the tensor-product B-splines B^u used to model the partial functions.

$$\hat{A}_{k,s}(\Theta, f_k) = \sum_u^U B^u(\Theta)\gamma_{k,s}^u + \sum_v^V B^v(f_k)\delta^v + \Phi_h \tag{6}$$

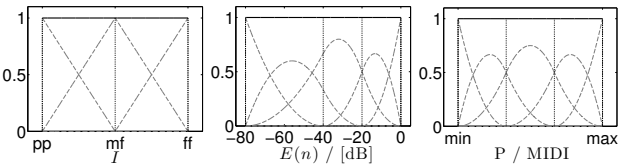

Figure 2: *Three B-Splines models used to create the tensor-product B-spline model for $B^u(\Theta)$*

Cutouts of the tensor-product B-splines are shown in fig. 3 to illustrate, how a hyperplane can be established in higher dimensional space using a tensor-product of several univariate B-splines.

For the resonance filter we employ an univariate B-spline model as shown in fig. 4 to continuously interconnect the positions of the frequencies of all partials present in an instruments data set. This allows for a smooth modeling of the filter function, while only observing a sampled version of its characteristic.

2.5. Noise Model

The noise model has been proposed in a similar manner as its harmonic counterpart, but with only one filter. The filter is established by separate functions for each cepstral coefficient l and temporal

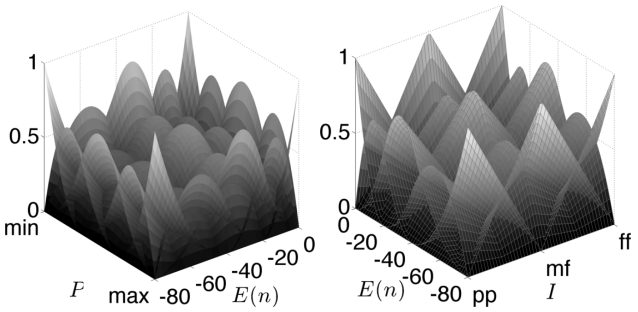

Figure 3: *Cutouts along two dimensions of the 3-dimensional tensor-product B-spline* $B^u(\Theta)$

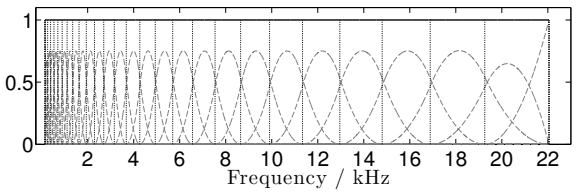

Figure 4: *A B-spline model used for* $B^v(f_k)$

segment s and each with respect to the control parameters Θ leading to the model in eq. 7. Again, we utilize a tensor-product B-spline to represent the behavior of each cepstral coefficient according to the gestural control parameters Θ together with an additional offset parameter Φ_r set some desired noise envelope minimum. Note that this offset parameter has to be in the cepstral domain.

$$\begin{aligned}\hat{C}_{l,s}(\Theta) = &\quad S_{l,s}(\Theta) \\ &= \sum_{w}^{W} B^w(\Theta)\epsilon_{k,s}^w + \Phi_r \end{aligned} \tag{7}$$

2.6. Parameter Estimation

To adapt the harmonic as well as the noise model to a selected data set of recordings of an instrument, the parameters to estimate are the B-spline weight parameters $\gamma_{k,s}^u$, δ^v and $\epsilon_{k,s}^w$. Even though, both models are linear, a closed-form solution is impractical to solve, as a single model can have up to 100k parameters and even small data sets exhibit several millions of data points, since each single partial in every frame represents one data point by its own, this leads to unmanageable matrix sizes even with single precision. Therefore, we utilize the scaled conjugate gradient method [24] (SCG) and define an error function in a least-means-square sense (8) and as all conjugate gradient techniques are offline methods, the error $R_h^{(\alpha)}$ as well as its gradient have to be accumulated for all data samples (α). The gradient is easy to determine and also the functions for the noise model are easy to derivate from the harmonic case.

