A SOURCE-FILTER MODEL FOR MUSICAL INSTRUMENT SOUND TRANSFORMATION

Marcelo Caetano, Xavier Rodet

Analysis/Synthesis Team, IRCAM
(caetano,rodet)@ircam.fr

ABSTRACT

The model used to represent musical instrument sounds plays a crucial role in the quality of sound transformations. Ideally, the representation should be compact and accurate, while its parameters should give flexibility to independently manipulate perceptually related features of the sounds. This work describes a source-filter model for musical instrument sounds based on the sinusoidal plus residual decomposition. The sinusoidal component is modeled as sinusoidal partial tracks (source) and a time-varying spectral envelope (filter), and the residual is represented as white noise (source) shaped by a time-varying spectral envelope (filter). This article presents estimation and representation techniques that give totally independent and intuitive control of the spectral envelope model and the frequencies of the partials to perform perceptually related sound transformations. The result of a listening test confirmed that, in general, the sounds resynthesized from the source-filter model are perceptually similar to the original recordings.

Index Terms— Source-filter model, sinusoidal models, musical instrument, timbre, sound transformations

1. INTRODUCTION

The source-filter (SF) model was originally proposed to explain speech production [1]. According to this model, speech is viewed as a glottal excitation signal (source) driving a time-varying linear filter that models the resonant characteristics of the vocal tract. The most well known SF system is based on linear prediction (LP) of speech [2], where the autoregressive filter is excited by either quasi-periodic pulses (during voiced speech), or noise (during unvoiced speech). The SF model has been applied to musical instrument sounds [3, 4], usually for instrument recognition [5] , transcription [6] , and source-separation algorithms [7]. In the context of sound transformations, SF models are very attractive because they represent source and filter independently, and therefore allow their independent manipulation [8]. Possible transformations using the SF model are cross-synthesis [4], formant-preserving pitch shifting [9], and sound morphing [10].

There are many possible representations of source and filter, from splines [3] to statistical models [6]. A popular SF model for musical instrument sounds represents the filter as a time-varying spectral envelope, while the source is obtained by inverse filtering [2, 8]. There are different possible spectral envelope estimation techniques, such as the channel vocoder [4], linear prediction [2], and cepstral smoothing [9]. The quality of the results depends directly on the accuracy of the representation. Ideally, the spectral envelope should be a smooth curve that approximately matches the peaks of the spectrum [5]. The estimation of the spectral envelope is intimately linked to the SF model [4, 9] because it corresponds to the identification of the parameters of the filter via deconvolution. The main goal of this deconvolution between source and filter by means of spectral envelope estimation is to eliminate the harmonic structure of the spectrum, which is associated with the source. However, a more compact and flexible representation of the excitation signal has been proposed as sinusoidal models for both speech [11] and musical instrument sounds [4].

This work presents an accurate source-filter (SF) model for musical instrument sounds that gives independent and intuitive control of the spectral envelope and frequency of the partials to perform sound transformations. The sounds are decomposed into a sinusoidal and a residual parts, which are modeled independently with the SF model. The sinusoidal component comprises a time-varying spectral envelope model (filter) and the frequencies of the partials (source), while the residual component is modeled as white noise (source) shaped by a time-varying spectral envelope (filter). We propose to use true envelope (TE) [9] to estimate the spectral envelope of the sinusoidal component, and use linear prediction [2] to estimate the spectral envelope of the residual. Advantages of the SF model over traditional sinusoidal models are accurate representation of source and filter for both the sinusoidal and residual components (which are also independent), independent manipulation of dynamic and spectral properties of source and filter, and a compact representation with perceptually related parameters. The SF representation was validated with a listening test. Participants were presented the original and SF representation of sounds and asked to assess their perceptual similarity.

The next section presents the SF model developed in this work, followed by the estimation of source and filter and resynthesis. Section 4 presents the results of the listening test. Finally we present the conclusions and future perspectives.

