STATISTICAL SYNTHESIS OF TRANSIENT AND PITCH-CHANGING SIGNALS

Sofia Cavaco
CITI, Departamento de Informática
Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
2829-516 Caparica, Portugal
scavaco@fct.unl.pt

ABSTRACT

We propose a statistical method for modeling and synthesizing sounds with both sinusoidal and attack transient components. In addition, the sinusoidal component can have pitch-changing characteristics. The method applies multivariate decomposition techniques (such as independent component analysis and principal component analysis) to learn the intrinsic structures that characterize the sound samples. Afterwards these structures are used to synthesize new sounds which can be drawn from the distribution of the real original sound samples. Here we apply the method to impact sounds and show that the method is able to generate new samples that have the characteristic attack transient of impact sounds.

1. INTRODUCTION

Many sound synthesis methods have been proposed over the years. On the one hand there are physical methods, whose models are derived from the properties of the object that produces the sound. These are knowledge based techniques whose models have parameters that are set according to the physics, dynamics and acoustics of the objects. However, those parameters may be hard to estimate when dealing with objects with complex geometries or materials. On the other hand there are signal modeling techniques. These are data-driven techniques that describe the acoustic structure of the sounds and do not require any knowledge of the physics, dynamics and acoustics of the objects. Signal modeling techniques use a set of data samples from which they estimate the parameters used in the models equations and synthesis method.

Most signal modeling techniques use a set of parameters that are estimated from one sound sample. Those parameters are used to later generate sounds. Since these methods use only a single sound sample, they are able to characterize that specific sound but they may fail to characterize the class of sounds of the same type (for instance, if there is a property of those sounds that is not very noticeable in that specific sound sample). In addition, these methods synthesize sounds that are actually not *new* sounds, in fact they consist of a modification of the original sound sample.

One of the difficulties of many signal modeling techniques is on the synthesis of attack transients. These broad band portions of the signal are characterized by a sudden and brief increase in energy, which may be hard to model and reproduce.

Here we propose a data-driven analysis and synthesis method that uses independent component analysis (ICA) and principal component analysis (PCA) to model and synthesize waveforms that contain both sinusoidal and attack transient components. The method is an extension of Cavaco and Lewicki's method for the analysis of the intrinsic structures of impact sounds [1]. That method is able to learn the acoustic properties (such as ringing, resonance, sustain, decay, and onsets) that characterize the sounds. The extension proposed here successfully generates sounds with attack transient and sinusoidal components, which can have pitch-changing characteristics. Moreover, the synthesized sounds consist of new samples drawn from the distribution of the real original sound samples.

The next section discusses some relevant related work. Section 3 reviews Cavaco and Lewicki's intrinsic structure analysis method. The analysis and synthesis extension proposed here and the results obtained are discussed in sections 4 and 5, respectively. Section 6 draws the conclusions and discusses some future work.

2. PREVIOUS WORK ON DATA-DRIVEN SYNTHESIS METHODS

Even though many synthesis methods exist here we focus only on the methods that are relevant to this work, namely on signal modeling techniques. These techniques describe the acoustic structure of the sound, independently of the properties of the object.

One of the best known signal modeling techniques is the phase vocoder [2, 3]. This technique successfully models and synthesizes harmonic signals with static pitch characteristics. Yet, it is not as successful when it comes to modeling pitch changing sounds (i.e., sounds that have partials with time-varying frequencies) and inharmonic sounds, since it does not model the slow frequency variations in the partials and the sinusoids are assumed to be harmonic. However, natural sounds can be inharmonic, and are typically

not purely periodic because their sinusoidal components can have slowly time-varying frequencies.

2.1 Sinusoidal modeling

Over the years, many extensions and alternatives to the phase vocoder have been proposed [4–8]. The most well known alternatives are the sinusoidal models, as the MQ modeling technique proposed by McAulay and Quatieri, and the PARSHL technique proposed by Serra [9–11]. Although, these two methods are very similar, MQ modeling and PARSHL were developed independently. While MQ modeling was developed to represent and synthesize speech, PARSHL focused on musical sounds. Nonetheless, the main ideas behind the two methods are similar. The two methods have an analysis module, which represents the sounds as a sum of sinusoids with slowly varying amplitude and frequency. Since these models use instantaneous frequencies (instead of a constant frequency for each sinusoid) they are able to characterize pitch changing sounds. These methods consider that each slowly varying sinusoidal component of the signal is represented by a horizontal ridge of energy in the signal's spectrogram, and they use a *peak tracking* algorithm to identify those horizontal ridges in the spectrogram. This algorithm looks for the local maxima in each frame of the spectrogram, and connects the peaks from different frames to form tracks. It then represents each track as a sequence of parameters that determine the track's instantaneous amplitudes and frequencies.

MQ modeling and PARSHL also have a synthesis module, which uses that representation previously obtained to synthesize the sounds. The synthesis can be done in the time domain with an oscillator bank: the oscillator bank generates a sinusoid for each track, and these sinusoids are added to obtain the final synthesized signal. Alternatively, and assuming the original phase values are preserved, the synthesis can be done in the frequency domain with the inverse Fourier transform.