$$R_h^{(\alpha)} = \frac{1}{2}\sum_{s}\sum_{n_s}\sum_{k=1}^{K}\left(A_k^{(\alpha)}(n) - \hat{A}_{k,s}(\Theta^{(\alpha)}(n), f_k^{(\alpha)})\right)^2 \tag{8}$$

2.7. Discussion

The most obvious issue of the harmonic model, is the summation of two filters which exhibits an unlimited amount of optimal solutions as any constant can always be added to one filter and subtracted from the other. But there are also ambiguous solutions due to the pitch dependency of the source filter function which can also be seen as a log-frequency dependency and therefore might lead to ambiguous results with the frequency dependent resonance filter. More ambiguous solutions may become possible, if more or different gestural control parameters are introduced.

Another major issue which needs to be addressed is the unequal distribution of the data with respect to the control parameters, meaning the existence of large areas in the model without any or just few data. This has to be assumed for all instruments, as it can be caused by partials only being present in *fortissimo*, but never within *pianissimo* signals, as well as by partials only present in lower pitches, but due to sample rate limits, never in higher pitches. Another reason for such data gaps can be caused by impulsively excited instruments, where, due to the shorter decay times of higher partials, no partial amplitudes can be found for regions of lower temporal intensity. Such data gaps can cause almost random behaviors within these regions during the adaptation process. This leads to highly malformed shapes within the model and therefore may result in an erroneous synthesis while transforming partial amplitudes.

Consequently, plausible shapes are needed to enable reasonable transformations.

3. CONSTRAINTS

To solve the issues which arise from the ambiguities and data gaps, we propose the use of constraints as additions to the model error function 8. These constraints are meant to measure deviations from desired characteristics, such that by means of minimization of the resulting objective function, not only model error minimization is pursued, but at the same time, the deviations from a desired behavior is minimized as well.

All these constraints can be formulated independently of the B-spline representation and an advantage of using B-splines is, that derivatives of arbitrary order can be taken easily. All constraints have to be weighted by means of a weighting factor, balancing the importance of the constraint. The choice of the weighting factor is a problem, as it needs to balance the impact of the constraint compared to the model error and the other constraints.

3.1. Constraint I

The first type of constraint samples a filter function with an arbitrary but high enough rate and squares the sum of all filter values, such that it measures the offset of the function. We apply this constraint only to the resonance filter as shown in eq. 9 with f being some frequency sampling vector. This constraint increases the value of the objective function for any offset of the resonance filter from zero and therefore its minimization solves the ambiguity problem of the summation of the two filters by fixing the second filter around 0dB.

$$C_{\mathbf{I}} = \frac{1}{2}\lambda^j\left(\sum_f F(f)\right)^2 \tag{9}$$

The constraints weighting parameter λ^j has to be adjusted according to the amount of data $f_k^{(\alpha)}$ and sampling points f, by means of obtaining a tradeoff for the constraint to be effective and still minimizing the error function properly. In eq. 10 we propose to use the ratio of the $l1$-norm of all squared B-spline values of all data points $f_k^(\alpha))$ and the $l1$-norm of the sum of the squared B-spline values of the sampling series f.

$$\lambda^j = \lambda_0^j \frac{\left|\left|\sum_{(\alpha)} \sum_s \sum_{n_s} \sum_k^K \left(B^v(f_k^{(\alpha)})\right)^2\right|\right|_1}{\left|\left|\sum_f \left(B^v(f)\right)^2\right|\right|_1} \quad (10)$$

This ratio is then adjusted with an initial weighting parameter λ_0^j which is independent of the actual amount of data and sampling points. It scales between 0, reflecting no constraint, and 1, making the constraint as strong as all data points together. We are using $\lambda = 0.001$ which works well for all our instrument sets.

3.2. Constraint II

The second class of constraints is designed to solve the data gap issues mentioned in 2.7, by means of extrapolation from regions containing meaningful data into sparsely filled or even empty areas. It further solves the ambiguity problem of the frequency dependency of both filter functions . We establish this constraint for the source filter function only.