2. THE SOURCE-FILTER MODEL FOR MUSICAL INSTRUMENT SOUNDS

An important assumption that is often made in the use of the SF model for speech is the independence of source and filter. The filter imprints its resonant characteristics on the spectrum of the source, but the frequencies of the partials are not affected by the interaction [2]. Independence between source and filter partially explains why the SF model is more rarely applied to musical instruments, whose source and filter are strongly coupled. The coupling between source and filter for musical instruments can be understood as the filter driving the source or, in more musical terms, tuning its pitch. In other words, the filter imposes not only its resonant characteristics to the source, but also the frequencies at which the source will resonate. Nevertheless, there are some special conditions under which the SF model can be applied to model the production of musical instrument

This work was supported by the Brazilian Governmental Research Agency CAPES (process 4082-05-2).

sounds. Slawson [12] proposes to look for subsystems of a musical instrument that are weakly coupled to the rest of the instrument and that may be largely responsible for the sound color of the instrument. Under these conditions, the filter can be associated with the resonant cavity of the instrument, and the source with the excitation method. In this work we neglect the strongly coupled component of musical instruments. So the source is supposed to be a quasi-periodic (or pitched) signal $x(t)$ that is fed into the weakly coupled component W_c of the system, which represents the resonator.

2.1. Signal Processing Modeling of Source and Filter

The musical instrument sound $y(t)$ is separated into a sinusoidal component $y_s(t)$ plus a residual component $y_r(t)$ where $y_r(t)$ is obtained by subtraction of the purely sinusoidal component $y_s(t)$ from the original sound $y(t)$. Both the sinusoidal and residual components are modeled as source and filter. The filter of both is modeled via spectral envelope estimation, while the sources are modeled separately. A flexible representation of the quasi-periodic excitation signal for speech [11] and musical instrument sounds [13, 4] is a sum of sinusoids plus a residual component as follows

$$x(t) = \sum_{k=0}^{K(t)} a_k(t) \exp\left[j\phi_k(t)\right] + x_r(t) = x_s(t) + x_r(t) \quad (1)$$

where $a_k(t)$ and $\phi_k(t)$ are the instantaneous excitation amplitude and phase of the k^{th} sinusoid, respectively, $K(t)$ is the number of sinusoids, which may vary in time, and $x_r(t)$ is the residual source. For most musical instrument sounds, a model where the sinusoids are harmonically related is a good approximation, giving

$$\frac{d}{dt}\phi_k(t) = 2\pi k t f_0(t) \quad (2)$$

where $f_0(t)$ is the instantaneous fundamental frequency. For musical instrument sounds, a further simplification of the excitation signal is convenient, assuming that its amplitude $a_k(t)$ is constant over time (and equal to unity, i.e., $a_k(t) = 1$). Based on these simplifications, the time-varying filter that models the resonant characteristics of the body of the musical instrument also approximates the effects of the shape of the excitation and of the transmission characteristics of the resonator. The time-varying transfer function of the filter can be written as

$$H(f,t) = |H(f,t)| \exp\left[j\psi(f,t)\right] \quad (3)$$

where $|H(f,t)|$ and $\psi(f,t)$ are respectively the amplitude and phase of the system. The processing of musical instrument sounds is usually done on a frame-by-frame basis, where each frame typically containing three periods of the waveform can be considered a stationary process [13]. In this case, inside a frame, the filter $H(f,t)$ is considered linear shift-invariant (LSI). Than the output of the system can be viewed as the convolution of the impulse response of the LSI filter and of the excitation signal as $y(t) = x(t) * h(t) = [x_s(t) + x_r(t)] * h(t) = y_s(t) + y_r(t)$. Recognizing then that the sinusoidal component of the excitation signal $x_s(t)$ is just the sum of $K(t)$ eigenfunctions of the filter $H(f)$, the following model is obtained

$$y_s(t) = \sum_{k=0}^{K(t)} |H[f_k(t)]| \exp\left[j\left(\phi_k(t) + \psi(f_k(t))\right)\right] \quad (4)$$

where $f_k(t) \approx k f_0(t)$ are the eigenfrequencies of the filter $H(f,t)$. The amplitude of the k^{th} harmonic is the system amplitude $|H(f_k(t))|$, which is also its eigenvalue. The phase of the k^{th} harmonic is the sum of the excitation phase $\phi_k(t)$ and the system phase $\psi[f_k(t)]$ and is often referred to as the instantaneous phase of the k^{th} harmonic. In our model, the amplitudes of the partials $|H(f_k(t))|$ are given by the spectral envelope curve, as shown in part b) of figure 1. Figure 1 compares the spectral representation of partials for the traditional sinusoidal modeling approach in part a), and for the SF model in part b). In sinusoidal modeling, each partial is assigned an amplitude and frequency values, while the SF model represents the amplitudes and frequencies of the partials intrinsically independently.