While PARSHL and the MQ method successfully model and synthesize inharmonic and pitch-changing sounds, they are inefficient when modeling and synthesizing signals with a broader spectrum, like noise and transients. These methods try to model the noise and transients as a sum of sinusoids, which is very ineffective and computationally expensive as typically these signals contain energy in the whole spectrum and would need to be modeled by a large number of sinusoids. As a response to this problem, Serra and Smith developed an extension of PARSHL which makes a distinction between the sinusoidal and broad band spectrum components in the signal [12,13]. This method, spectral modeling synthesis (SMS), combines sinusoidal modeling and noise modeling to represent and synthesize sounds with both sinusoidal and noise components [13].

The noise modeling part of SMS, represents the non-sinusoidal portions of the signal, which include the excitation energy that is not transformed into stationary vibrations of the sound source. This module assumes that the non-sinusoidal components consist of stochastic signals, which do not require a precise description of the time-varying magnitude shape of each frequency bin and that can be represented by a density function that describes the expected magnitude of each frequency bin over time. It uses a time-varying frequency-shaping filter to represent the density function and applies it to white noise to represent the stochastic components of the signal.

SMS successfully models and synthesizes inharmonic and pitch-changing sounds that have broad band spectrum components with stochastic characteristics, like the sound of the bow sliding against the strings of an instrument, or the sound of breath in a wind instrument. However, this technique may fail to effectively model and synthesize the transient portion of the signals. Transients are not well represented by sinusoidal modeling due to their broad band spectrum characteristics (as they would have to be represented by a quite large quantity of sinusoids), and they are not well represented by the noise model of SMS because they need precise time synchronization between the various frequency components in their representation. As a result, when attack transients are modeled and synthesized with the techniques described here, they lose their characteristic sharpness and sound more like noise than like attacks. As a response to this problem, some methods have been developed that treat transients as a separate kind of signal [14–17]. Below we describe one of these methods, namely the transient modeling synthesis.

2.2 Transient modeling synthesis

Transient modeling synthesis (TMS) is an analysis/synthesis method proposed by Verma and colleagues to model and synthesize transient sounds [16, 17]. TMS can be combined with SMS to model the sinusoidal, attack transients and noise portions in the sounds.

While a sinusoid is a slowly varying curve in the time domain and a sharp peak in the frequency domain, a transient is sharp in the time domain and it can be represented by a slowly varying curve in the frequency domain. In order to model the transients as slowly varying curves, TMS uses a two-step space transformation. In the first step, it computes the discrete cosine transform (DCT) of the waveform, to represent the signal in a new space of cosine basis functions (or simply, frequency) by amplitude. The DCT is used because it represents the transients as sinusoids in the frequency domain. This transform maps the signal into a frequency by amplitude space and retains the phase information.

In the second step, TMS computes the magnitude spectrogram of the DCT of the signal, to represent the signal in a frequency by time space. The signal is represented in a space of spectrogram frame number (in the horizontal axis) by discrete Fourier transform (DFT) bin number (in the vertical axis). The spectrogram frame represents a window of DCT bins, and in turn, a DCT bin corresponds to a cosine basis function, or frequency. Thus, the frame stands for frequency. The DCT of an impulse at the beginning (or left side) of the time window is represented by a low frequency sinusoid. Impulses that appear later in time (i.e., towards the right side of the window) are represented by higher frequency sinusoids. Thus, there is a correspondence between time and frequency of the sinusoids. The

DFT bins correspond to the frequencies of these sinusoids, and therefore, they also correspond to time.

In order to model the slowly varying sinusoids on the spectrogram of the DCT, TMS uses a peak tracking algorithm (described in section 2.1 for sinusoidal modeling). It identifies these horizontal lines, and represents them by tracks that consist of sequences of parameters that determine the instantaneous amplitudes, and the onset times of the transients (coded in terms of frequency, i.e., DCT basis functions).

Once the transients are modeled by these sequences of parameters, and after optional modifications to the parameters, the transients can be synthesized. This can be done with a bank of oscillators to generate a sinusoid for each track, which are added to obtain a signal in the DCT domain. Finally, the inverse DCT transforms the signal back into the time domain. The result is a waveform that contains the transients of the original signal.

2.3 Other methods for modeling and synthesis of sinusoids and transients

The previous sections discuss some techniques to model and synthesize the sinusoids and transients in the signals. Yet, these are not the only options that may be considered, and more work has been proposed in this area. For example, Depalle and Hélie proposed a parametric method to extract the sinusoids from the spectrogram [18]. George and Smith proposed the analysis-by-synthesis overlap-add method, which extracts sinusoids in an iterative way [19]. Also other methods have been proposed to model transients. For instance, Christensen and van de Paar model transients with a sum of sinusoids, whose amplitudes are modulated by gamma envelopes [20]. Nsabimana and Zölzer proposed improvements to TMS [21].

3. THE INTRINSIC STRUCTURES ANALYSIS METHOD

The goal of the Intrinsic Structures Analysis (ISA) method proposed by Cavaco and Lewicki is to represent the structures in the sounds [1]. Instead of extracting pre-defined features of the sounds, like Mel-frequency cepstral coefficients or other short-time features, the method learns a set of temporal and spectral features from the data. For that end, the method represents the sounds with magnitude spectrograms ($\mathbf{S}^k$, where k is the index of the sound) and then it extracts sound features from the transposed spectrograms.

In more detail, the method performs ICA [1] (or PCA) on the concatenation of the transposed spectrograms:

$$\left(\left(\mathbf{S}^1\right)^T, \ldots, \left(\mathbf{S}^{\mathrm{K}}\right)^T\right), \quad (1)$$

where $(\mathbf{S}^k)^T$ is the transpose of the magnitude spectrogram of sound k, and this expression represents a huge matrix that consists of the concatenation of the transposed spectrograms in the data set. Thus the columns in this matrix are the transpose of the bins in the data set, that is in the spectrograms $\mathbf{S}^1, \ldots, \mathbf{S}^K$.