Again, the constraint samples the filter function at an arbitrary grid Θ, but sums the squares of the z-th order partial derivatives of the filter function with respect to one of its dimensions as denoted in eq. (11). With $z = 1$ or $z = 2$ the constraints value therefore increases with either increasing slope or curvature respectively, but both times along some selected dimension $i \in \{P, I, E(n)\}$. Hence, minimizing the slope of the surface along one of its dimensions means flattening its shape and extrapolating constantly, whereas minimizing its curvature can be used to linearly extrapolate along one dimension. Moreover, specifically targeted for impulsively excited signals, we introduce a function $\eta^{i,z}(\Theta_i)$ to locally emphasize a constraints weighting parameter $\lambda_{k,s}^{i,z}$, which allows to not only adjust a weight constantly for a whole constraint, but according to a certain control parameter i. This allows for adjustments of the impact of a certain constraint, depending on the value of a gestural control parameter, since these instruments show rapid spectral and intensity changes. In other words, we like to have stronger constraints for lower values of the local intensity envelope than for higher values to smoothly fade the partial amplitudes of higher index. In our case, $\eta^{i,z}(\Theta_i)$ is linear along i and constant along all others and only used for impulsively excited instrument signals.

We established this constraint generically for all orders of partial derivatives and dimensions of the filter function.

$$C_{\mathbf{II}}^{i,z} = \frac{1}{2} \sum_s \sum_{k=1}^K \lambda_{k,s}^{i,z} \sum_{\Theta} \eta^{i,z}(\Theta_i) \left(\frac{\partial^z}{\partial \Theta_i^z} S_{k,s}(\Theta)\right)^2 \quad (11)$$

Again, the weighting parameters of the constraints need to be processed in a manner similar to the first constraint. But as we apply this constraint to the source filter function only, the according weight has to be summed separately for each temporal segment s and partial index k and also independently for all selected values of dimension i and derivative order z, establishing a weight $\lambda_{k,s}^{i,z}$ with respect to these parameters.

Eq. (12) shows how an initial weighting parameter $\lambda_0^{i,z}$ is being scaled with the ratio of the $l1$-norm of the data dependent sum and and the $l1$-norm of the sum over the sampled filter function, hence making the initial constraints weight $\lambda_0^{i,z}$ independent of the data and dependent only on the selected dimension and derivative order.

$$\lambda_{k,s}^{i,z} = \lambda_0^{i,z} \frac{\left|\left|\sum_{(\alpha)} \sum_{n_s} \left(B^u(\Theta^{(\alpha)}(n_s))\right)^2\right|\right|_1}{\left|\left|\sum_{\Theta} \left(\frac{\partial^z}{\partial \Theta_i^z} B^u(\Theta)\right)^2\right|\right|_1} \quad (12)$$

The ambiguity due to the log-frequency dependency of the source filter and the frequency dependency of the resonance filter can be dissolved by the assumption that a source varies slower with the log-frequency than a resonance filter varies with frequency. This follows the idea of an excitation signal being similar for different pitches, but a resonance body may exhibit nearby resonances and anti-resonances. Such a constraint can be created by using $i = P$ and $z = 1$, forcing the surface of the source to be rather constant along the pitch dimension.

Extrapolation as described above is being carried out for the global I as well as the local intensity E, since the pitch is already constrained using $z = 1$ for constant behavior. For the global intensity we choose $z = 2$ to obtain some fade-out, rather than a constant behavior while extrapolating *ff* to *pp*. The local intensity envelope is always being constrained using a trade-off between $z = 1$ and $z = 2$ for the effect of some smooth fade-out of the partial amplitude values. Additionally, for impulsive signals, these constraints are given boosted weights for lower local intensity values using $\eta^{i,z}(\Theta_i)$ for improved modeling of its rapid changes and huge gaps for lower values of E.

The initialization of the lambda parameters is crucial for all training results and since we do not yet have an automatic selection of suitable parameter configurations, several manual step by step adjustments are required to achieve good training results which can be used for sound synthesis.