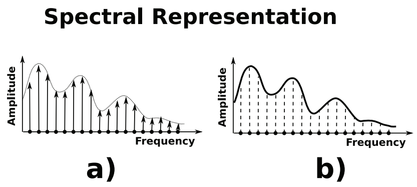


Fig. 1. Spectral representation of partials. The figure shows the traditional sinusoidal representation with the frequency values and amplitudes tied to each other in part a). Part b) depicts our representation, where the amplitudes of the partials are represented independently with a spectral envelope model.

3. ESTIMATION OF SOURCE AND FILTER

The estimation of the source and filter parts for both the sinusoidal and residual components is a key aspect of the method. The quality of the results depends largely on the accuracy of the representation. Each component is modeled (and processed) separately as follows.

3.1. Sinusoidal Component

For musical instrument sounds, the sinusoidal component contains most of the acoustic energy present in the signal because musical instruments are designed to have very steady and clear modes of vibration. The sinusoidal component $y_s(t)$ is modeled as sinusoidal tracks (source) $x_s(t)$ and the response $h_s(t)$ of the resonance cavity W_c for each frame. The frequencies of the sinusoids $f_k(t)$ are estimated from the spectrum using quadratic interpolation [11, 13]. The filter response $h_s(t)$ is estimated as the spectral envelope of the spectrum of the sinusoidal component $Y_s(f)$. The spectral envelope estimation method used is extremely important. Wen and Sandler [4] propose an algorithm base on the channel vocoder to model the filter part. However, Röbel [9] showed that "true envelope" (TE) outperformed the other spectral envelope estimation methods tested. TE can be interpreted as the best bandlimited interpolation of the spectral peaks, minimizing the estimation error for the peaks of the spectrum. Thus "true envelope" was chosen to estimate the spectral envelope curve of $Y_s(f)$.

The proposed SF representation replaces the amplitude values of the sinusoidal model with the spectral envelope curve. Figure 2 presents a comparison of the sinusoidal and the SF representation of the amplitudes of partials. The top part of figure 2 shows the original spectrum (grey) and the partials (red), i.e., the spectral peaks selected by the peak-picking algorithm. At the bottom part, we see the partials from sinusoidal analysis (red) and the spectral envelope

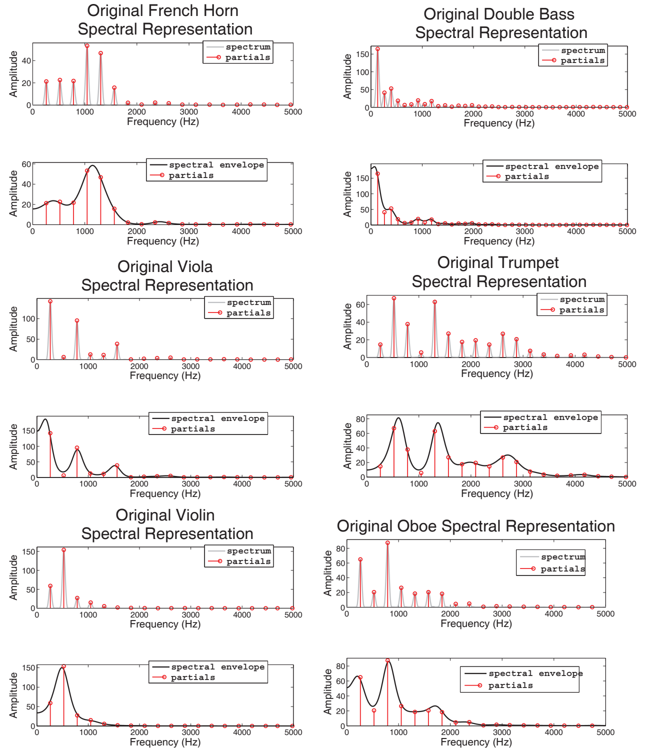
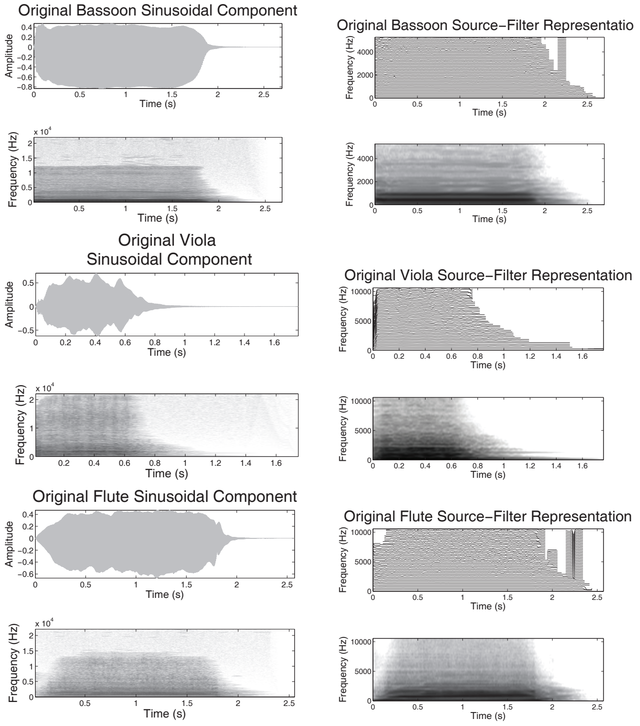