As a result the method learns a set of temporal basis functions $\mathbf{\Phi}$ that describe the temporal regularities in the bins in the data set. Along with that, the method also extracts a set of weights, such that each bin is now represented by a set of weights (one weight per basis function). If we consider all the weights related to the bins of one spectrogram and one basis function, we have a (spectral) vector, which we call *spectral source signal*.

Thus, the concatenated transposed spectrograms can be expressed as the linear combination of these spectral source signals:

$$\left(\left(\mathbf{S}^1\right)^T, \ldots, \left(\mathbf{S}^{\mathrm{K}}\right)^T\right) = \mathbf{\Phi}\left(\mathbf{C}^1, \ldots, \mathbf{C}^{\mathrm{K}}\right) \ , \quad (2)$$

where $\mathbf{\Phi}$ is a matrix with one temporal basis function per column, K is the number of sounds, and, for every $1 \leqslant k \leqslant \mathrm{K}$, $\mathbf{C}^k$ is a matrix of spectral source signals: there is one such matrix for each spectrogram and the rows of these matrices contain the source signals. The ith row of every matrix $\mathbf{C}^k$ has the weights to the ith basis function in $\mathbf{\Phi}$ (that is, the ith column of $\mathbf{\Phi}$).

Figure 1 shows an example of the temporal basis functions and spectral source signals obtained by PCA of the spectrograms of a set of ten sounds from impacts on an aluminum rod (where the spectrograms are arranged as in equation 1). Figure 1a shows 3 basis functions, that is, 3 columns from $\mathbf{\Phi}$ (plotted horizontally), and figure 1b shows 3 spectral source signals. While the basis functions in $\mathbf{\Phi}$ are common to all spectrograms, each spectral source signal is associated to a specific basis function and a specific spectrogram. For instance, in this example, $\mathbf{c}_1^{Al1}$ (the first row in $\mathbf{C}^{Al1}$) is a vector that contains the weights associated to $\mathbf{S}^{Al1}$ and $\boldsymbol{\phi}_1$.

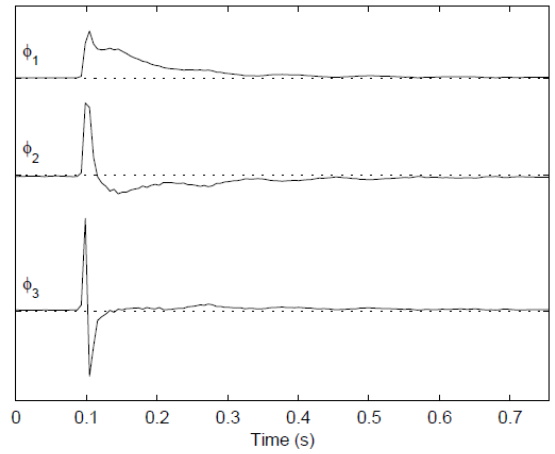
Since there still are temporal regularities that can be further explored in the matrices $\mathbf{C}^k$, the method has a second stage that consists of modeling those regularities. For that purpose, the method first constructs new matrices of source signals, $\mathbf{D}^i$, such that each matrix contains the source signals related to one basis function. Matrix $\mathbf{D}^i$ contains the source signals related to the ith basis function in $\mathbf{\Phi}$, that is, each column in this matrix is the ith spectral source signal for a different sound,

$$\mathbf{D}^i = \left(\mathbf{c}_i^1, \ldots, \mathbf{c}_i^K\right), \quad (3)$$

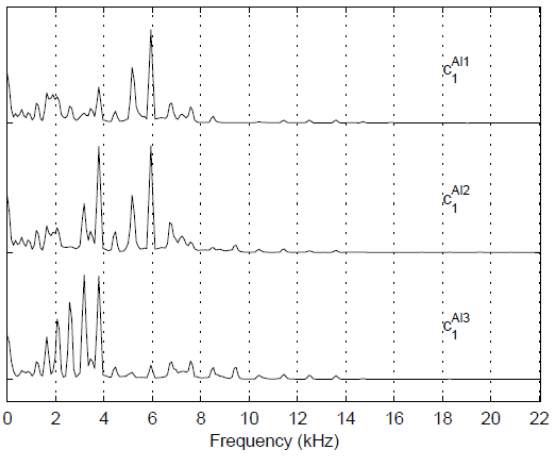
where $\mathbf{c}_i^k$ is the ith source signal (that is, row) in $\mathbf{C}^k$. There will be as many $\mathbf{D}^i$ matrices as basis functions in $\mathbf{\Phi}$. (Note that, $\mathbf{c}_i^k$ is a column in $\mathbf{D}^i$ and a row in $\mathbf{C}^k$. Thus, since here vectors are column vectors, we should have used the transposed, $(\mathbf{c}_i^k)^T$, when referring to the row of $\mathbf{C}^k$, but we will drop the transpose symbol whenever there is no ambiguity.)