Finally, the resulting objective function to optimize the harmonic model according to the data samples (α) and constraints $C_{\mathbf{I}}$, $C_{\mathbf{II}}^{i,z}$ is denoted as eq. 13.

$$O_h = \sum_{(\alpha)} R_h^{(\alpha)} + \sum_{i,z} C_{\mathbf{II}}^{i,z} + C_{\mathbf{I}} \quad (13)$$

4. GRADIENT NORMALIZATION

As we already stated, the harmonic as well as the noise model are both linear models and therefore their objective functions exhibit a global minimum and both possess the shape of an ellipsoid. SCG is a conjugate gradient optimization method designed for complex and high dimensional non-linear problems. In the present case the problem is linear, but very high-dimensional, such that an analytic solution is not feasible. The convergence of conjugate gradient methods is related to the condition number of the Hessian that characterizes the problem. The condition number of the present problem is unfortunately rather high, causing SCG to take several thousand iterations until convergence for a linear problem with up to 100k parameters. The high condition number is related to the uneven distribution of the data points, according to our model space of control parameters and partial frequencies, which had already been discussed in section 2.7.

A well known solution for such cases is the preconditioning of the problem, such that the condition number is reduced [25]. For the present problem, we propose to follow a rather simple approach of preconditioning, that consists of scaling all equations, such that the diagonal elements of the Hessian matrix are all identical and equal to 1. Our practical experience with this preconditioning has shown a significant increase in convergence speed, reducing the required number of iterations from a few 1000 to 25 or 50 for the noise and harmonic parameter optimization respectively. To normalize all second derivatives on the diagonal of the Hessian, we introduce a substitution of all B-spline parameters of the harmonic model as shown in eq. (14) and (15) into a new B-spline parameter $\bar{\gamma}^u_{k,s}$ and an additional data dependent weighting parameter $c^u_{k,s}$. The equations for the noise case can easily be derived from the harmonic case.

$$\gamma^u_{k,s} = \bar{\gamma}^u_{k,s} \cdot c^u_{k,s} \quad (14)$$

$$\delta^v = \bar{\delta}^v \cdot c^v \quad (15)$$

Now, the new B-spline parameters $\bar{\gamma}^u_{k,s}$ need to be estimated from the input data, with respect to all constraints and therefore, all gradients of the objective function need to be reformulated using the new weighting parameters, which is left to the reader. The introduced parameters $c^u_{k,s}$ and c^v need to be calculated from eq. 16 and eq. 17 by solving all equations for the new c parameters.

$$\frac{\partial^2 O_h}{\partial \bar{\gamma}^{u\,2}_{k,s}} = \sum_{(\alpha)} \frac{\partial^2 R_h^{(\alpha)}}{\partial \bar{\gamma}^{u\,2}_{k,s}} + \sum_{i,z} \frac{\partial^2 C^{i,z}_{\mathbf{II}}}{\partial \bar{\gamma}^{u\,2}_{k,s}} = 1 \quad (16)$$

$$\frac{\partial^2 O_h}{\partial \bar{\delta}^{v\,2}} = \sum_{(\alpha)} \frac{\partial^2 R_h^{(\alpha)}}{\partial \bar{\delta}^{v\,2}} + \frac{\partial^2 C_{\mathbf{I}}}{\partial \bar{\delta}^{v\,2}} = 1 \quad (17)$$

Solving the equations 16 and 16 for $c^u_{k,s}$ and c^v gives the requested normalization of the diameters of the error ellipsoid.

5. MODELING RESULTS

To demonstrate the universality of our approach for sound transformations utilizing a generalized instrument model, we selected several different musical instruments, to cover various specific models of excitation and gestural control. These instruments are trumpet, clarinet, violin and a grand piano. In the figures, we present cutouts of the estimated hyperplanes of the source filter $S_{k,s}(\Theta)$ for a certain partial k and temporal segment s with respect to two out of the three control parameters and estimated univariate resonance filters $F(f)$.