Fig. 2. Spectral view of the source-filter model. Each figure shows the traditional sinusoidal representation at the top and the source-filter representation at the bottom for one analysis frame.



(a) Waveform and Spectrogram

(b) Source and Filter

Fig. 3. Comparison between the spectro-temporal view of the sinusoidal and SF representations. Part a) shows the waveform (top) and spectrogram (bottom). Part b) shows the source (top) and filter (bottom). The source is represented as the temporal variation of the frequencies of the partials, while for the filter the higher amplitudes are darker, like the spectrogram.

curve estimated with “true envelope” (black) representing the amplitude of the partials. The partials now are simply the frequency values at which we “sample” the spectral envelope curve.

Both representations retain essentially the same information (amplitude and frequency of partials) in different ways. The spectral envelope curve is only an interpolation of the amplitudes of the partials using a cepstral model. The sinusoidal model has a more accurate representation of the amplitudes of the partials, while the SF representation presents small errors in the values of the amplitudes inherited from the spectral envelope estimation procedure. On the other hand, the SF model is very flexible when we want to transform source and filter independently. Figure 3 shows the SF model from a spectro-temporal perspective. The sinusoidal representation is not very intuitive to manipulate coherently because there are too many degrees of freedom. Changing the value of only one sinusoidal track renders results that are not perceptually relevant. The SF model is more perceptually intuitive to manipulate because changing the parameters of the spectral envelope representation usually leads to a smoother distribution of spectral energy.

3.2. Residual Component

The residual signal $y_r(t)$ is modeled as a white noise source $x_r(t)$ driving the response of the system W_c. Thus, the response of the resonant cavity to the excitation $x_r(t)$ is modeled as the spectral envelope of $Y_r(f)$ using linear prediction because it provides a better estimate for noise. In this case, the filter is modeled using a spectral envelope curve that follows the average energy of the magnitude spectrum rather than fit the amplitudes of the spectral peaks. The SF residual is mixed into the SF sinusoidal after resynthesis.

3.3. Resynthesis

The result of TE estimation is a set of cepstral coefficients representing the estimated spectral envelope curve. Next, this cepstral based representation is converted to a linear prediction based representation using the spectral power density method. The conversion needs to be done to retrieve the amplitudes of the partials from the LPC representation of the filter. The LPC representation is necessary upon resynthesis because the sinusoids used to represent the partials are the eigenfunctions of LSI systems. In other words, we can simply “sample” the LPC representation of the filter part of the SF model and we obtain the corresponding amplitudes. In mathematical terms, the sinusoidal source $x_s(t)$ is expressed as a sum of sinusoids as in equation 1 with $a_k = 1, \forall k$. Then, if we express the LPC representation of the filter by $H(f)$, we can obtain the amplitudes of the partials of $y_s(t)$ from the frequency values f_k and $H(f)$ using equations 2 and 4 as follows

$$\sum_{k=0}^{K(t)} \left|H\left(f_k\left(t\right)\right)\right| = H\left(\sum_{n=1}^{K(t)} 2\pi f_k t + \psi_k\right) \qquad (5)$$

Thus $y_s(t)$ can be represented as

$$y_s(t) = \sum_{k=0}^{K(t)} s_k(t) H(2\pi f_k t + \psi_k) \qquad (6)$$

where $s_k(t) = \sin(2\pi f_k t + \psi_k)$ is a slowly varying sinusoid with frequency f_k in Hertz. The phase ψ_k is reconstructed integrating the instantaneous frequencies $2\pi f_k t$.