The method then applies ICA or PCA to each matrix $\mathbf{D}^i$. As a result, it learns a set of spectral basis functions, represented in the columns of $\mathbf{\Psi}^i$, and a set of weights $\mathbf{U}^i$ such

[1] We used the *fastica* software package [22]. Even though in the text we mention that we use the magnitude spectrogram, in fact we use both the magnitude spectrogram $\mathbf{S}^k$ and its negative, i.e., we used the extended matrix $(-\mathbf{S}^k, \mathbf{S}^k)$. Therefore the mean is zero, and this is done because ICA models the variation around the data mean. This way the results from ICA describe the signal rising and falling from zero, rather than the spectrogram mean.

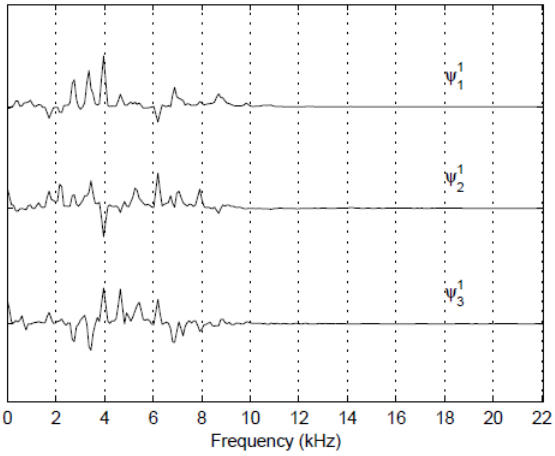


(a)

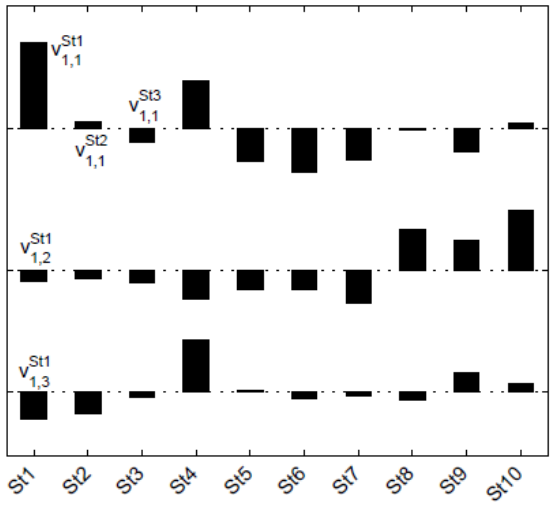


(b)

Figure 1. Temporal basis functions in $\mathbf{\Phi}$ and spectral source signals in $\mathbf{C}^k$ (with $1 \leqslant k \leqslant 3$) obtained by PCA of a set of 10 impacts on an aluminum rod ({Al1, ..., Al10}) [1]. (a) The first three basis functions in $\mathbf{\Phi}$ (that is, the first 3 columns in this matrix, which for convenience here we plotted horizontally): ϕ_1, ϕ_2 and ϕ_3. (b) The first spectral source signal for sounds Al1, Al2 and Al3 (that is, the first row in matrixes $\mathbf{C}^{Al1}$, $\mathbf{C}^{Al2}$ and $\mathbf{C}^{Al3}$): $\mathbf{c}_1^{Al1}$, $\mathbf{c}_1^{Al2}$ and $\mathbf{c}_1^{Al3}$.



(a)



(b)

Figure 2. Spectral basis functions in $\mathbf{\Psi}^1$ and weights obtained by PCA of the source signals which were obtained by PCA of a set of 10 impacts on a steel rod ({St1, St2, ..., St10}) [1]. (a) First three spectral basis functions from $\mathbf{\Psi}^1$ (that is, the first 3 columns in this matrix, which for convenience here we plotted horizontally). (b) Weights in $\mathbf{U}^1$, where $v_{1,j}^k$ is the weight for sound k and basis function ψ_j^1.

that

$$\mathbf{D}^i = \mathbf{\Psi}^i \mathbf{U}^i \,. \tag{4}$$

$\mathbf{\Psi}^i$ contains the spectral basis functions that reflect the regularities in the spectral source signals associated to the ith basis function in $\mathbf{\Phi}$. While the jth row in $\mathbf{U}^i$ is related to a the jth basis function (that is, column) in $\mathbf{\Psi}^i$, the kth column in $\mathbf{U}^i$ is related to sound k. Now, we can reorganize the weights in matrices $\mathbf{U}^i$ and build matrix $\mathbf{V}^k$ that contains all the weights related to sound k: each column in $\mathbf{V}^k$ is the kth column of a different $\mathbf{U}^i$.

Figure 2 shows an example of the spectral basis functions and weights obtained by PCA of the spectral source signals in $\mathbf{D}^1$, where, in turn, the source signals were obtained by PCA of the spectrograms of a set of ten sounds from impacts on an steel rod (where the spectrograms are arranged as in equation 1). Figure 2a shows 3 basis functions, that is, 3 columns from $\mathbf{\Psi}^1$ (plotted horizontally). Figure 2b shows the weights in $\mathbf{U}^1$, that is, it shows the first column of each matrix $\mathbf{V}^{St1}, \ldots, \mathbf{V}^{St10}$.

To conclude, the ISA method represents the sounds with a model that consists of the triple $(\mathbf{\Phi}, \mathbf{\Psi}, \mathbf{V})$, where $\mathbf{V}$ contains $\mathbf{V}^1, \ldots, \mathbf{V}^K$, $\mathbf{\Psi}$ contains the sets $\mathbf{\Psi}^1, \ldots, \mathbf{\Psi}^I$, and I is the number of basis functions in $\mathbf{\Phi}$. For a more complete explanation of this model and more figures with results, please refer to [1].