Additionally, all figures depict data points from the input signals in accordance to their mapping onto the filters, to show the fitting of the B-spline models to their respective data. This means, the displayed amplitudes of the partial data are processed by subtracting the not-displayed filter from the input data like shown in eq. (18) for the source and in eq. (19) for the resonance filter figures.

$$\tilde{A}^{(\alpha)}_{k,s}(n_s) = A^{(\alpha)}_k(n_s) - F(f_k) \quad (18)$$

$$\tilde{A}^{(\alpha)}_{k,s}(n_s) = A^{(\alpha)}_k(n_s) - S_{k,s}(\Theta^{(\alpha)}) \quad (19)$$

Note, that for the source filter only partial amplitudes of a fixed partial index k and temporal segment s are shown, since the source hyperplane is specific for each k and s but with respect to the control parameters. On the contrary, within the figures of the resonance filters, the amplitudes of all partials k from both temporal segments are displayed according to their frequency. It can be

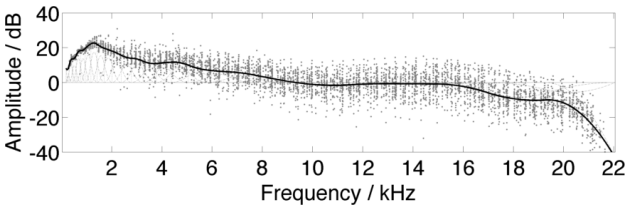

Figure 5: *Estimated filter $F(f)$ of the trumpet (solid) and according data points $\tilde{A}^{(\alpha)}_{k,s}(n_s)$ (grey)*

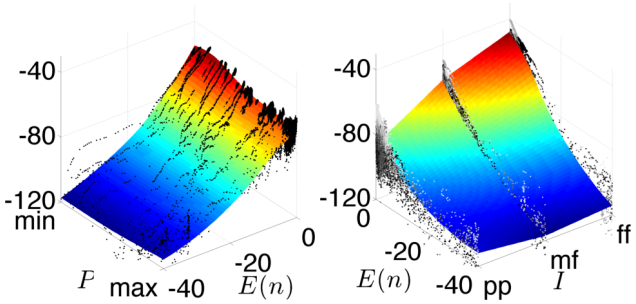

Figure 6: *3-dimensional cutouts of the four-dimensional source filter function for the 10-th partial and 2nd temporal segment of the trumpet. The plane represents the source model $S_{k,s}(\Theta)$, with the left showing the plane for* mf *and the right denoting the plane for medium pitches. Data points fade from black to white indicating their decreasing influence to the shown plane regarding the B-splines of the left-out dimension.*

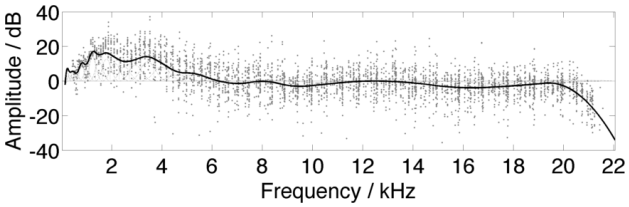

Figure 7: *Estimated filter $F(f)$ of the clarinet (solid) and according data points $\tilde{A}^{(\alpha)}_{k,s}(n_s)$ (grey)*

observed from the figures, that all models exhibit a pretty strong variance of the data regarding the filter functions. This is related to timbre variations of each instrument, which are not covered by our selected control parameters. As these variances are not modeled, they will therefore retain in the signals for all transformations.

One major limitation of our model is due to the still missing automatic adjustment of the configuration of the B-spline models as well as the initialization of the constraint weighting parameters. We therefore need to optimize these parameters manually by judging the training results visually.

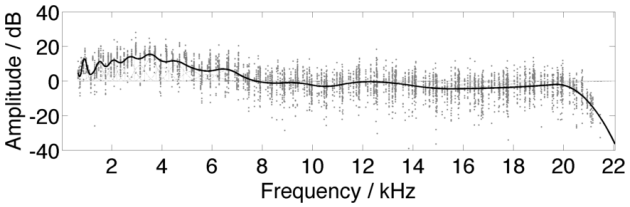

Figure 8: *Estimated filter $F(f)$ of the violin (solid) and according data points $\tilde{A}_{k,s}^{(\alpha)}(n_s)$ (grey)*

