Statistical Modeling and Synthesis of Intrinsic Structures in Impact Sounds

Sofia C.F.M. Cavaco

CMU-CS-07-138

July 2007

School of Computer Science
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:

Michael S. Lewicki, Chair
Roger B. Dannenberg
Richard M. Stern
Daniel P.W. Ellis, Columbia University

Submitted in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy

This research was sponsored by Fundação para a Ciência e a Tecnologia (Portugal), Fundação Calouste Gulbenkian (Portugal), the National Science Foundation (NSF) Grant No. IIS0238351, the Office of Naval Research (ONR) Grant No. N000140110677, and the W. M. Keck Foundation. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Acknowledgements

I would like to thank my supervisor, Dr. Michael Lewicki, for all the help, guidance and encouragement he gave me throughout the years I spent at Carnegie Mellon University. I also thank him for always giving me the freedom to explore the paths I believed in and found most interesting.

I would also like to thank everyone else in the Laboratory for Computational Perception and Statistical Learning: Dr. Evan Smith, Yan Karklin, Xuejing Chen, Daniel Leeds, and especially Dr. Eizaburo Doi and Doru Balcan for all the very useful discussions and interactions. Many thanks also to Doru, Yan and Daniel for proofreading parts of this dissertation.

It was a great pleasure to work and interact with Dr. Roger Dannenberg, from whom I have learned a great deal on audio and acoustics. I thank him for having me as a TA for both of his computer music courses. I also thank him for all the help and very useful feedback he gave me. I would like to thank Drs. Richard Stern, Daniel Ellis, Tom Cortina, and Mark Kahr for their interest in my work and all the valuable advice, comments, and feedback they gave me. I also would like to thank Dr. Luís Monteiro, who encouraged and supported my coming to the United States to pursue my Ph.D. studies.

Lastly but not least, I would like to thank my family. I thank my sisters, Patrícia and Marta, for always being there for me, especially for the support and encouragement they gave me during the first years of my Ph.D. studies. Many thanks to Giles for proofreading most of this dissertation. I thank my parents, Teresa and António, for their support, encouragement and visits, but mostly I thank them for raising me to not be afraid of changes and the unknown, and to believe in and pursue my dreams even when the path to achieve them is not the most comfortable one. I thank my mother in law, Helena, for her support and many visits. Finally, I thank my husband Tiago for being so supportive and understanding. I thank him also for all the help he gave me throughout my Ph.D. studies, and for making my stay in Pittsburgh so enjoyable and helping me make a home so far from Portugal.

Contents

Chapter 1

Introduction

Models of sounds have proven useful in many fields, such as in the study of sound perception. When one object impacts another, one can readily perceive many intrinsic properties of the resulting sound that allow us to distinguish differences in material, surface properties, and even object shape, in spite of wide variations in the raw acoustic waveform. Some studies have investigated which acoustic properties of impact sounds are perceptually relevant and give the perception of physical properties such as material, hollowness or geometry [Avanzini and Rocchesso, 2001b, Klatzky, Pai, and Krotkov, 2000, Lutfi and Oh, 1997, Lakatos, McAdams, and Caussé, 1997, Lutfi, 2001, Freed, 1990, Gaver, 1988]. However these studies are limited by the models they use to represent the sounds, as they cannot determine the relevance of features that are not explicitily represented in the models' equations.

Models of sounds are also useful in other fields, such as sound recognition, identification of properties of the sound-producing objects (like material or length), identification of the events (like falling or rolling), and sound synthesis and computer music. The models give sound synthesis and computer music the flexibility to manipulate specific parameters of the sounds. Also, instead of relying on recordings of sounds, virtual reality, computer animation and computer games, can use sound modeling and synthesis to produce new sounds. For example, if the user of a virtual reality environment strikes an object, the synthesis algorithm can use a model of impact sounds to produce the sound of the object being struck.

A struck object produces sound that depends on the way the object vibrates. This sound is determined by physical properties of the object, such as its size, geometry and material, and also by the characteristics of the event, like the force and location of impact. It is possible to derive physical

models of impact sounds given the relationship between the physical and dynamic properties of the object, and the acoustics of the resulting sound (for more details on physical models see, for instance, [Roads, 1996], also examples of physical models will be given in chapter 2). However, physical models are limited because of the *a priori* knowledge they require and because they do not successfully model all the complexities of real sounds.

Natural sounds of the same type have a rich variability in their acoustic structure. For example, different impacts on the same object can generate very different acoustic waveforms. In spite of these variations, when these sounds are heard they are often perceptually very similar, that is, impacts on the same object (or on similar objects) have some common intrinsic structures that listeners can identify. Our goal here is not to study sound perception, but rather to construct a method that learns the common intrinsic structures of similar sounds as well as their variability.