4. MODELING INTRINSIC STRUCTURES OF IMPACT SOUNDS WITH ACCURATE TRANSIENT SYNTHESIS

Using the ISA method, it is possible to model all parts (attack, sinusoids, etc.) of impact sounds. The learned model is able to identify and describe the structures of the

sounds, which can be used to synthesize the pure tones of the sounds, for instance using sinusoidal modeling and synthesis [9–11]. However, the same is not true when it comes to synthesizing the attack transients of the sounds. This is due to the lack of phase information in the magnitude spectrograms. Here, we extend the ISA method such that, the model it learns can be used to synthesize both the sinusoidal and transient parts in the sounds.

The extended method, which we name the Intrinsic Structure Analysis and Synthesis (ISAS) method, is composed of an *analysis stage* that decomposes the sounds into basis functions that represent their structures, and a *synthesis stage* that uses those basis functions to produce waveforms. The analysis stage starts by dividing each signal into two parts: the *sinusoidal sub-signal*, which we call s and contains the pure tones from the original signal, and the *transient sub-signal*, which we call a (from attack) and which contains the transients from the original signal. Afterwards, the sub-signals are treated differently. The sinusoidal sub-signals are analyzed by the ISA method, which extracts the temporal basis functions $\mathbf{\Phi}$, the spectral basis functions $\mathbf{\Psi}$ and weights $\mathbf{V}$. The transient sub-signals are analyzed by the transients method described in section 4.1, which also learns sets of spectral and temporal basis functions, along with the weights. The synthesis stage receives the basis functions learned by the ISA method and transients method, and uses them to synthesize new sounds (section 4.2).

4.1 The transients method

In order to avoid the problems introduced by the lack of phase information, the transients method does not use a spectrogram to represent this sub-signal. Instead, it uses the representation proposed by Verma and colleagues in the context of TMS, that is, the spectrogram of the DCT (see section 2.2) [16, 17]. So, the analysis of the transients sub-signal starts by representing ensembles of transients signals (or a single transients signal) with the spectrogram of the DCT of the signals' waveforms, which here we call $\mathbf{Z}^k$. (Note that the frames of $\mathbf{Z}^k$ correspond to frequencies of the signal, and the bins correspond to time, while in $\mathbf{S}^k$, the frames correspond to time slices, and the bins to frequency intervals.) This allows representing the transients by a periodic signal (composed of cosine waves) that is easily modeled, modified and synthesized, while preserving the transient characteristics of the signal.

The next step consists of concatenating the spectrograms in the same way as with the ISA method, such that the data matrix $\mathbf{X}$ is the concatenation of transposed spectrograms, $\mathbf{X} = ((\mathbf{Z}^1)^T, (\mathbf{Z}^2)^T, ...(\mathbf{Z}^K)^T)$. Then the method applies ICA or PCA to this matrix (whose concatenated frames, i.e. the rows of $\mathbf{X}$, are the signal mixtures) and as a result, it learns a set of basis functions and a set of source signals (i.e., vectors of weights). While at this step, the ISA method learns the temporal basis functions $\mathbf{\Phi}$, the transients method learns a set of spectral basis functions, which we call $\mathbf{\Upsilon}$. The basis functions learned at this step are spectra because the frames of the matrices $\mathbf{Z}^k$ correspond to frequencies of the original signal.

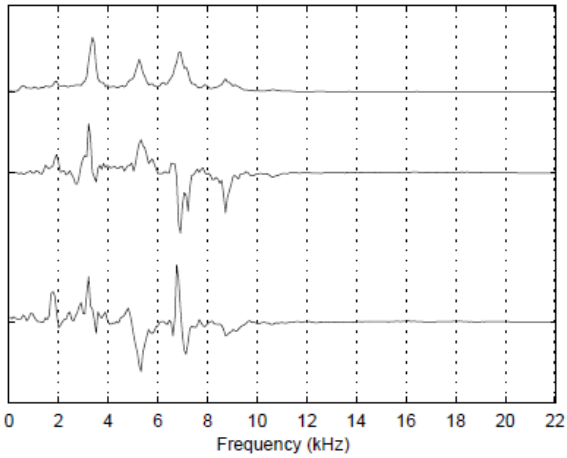


Figure 3. Spectral basis functions $\mathbf{\Upsilon}$ learned by PCA of the set of 10 transients sub-signals from impacts on a steel rod ($\{$St1, St2, ..., St10$\}$). The figure shows the first three (most dominant) basis functions from top to bottom: Υ_1, Υ_2 and Υ_3.

As an illustration, Figure 3 shows the most dominant basis functions in $\mathbf{\Upsilon}$ learned by PCA of transients from impacts on a steel rod. As one could expect, these basis functions have some broad band characteristics (for instance, comparing to figure 2a here the partials are not as well defined), which is consistent with the broad band characteristics of the transients.

The source signals obtained at this stage are time-varying functions because the bins of $\mathbf{Z}^k$ correspond to time of the original signal. Just like with the ISA method, the transients method has a further stage which consists of analyzing the structures in these temporal source signals (vectors of weights) obtained in the first stage along with $\mathbf{\Upsilon}$. The process is the same as explained in section 3 but now the method uses matrices of temporal source signals instead of the matrices $\mathbf{C}^k$. As a consequence, the basis functions learned at this second stage are time-varying functions, which we call $\mathbf{\Gamma}$. As an example, figure 4 shows temporal basis functions learned by PCA. (The same sounds were used both in figure 3 and figure 4.) These basis functions characterize the temporal structure of transients and, as it can be observed they are very sharp (they have very sudden increases of energy and fast decays) which is consistent with the characteristics of transients. Associated with each temporal basis function in $\mathbf{\Gamma}$ and each transients sub-signal, there is a weight that scales the basis function. Here, $\mathbf{O}$ is the set of those weights.