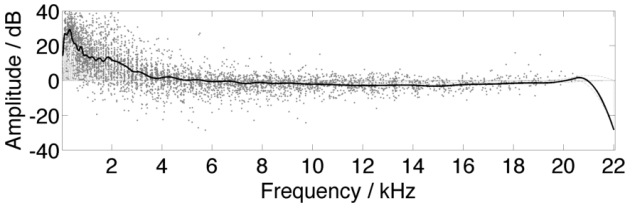

Figure 9: *Estimated filter $F(f)$ of the piano (solid) and according data points $\tilde{A}_{k,s}^{(\alpha)}(n_s)$ (grey)*

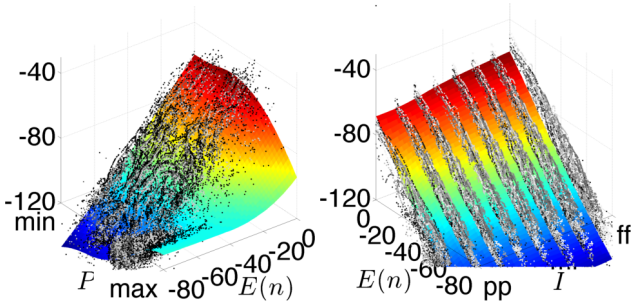

Figure 10: *3-dimensional cutouts of the four-dimensional source filter function for the 5-th partial and 2nd temporal segment of the piano. The plane represents the source model $S_{k,s}(\Theta)$, with the left showing the plane for* ff *and the right denoting the plane for medium lowest pitches. Data points fade from black to white indicating their decreasing influence to the shown plane regarding the B-splines of the left-out dimension.*

6. SYNTHESIS

6.1. Filter Envelope Processing

Since we use spectral domain filtering for our sound synthesis, the estimated partial amplitudes $\hat{A}_{k,s}$ as well as the cepstral coefficients $\hat{C}_{l,s}$ need to be transformed into spectral envelopes with equidistant points ranging from 0Hz to half the sampling rate. In both cases, we employ the cepstral smoothing method, as the estimated noise coefficients already represent a smoothed envelope and we therefore just need to process a single DFT of the cepstral coefficients to obtain the desired filter envelope. For the harmonic envelope, we need to smoothly interpolate between the partial amplitudes, but as they are already in log domain, a small IDFT with a cepstral domain filtering, followed by another small DFT is sufficient to generate a smoothed envelope passing through the partial amplitudes.

6.2. Source Signal Creation

Before we are able to utilize our models for sound transformations, we filter all segregated harmonic $x_h(t)$ and noise input sounds $x_n(t)$ of the instruments database with the inverse of the estimated envelopes, using their associated gestural control parameters Θ. This removes the timbre, estimated by the models, from each signal and creates a harmonic signal $\bar{x}_h(t)$ with almost white distributed partial amplitudes and an almost white noise signal $\bar{x}_n(t)$. This procedure will remove all sound features from the input signals, which are covered by the instrument models, but leaves everything back in the created signals, which is not modeled.

These source signals can then be transformed with filter envelopes, generated using the model with manipulated gestural control parameters, to achieve signals with altered gestures. They will exhibit the estimated target timbre, but with preserved variations, which are not covered by the models. This makes the synthesis sound natural and not static.

This concludes our proposed model, as we we have now generated all the needed components for our proposed sound synthesis. These components are namely the harmonic and noise source signals as well as the two harmonic filter functions and finally the noise filter function.

6.3. Sound Transformation

The transformation of sounds using our proposed extended source-filter model has first to accomplish the resynthesis of the original sounds without the introduction of artifacts. We denote it the neutral synthesis and it can be achieved by applying filter envelopes to source signals, generated by using the associated gestural control parameters without any manipulation. This has to deliver sounds which are indistinguishable from the originals.

Sound synthesis exhibiting transformations requires manipulation of the gestural control parameters. For example, pitch or global note intensity changes. Transformations due to manipulations of the note intensity effects in altered spectral envelopes according to the envelopes learned from the according sound samples. Changes of the pitch, however, requires an additional transposition step of the harmonic source signal, as the envelopes only serve for the filtering of the spectral shape and hence, the desired pitch shift has to be done by some additional algorithm.

Additionally to sound transformations using the instrument model belonging to the source signals, all components can be exchanged with components from other instruments to create new hybrid instruments with a huge variety of combinations of components.