In chapter 2, we present a statistical (and consequently, data-driven) method for learning the intrinsic features that govern the acoustic structure of impact sounds and which may be not easily computable. The method aims to characterize the structures that are common to sounds of the same type as well as their variability. It requires no *a priori* knowledge and it aims for low dimensional characterizations of the sounds. In addition, the method is not restricted to learn an explicit set of properties of the sounds (e.g., basic features such as decay rate and average spectra); instead, it learns the properties that best characterize the statistics of the data. As it will be seen in chapter 2, the method uses the learned features to create mathematical models of the sounds. These mathematical models are able to represent the complexities and natural variability in the structures of the sounds, which is a major advantage over previous knowledge-based models, and represent properties of the sounds such as ringing, resonance, sustain, decay, and sharp onsets.

The method, as is presented in chapter 2, is useful in many fields, like sound recognition and clustering, and the study of the intrinsic structures of impact sounds. In chapter 3, the method is extended so that it can synthesize realistic sounds. An important part of this research is the synthesis of impact sounds using the features learned by the method. The possibility of manipulating the learned features in order to modify the original sounds, or even to create new sounds (for example, from the interpolation of the representation of recorded sounds), and the possibility of synthesizing sounds from only a few features, is of relevance to sound synthesis, virtual reality, multimedia, sound compression, and the study of sound perception, among other fields.

In chapter 4, we discuss some tests and user studies that validate the method. These tests and user studies demonstrate that the sounds modeled and synthesized by our method are realistic, as they are perceived more often as real than as synthesized. Finally, in chapter 5, we summarize the main results of this work.

Chapter 2

Modeling Intrinsic Structures of Impact Sounds

In this chapter we present a statistical data-driven method for learning intrinsic structures of impact sounds. The method applies principal and independent component analysis to learn low-dimensional representations that model the distribution of both the time-varying spectral and amplitude structure. As a result, the method is able to decompose sounds into a small number of underlying features that characterize acoustic properties such as ringing, resonance, sustain, decay, and onsets. The method is highly flexible and makes no *a priori* assumptions about the physics, acoustics, or dynamics of the objects. In addition, by modeling the underlying distribution, the method can capture the natural variability of ensembles of related impact sounds.

2.1 Models of sounds

As mentioned before in chapter 1, models of sounds are useful in many fields, such as sound recognition, identification of events or properties (like material or length) of the objects involved, sound synthesis, virtual reality and computer graphics. Physical models of sounds (that is, mathematical models that describe the mechanical and acoustic properties of the objects) have been proposed in the literature. However, physical models are limited because of the *a priori* knowledge they require and because they do not successfully model all the complexities of real sounds.

The resonance model proposed by Gaver [1988, 1994] consists of a sum of amplitude-decaying

sine waves that model impact sounds:

$$y(t) = \sum_{n=1}^{N} \alpha_n e^{-\delta_n t} \sin(\omega_n t), \tag{2.1}$$

where ω_n is the frequency of partial n, α_n is the initial amplitude of this partial and $e^{-\delta_n t}$ is decay function of the same partial. The values of parameters ω, α and δ can be set from mathematical expressions derived from physics for a limited set of very simple geometries for which the functions of frequency, amplitude and decay are known. It is also possible to deal with more complex geometries by fitting the parameters to recorded sounds, in which case this may be seen as a particular example of Prony's method – Prony's method, which has its roots in Baron de Prony's method for describing the gas expansion behavior in a chemistry experiment [Prony, 1795], consists of a set of methods that decompose the signals into linear combinations of exponentially decaying sinusoidal waves with varying frequency, amplitude, phase and damping factor [Roads, 1996, Kay and Marple Jr., 1981]. In order to set the parameters, Pai and his colleagues create a mesh of the surface of the object, then they record impacts on the object's locations that correspond to the mesh's points of intersection, and fit the parameters to these sounds [Pai, van den Doel, James, Lang, Lloyd, Richmond, and Yau, 2001, van den Doel, Kry, and Pai, 2001].

A limitation of this simplified, knowledge-based model is that it fails to account for the rich structure and variability of real impact sounds. For instance, it fails to model the complex structure of the attack and the variability of sounds resulting from roughness in the surfaces. A solution that uses a stochastic model to simulate the structure of the attack, was proposed by van den Doel et al. [2001] to overcome this problem; however some knowledge about the surfaces of the objects and their contact dynamics is still required. Other physical models have been proposed [e.g. Lambourg, Chaigne, and Matignon, 2001, Avanzini and Rocchesso, 2001a,b], but as with the above-noted models, they require knowledge of the acoustics, as well as the physics, dynamics of contact, and the surface texture of the objects.