To summarize, the transient method is very similar to the ISA method but instead of initially representing the sounds with a spectrogram of the waveform, it represents them with the spectrogram of the DCT of the waveform ($\mathbf{Z}^k$). It then represents the spectral and temporal structures in these spectrograms with sets of spectral and temporal basis functions. To learn these basis functions, it uses the same processes as the ISA method. Since the frames of $\mathbf{Z}^k$ correspond to frequencies of the original signal, and the bins correspond to time, the transient method first obtains a set

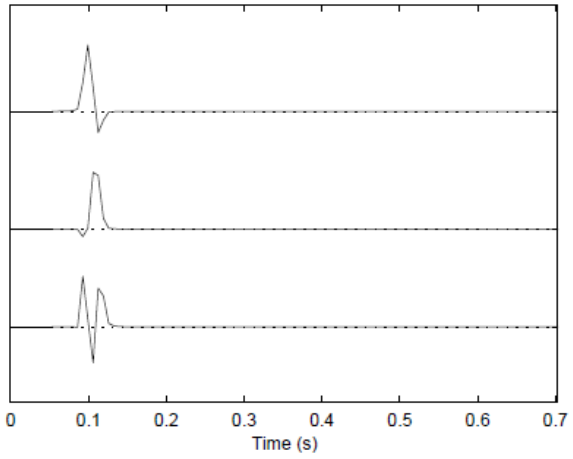


Figure 4. Temporal basis functions $\mathbf{\Gamma}$ learned by PCA of the temporal source signals, which in turn were obtained by PCA of the set of 10 transients sub-signals from impacts on a steel rod ($\{$St1, St2, ..., St10$\}$). The first three (most dominant) basis functions are shown from top to bottom: γ_1, γ_2 and γ_3.

of spectral basis functions $\mathbf{\Upsilon}$, and a set of temporal source signals. Afterwards, it learns a set of temporal basis functions $\mathbf{\Gamma}$, and a set of weights $\mathbf{O}$ (by analyzing the temporal source signals obtained in the first stage). The transients method represents the sounds with a model that consists of the triple $(\mathbf{\Upsilon}, \mathbf{\Gamma}, \mathbf{O})$.

4.2 The synthesis stage

Once the analysis stage has completed, we have sets of basis functions ($\mathbf{\Phi}$, $\mathbf{\Psi}$, $\mathbf{\Upsilon}$ and $\mathbf{\Gamma}$) that describe the structures in sinusoidal and transients sub-signals. These basis functions and the weights in $\mathbf{V}$ and $\mathbf{O}$ can be used to synthesize sounds. The synthesis is sub-divided into two methods: one that synthesizes a sinusoidal waveform s' (section 4.2.1) and another that synthesizes a transients waveform a' (section 4.2.2). After these two waveforms have been generated, the synthesis stage combines them to produce the final synthesized signal,

$$y(t) = s'(t) + a'(t). \tag{5}$$

Optionally, the basis functions and weights can be modified in order to obtain new sounds. For instance, the synthesis stage can consider only a subset of basis functions or it can modify in some way the shape of one (or more) basis functions. Also, instead of using the original weights, the synthesis stage can use newly generated weights. If those new weights are drawn from the distributions in $\mathbf{V}$ and $\mathbf{O}$, the method can obtain new realistic impact sounds.

4.2.1 Synthesis of the sinusoidal sub-signal

This stage uses the basis functions in $\mathbf{\Phi}$ and $\mathbf{\Psi}$ to synthesize a sinusoidal signal s'. Since the sinusoidal sub-signal s is represented by a magnitude spectrogram, the original phase information is not retained. This loss of phase information is not critical for synthesizing s' because phase information is not perceptually significant in the periodic regions of the sounds. Here we use an algorithm similar to sinusoidal modeling and synthesis, as in the MQ method and PARSHL [9–11]. The method builds a spectrogram from the information in the sets of basis functions $\mathbf{\Phi}$ and $\mathbf{\Psi}$, and weights $\mathbf{V}$. Then it uses the peak tracking algorithm of sinusoidal modeling to extract the parameters that represent the tracks of peaks in the spectrogram. These parameters are then used by sinusoidal synthesis (with a bank of oscillators) to obtain a sinusoidal waveform s'.

4.2.2 Synthesis of the transients sub-signal

This stage uses the basis functions in $\mathbf{\Upsilon}$ and $\mathbf{\Gamma}$ to synthesize a transients signal a'. It starts by building a spectrogram from the information in the basis functions in $\mathbf{\Upsilon}$ and $\mathbf{\Gamma}$, and weights $\mathbf{O}$. Then it uses a process similar to TMS, to produce a transients waveform a' from this spectrogram: It uses sinusoidal modeling to model the energy tracks in this spectrogram, and converts the tracks into a waveform by sinusoidal synthesis and an inverse DCT.