4. EVALUATION OF THE SF MODEL

We should consider two important aspects in the spectral modeling part, accuracy of representation and ease of manipulation. Ideally, the model should represent the original sound accurately and allow independent and coherent manipulation of different parts of the model. One way to test the accuracy of the representation is to resynthesize a sound from the parameters of the model representation and compare it with the original sounds. An accurate model should render sounds that are perceptually similar to the original recordings (or at least close enough depending on the intended application). On the other hand, the ease of manipulation is essential when performing sound transformations. If a representation has too many independent parameters, it becomes cumbersome to manipulate all of them coherently to obtain a certain result. Perceptually speaking, a coherent manipulation of the amplitude values of a spectral representation would change the values of the amplitudes in such a fashion that the balance of the distribution of spectral energy is changed, rather than the amplitudes of isolated partials independently from the partials around it.

The result of a simple listening test that aimed to evaluate how perceptually similar the sounds resynthesized from the SF representation are to the original recordings is explained next. The listening test presented 20 pairs of musical instrument sounds from the Vienna symphonic library (`http://www.vsl.co.at/en/65/71/84/1349.vsl`) and asked the participant to rate the perceptual similarity between them using a scale from 1 to 5 (5 being identical). The Vienna sound database contains samples from most musical instruments commonly found in an orchestra recorded under controlled conditions and played by professional instrumentalists. There are woodwind, brass, plucked and bowed string instrument samples in the database covering the normal pitch range of each instrument. The listening test is available online `http://recherche.ircam.fr/anasyn/caetano/survey/similarity.html`. In total, the results of 80 participants aged between 22 and 67 were used.

In general, the implementation of SF model used in this work was rated between 4 and 3. Except for the bass trumpet sound, which was very "brassy". Thus the results of the similarity test validate the SF model as perceptually similar to the original sounds. Interestingly, most listeners reported using the noisy residual to assess the differences. Some listeners even referred to the sounds as coming from particular instruments, even though this information was not given.

5. CONCLUSIONS AND FUTURE PERSPECTIVES

This work presented a source-filter (SF) model for musical instrument sound transformation that is accurate, compact and flexible. The SF model is based on sinusoidal plus residual decomposition of the original sounds. The sinusoidal source is modeled as sinusoidal tracks driving a time-varying filter estimated with "true envelope". The residual is represented as white noise (source) shaped by a time-varying autoregressive filter. The SF model is very appropriate for sound transformations because of its independent and compact spectral representation. Most sound transformation techniques found in the literature apply sinusoidal models in part due to the quality of the representation of a broad class of sounds. However, a drawback of using sinusoidal models in sound transformations is that the number of parameters is proportional to the number of partials. The SF representation, in turn, models the amplitude of the partials with a spectral envelope curve, which has a limited number of parameters that depend solely on the fundamental frequency of the spectrum, not on the number of partials. The parameters of spectral envelope models normally give smooth continuous transformed spectral envelope curves when manipulated. The SF representation was validated with a perceptual similarity test. Participants were presented the original and SF representation of sounds and asked to assess their perceptual similarity. In general the SF representation was judged perceptually similar to the original recordings. Future perspectives of this work would include using the SF model developed here in musical instrument modeling for instrument recognition, transcription, and sound source separation.

6. REFERENCES

[1] L. Rabiner, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[2] J. D. Markel and A. H. Gray Jr., *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

[3] H. Hahn, A. Röbel, J. J. Burred, and S. Weinzierl, "A source-filter model for quasi-harmonic instruments," in *Proc. DAFx*, 2010.

[4] X. Wen and M. Sandler, "Source-filter modeling in the sinusoidal domain," *J. Audio Eng. Soc.*, vol. 58, no. 10, 2010.

[5] J.J. Burred, A. Röbel, and T. Sikora, "Dynamic spectral envelope modeling for timbre analysis of musical instrument sounds," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 18, no. 3, March 2010.

[6] A. Klapuri, "Analysis of musical instrument sounds by source-filter-decay model," in *Proc. ICASSP*, 2007.

[7] A. Klapuri, T. Virtanen, and T. Heittola, "Sound source separation in monaural music signals using excitation-filter model and em algorithm," in *Proc. ICASSP*, 2010.

[8] D. Arfib, F. Keiler, and U. Zölzer, "Source-filter processing," in *DAFX: Digital Audio Effects*, Udo Zölzer, Ed., pp. 279–320. John Wiley & Sons, second edition edition, 2011.

[9] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Proc. DAFx*, 2005.

[10] M. Caetano and X. Rodet, "Sound morphing by feature interpolation," in *Proc. ICASSP*, 2011.

[11] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust, Speech, and Sig. Proc.*, vol. ASSP-34, pp. 744–754, 1986.

[12] W. Slawson, *Sound Color*, University of California Press, Berkerley, 1985.

[13] X. Serra and J. O. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 49–56, 1990.