Though, the proposed instrument model still has some limitations. First, as we do not add artificial partials to the source signals while transforming, sound synthesis is rather limited for transformations of sounds from lower to higher global note intensities and pitch transpositions from higher to lower notes, as such transformations would require the addition of partials. Another limitation results from the still missing automatic adjustment of the temporal intensity envelope, which leads to less natural synthesis results, especially while pitch shifting impulsive signals for more than an octave or changing their note intensity significantly.

6.4. Synthesis Results

Sound examples demonstrating the capabilities of our proposed instrument model are being made available through our demo web page `http://anasynth.ircam.fr/home/media/sor2-instrument-model-demo`.

7. CONCLUSION

We have presented a deeply revised version of our recently proposed instrument model for sound transformation and hybridization using a sample-based synthesis. This included an improved mathematical formalism and the introduction of generalized classes of constraints for enhanced modeling of an instruments timbre variations, due to selected gestural control parameters. We further revised the adaptation algorithm with a preconditioning technique to significantly reduce the amount of iterations, needed for the gradient decent method. Finally, a synthesis scheme has been described to demonstrate the application of our proposed instrument model for sample-based sound synthesis with independent control of the gestural parameters, including natural sounding transformations as well as the creation of new hybrid instruments.

8. ACKNOWLEDGMENTS

Many thanks to the anonymous reviewers for their precious remarks and suggestions.

9. REFERENCES

[1] Diemo Schwarz, "A system for data-driven concatenative sound synthesis," in *Digital Audio Effects (DAFx)*, Verona, Italy, December 2000, pp. 97–102.

[2] E. Lindemann, "Music Synthesis with Reconstructive Phrase Modeling," *Signal Processing Magazine, IEEE*, vol. 24, no. 2, pp. 80 – 91, 2007.

[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] Homer Dudley, "Remaking Speech," *J. Acoust. Soc. Am.*, vol. 11, no. 2, pp. 169 – 177, 1939.

[5] Wayne Slawson, *Sound Color*, University of California Press, Berkerley, 1985.

[6] 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.

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

[8] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[9] D. G. Childers, "Glottal source modeling for voice conversion," *Speech Communication*, vol. 16, no. 2, pp. 127–138, 1995.

[10] Arantza del Pozo, *Voice Source and Duration Modelling for Voice Conversion and Speech Repair*, Ph.D. thesis, Cambridge University, Engineering Department, April 2008.

[11] Javier Perez Mayos, *Voice Source Characterization for Prosodic and Spectral Manipulation*, Ph.D. thesis, Universitat Politecnica de Catalunya, July 2012.

[12] Gilles Degottex, Pierre Lanchantin, Axel Roebel, and Xavier Rodet, "Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis," *Speech Communication*, vol. 55, no. 2, pp. 278 – 294, 2012.

[13] Sean O'Leary, *Physically Informed Spectral Modelling of Musical Instrument Tones*, Ph.D. thesis, The University of Limerick, 2009.

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

[15] 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.

[16] X. Serra, *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*, Ph.D. thesis, Stanford University, 1989.

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

[18] Marcelo Caetano, *Morphing Isolated Quasi-Harmonic Acoustic Musical Instrument Sounds Guided by Perceptually Motivated Features*, Ph.D. thesis, Universite Pierre et Marie Curie, UPMC, Universite Paris VI, 2011.

[19] H. Hahn and A. Röbel, "Extended source-filter model of quasi-harmonic instruments for sound synthesis, transformation and interpolation," in *Sound and Music Computing Conference (SMC) 2012*, Copenhagen, Denmark, July 2012.

[20] 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.

[21] H. Hahn and A. Röbel, "Joint f0 and inharmonicity estimation using second order optimization," in *Sound and Music Computing Conference (SMC) 2013*, August 2013.

[22] C. de Boor, *A Practical Guide to Splines*, Springer, 1978.

[23] K. Höllig, *Finite Element Methods with B-Splines*, Society for Industrial and Applied Mathematics, SIAM, 2003.

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

[25] Jonathan R. Shewchuk, "An introduction to the conjugate gradient method without the agonizing pain," Tech. Rep., School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1994, Available at `http://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf`.