5. RESULTS

In this section we discuss the results obtained by the ISAS method. The data used here to train the model, that is, to learn the basis functions $\mathbf{\Phi}$, $\mathbf{\Psi}$, $\mathbf{\Upsilon}$ and $\mathbf{\Gamma}$, consists of sounds from impacts on metal rods. A wooden rod, with a much shorter length but the same diameter, was used as a mallet. Since the rods were hit by hand, there were slight variations on the impact location and force. The sounds were digitized using a sampling frequency of 44 100 Hz.

The ISA method is able to represent both the steady and attack structures in the sound, and its results can be used to synthesize the steady portions of the sound. However, because of the loss of phase information inherent to this method, the same is not true for the transients: when synthesized, they will sound less sharp than the real transients. To illustrate these limitations, the top line of figure 5 shows a waveform, y_{ISA}, obtained by the ISA method and by the synthesis procedure described in section 4.2.1. Here, the sets of basis functions $\mathbf{\Phi}$, $\mathbf{\Psi}$ and coefficients $\mathbf{V}$, were learned by the ISA method (with ICA of the transposed spectrograms of ten sounds from impacts on a zinc plated steel rod, and with PCA of the spectral source signals). These basis functions and the weights obtained for sound $Zn1$, that is, $\mathbf{V}^{Zn1}$, were used to produce a spectrogram, which was modeled and turned into waveform y_{ISA} by the synthesis procedure described in section 4.2.1. This waveform contains the steady, slower decaying, portion of the sound, but lacks the initial sharp and big increase of energy that is characteristic of the attack portion of these impact sounds.

As explained above, the ISAS method produces two waveforms, a sinusoidal waveform s' and a transients waveform a', which can be combined to produce the final synthesized waveform y. Let us first look into the sinusoidal waveform s'. To illustrate the sinusoidal waveforms obtained, figure 6 shows a sinusoidal waveform s' obtained by the ISAS method. As above, here, the sets of basis functions $\mathbf{\Phi}$, $\mathbf{\Psi}$ and coefficients $\mathbf{V}$, were learned with ICA of the

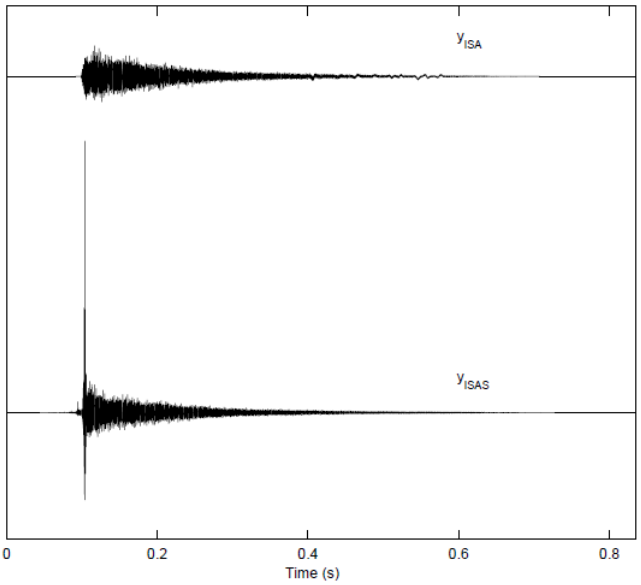


Figure 5. Synthesized signals with a training data set of 10 sounds from impacts on a zinc plated steel rod ({Zn1, Zn2, ..., Zn10}). The weights in $\mathbf{V}^{Zn1}$ were used to synthesize both waveforms. (Top) y_{ISA} was synthesized by sinusoidal modeling and synthesis and the ISA method. (Bottom) y_{ISAS} was synthesized by the ISAS method.

transposed spectrograms of ten sinusoidal sub-signals from impacts on a zinc plated steel rod, and PCA of the spectral source signals. These basis functions and the coefficients in $\mathbf{V}^{Zn1}$ were used to produce a spectrogram, which was then used to obtain a waveform s' with the synthesis procedure described in section 4.2.1. As can be observed, s' is quite similar to y_{ISA} in figure 5.

Now, in contrast to the ISA method, the ISAS method can deal with the synthesis of attack transients. The bottom waveform in figure 5 shows a waveform, y_{ISAS}, obtained by the ISAS method. This waveform consists of the sum of the synthesized sinusoidal and transients waveforms (see equation 5). The synthesized sinusoidal waveform consists of the waveform s' shown in figure 6. In order to synthesize the transients waveform a', the ISAS method used the transients method described in section 4.1 with ten transients sub-signals from impacts on a zinc plated steel rod, along with the synthesis method described in section 4.2.2 and the weights obtained for sound $Zn1$. As it can easily be observed in the figure, this waveform starts with an attack transient, which is the very brief part of the sound at around 0.1 s with a very sharp increase and decrease of energy. This attack transient is followed by a slower decaying portion, which corresponds to the sinusoidal part of the sound. This demonstrates that our goal here was successfully achieved, that is, the ISAS method can synthesize waveforms that preserve the transient characteristics of the signals.

6. CONCLUSIONS

Here we proposed a statistical analysis and synthesis method that is able to deal with both the sinusoidal and attack transient portion of the sounds and does not require any knowledge of the physics, dynamics and acoustics of the objects. The method is an extension to the ISA method [1]. The ISA method is able to model the intrinsic structures of all parts of the sounds (such as the attack, decay and sustain portions, and even other interesting properties such as ringing). The structures (basis functions) learned by that method can be used to synthesize the sinusoidal portions of impact sounds. However, due to the loss of phase information inherent to the ISA method, the attack portion of the sounds cannot be successfully synthesized.

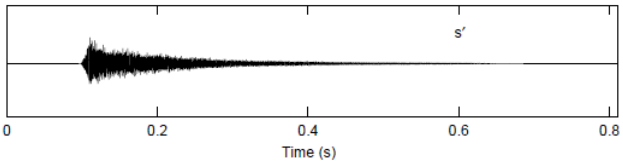


Figure 6. Sinusoidal waveform s' synthesized by the ISAS method with a training data set composed of the sinusoidal sub-signals extracted from 10 sounds from impacts on a zinc plated steel rod ({Zn1, Zn2, ..., Zn10}).

On the other hand, the extension proposed here (the ISAS method) is able to synthesize waveforms that contain both attack transients and slower decay (sinusoidal) portions, which can also have pitch-changing characteristics. This extension treats the sinusoidal and transients parts of the signal in different ways: it analyses and synthesizes them differently. It starts with a different initial representation for each case. The sinusoidal portions are represented with a magnitude spectrogram and the transient portions are represented with the spectrogram of the DCT. This allows representing the transients by a periodic signal that can then be analyzed in the same way as the sinusoidal parts. The method then learns the spectral and temporal structures that represent each portion of the sounds. These structures can then be used to synthesize new sounds that contain both the sinusoidal and attack transient portions.

Other methods have been proposed to synthesize sinusoidal and transient sounds, such as SMS and TMS [9–11, 16, 17], but there are some key differences. Most of those signal modeling techniques analyse one sound sample and generate a new sound that consists of a modification of the original sound. The ISAS method works in a different way: it analyzes a set of sound samples. Since it analyses the structures of a set of sounds and not those of a single sound, it is able to extract the intrinsic properties of that class of sounds. Thus, given a distribution of sound samples from the same class, the ISAS method is able to generate new sounds from within that distribution.

As future work, we are studying the possibility of generating sounds from the interpolation of different classes. For instance, given a set of impacts on the edge of an object and another set of impacts on the center of the object, we are studying the possibility of generating sounds from impacts on intermediate locations. Also, the sounds used here were from impacts on metal rods, but we are planning to use the method to generate other types of sounds, like impacts on other objects.

Acknowledgments

We would like to thank Dr. M. Lewicki for his advice, and also Dr. L. Holt and Dr. V. Ming for help recording the impact sounds used here.

This work was supported by grant UTA-Exp/MAI/0025/2009 from Fundação para a Ciência e a Tecnologia (Portugal) and fellowships from Fundação para a Ciência e a Tecnologia and Fundação Calouste Gulbenkian (Portugal).

7. REFERENCES

[1] S. Cavaco and M. Lewicki, "Statistical modeling of intrinsic structures in impact sounds," *Journal of the Acoustical Society of America*, vol. 121, no. 6, pp. 3558–3568, June 2007.

[2] J. Flanagan and R. Golden, "Phase vocoder," *Bell System Technical Journal*, vol. 45, pp. 1493–1509, November 1966.

[3] M. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 243–248, June 1976.

[4] J. Marques and L. Almeida, "New basis functions for sinusoidal decomposition," in *Proceedings of EUROCON*, Stockholm, Sweden, 1988.

[5] D. Griffin and J. Lim, "Multiband excitation vocoder," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 36, issue 8, 1988, pp. 1223–1235.

[6] M. Puckette, "Phase-locked vocoder," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 1995.

[7] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, May 1999.

[8] J. Laroche and M.Dolson, "New phase vocoder techniques for pitch-shifting, harmonizing and other exotic effects," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 1999, pp. 91–94.

[9] R. McAulay and T.Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, no. 4, pp. 744–754, August 1986.

[10] R. McAulay and T. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 4, pp. 121–173.

[11] J. Smith and X. Serra, "PARSHL: an analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," in *Proceedings of the International Computer Music Conference*, 1987, pp. 290–297.

[12] X. Serra and J. Smith, "Spectral modeling synthesis: a sound analysis/synthesis based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[13] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, C. Roads, S. Pope, A. Picialli, and G. De Poli, Eds. Swets & Zeitlinger Publishers, 1997.

[14] P. Masri, "Computer modeling of sound for transformation and synthesis of musical signal," Ph.D. dissertation, University of Bristol, 1996.

[15] M. Ali, "Adaptive signal representation with applications in audio coding," Ph.D. dissertation, University of Minnesota, 1996.

[16] T. Verma, S. Levine, and T. Meng, "Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals," in *Proceedings of the International Computer Music Conference*, September 1997, pp. 164–167.

[17] T. Verma and T. Meng, "Extending spectral modeling synthesis with transient modeling synthesis," *Computer Music Journal*, vol. 24, no. 2, pp. 47–59, 2000.

[18] P. Depalle and T. Hélie, "Extraction of spectral peak parameters using a short-time Fourier transform modeling and no sidelobe windows," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, October 1997.

[19] E. George and M. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones," *Journal of the Audio Engineering Society*, vol. 40, no. 6, pp. 497–515, June 1992.

[20] M. F. Christensen and S. van de Par, "Efficient parametric coding of transients," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1340–1351, 2006.

[21] F. Nsabimana and U. Zölzer, "Analysis/synthesis of transients in audio signals," presented at Jahrestagung für die Akustik DAGA'06, Braunschweig, Germany, March 2006.

[22] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley and sons, inc., 2001.