In order to obtain a detailed description of the modes of vibration and parameters of objects with complex geometries, some knowledge-based techniques use rigid body simulators developed for computer graphics [O'Brien, Cook, and Ess, 2001, O'Brien, Shen, and Gatchalian, 2002, James, Barbič, and Pai, 2006]. These approaches permit the synthesis of very realistic sounds; however, they are computationally intensive and they require a detailed description of the objects.

A more fundamental limitation of all these approaches, however, is that it is difficult to derive

from natural impact sounds intrinsic acoustic properties beyond those that are explicitly modeled by the equations. For instance, how can a ringing property or a non-exponential decay be modeled by equation 2.1?

This leads to another motivation for this work, which is the extraction of intrinsic features from sounds. Algorithms have been developed to extract basic features of impact sounds, such as the decay rates or the average spectra, but these approaches fail to capture the acoustic richness and variability that is characteristic of natural impact sounds.

Here, we propose a statistical data-driven method for learning the intrinsic features that govern the acoustic structure of impact sounds, and which we call the Intrinsic Structure Analysis (ISA) method (this method was originally presented in [Cavaco and Lewicki, 2007]). The method aims to characterize the structures that are common to sounds of the same type (for instance, if the impacts on the same rod have a ringing property, the method should be able to learn a characterization of this intrinsic structure) as well as their variability (using the same example, the method should also capture the subtle variability of the ringing property in different impacts). At the same time, it aims for low dimensional representations of the sounds. This method requires no *a priori* knowledge and is used to create models of impact sounds that represent a rich variety of structure and variability in the sounds. The method is not restricted to learn an explicit set of properties of the sounds, and it has shown to be able to learn properties such as ringing, resonance, sustain, decay and sharp onsets.

To the best of our knowledge, this is the first statistical (and consequently, data-driven) approach for modeling distributions of impact sounds. While some previous methods [e.g. van den Doel et al., 2001, Pai et al., 2001] use real sounds to fit parameters such as the amplitudes, frequencies and decay rates of the vibration modes, those are actually physical modeling approaches (and not data-driven) because their equations are knowledge based. As it will be seen in section 3.2, other (non-physical) methods that extract their equations' parameters from real sounds have been proposed in the literature. However, these methods are tightly coupled with synthesis and, consequently, their goals are very different from the ISA method's goals. Furthermore, they were usually developed to analyze, modify and synthesize only one sound at a time. On the contrary, the ISA method can deal with ensembles of sounds. Some statistical approaches to model parameters of the sounds have been proposed. For example, Serra and Smith [1990] model the frequencies of the noise portion of the sound with a density function that describes their expected magnitude over time, and Desainte-

Catherine and Hanna [2000] use a statistical approach to extract parameters that describe noise-like sounds. The main difference between these approaches and the ISA method is that the former use statistical approaches to model specific parameters or portions of the sounds, while the ISA method learns the features (structures) that it models, and it can model the distribution of ensembles of sounds.

2.2 Modeling intrinsic structures

Our goal is to learn the intrinsic structure of sounds: we aim to decompose sounds in terms of the set of component signals that best describes them. For convenience, we assume the sounds are initially represented by a spectrogram, $\mathbf{S}$. (Here we will refer to the rows of $\mathbf{S}$, which are the power of frequencies over time, as *frequency bins* or *bins*, and we will refer to the columns of $\mathbf{S}$, each of which is the power spectrum at a given time, as *frames*). Even though our method can be applied to a broader variety of sounds, here we will focus on impact sounds. To illustrate the data, figure 2.1 shows the spectrograms of two impact sounds on an aluminum rod (more details on how these sounds were produced and digitized are given in section 2.3).

Natural sounds of the same type have a rich variability in their acoustic structure. For example, different impacts on the same rod can generate very different acoustic waveforms. In natural environments there is variability due to reverberation and background noise, but even when the sounds are recorded in anechoic conditions there is variability that is due to factors such as the slight variations in the impact force and location (see section 2.3 for details on the recording conditions). Figures 2.1 to 2.3 show that, even though different impacts on the same rod have very similar spectra, the relative power and temporal behavior of the partials varies from one instance to the other. These differences cannot be explained by a simple variation in amplitude of the whole spectrogram.

In spite of these variations, when these sounds are heard they are often perceptually very similar, that is, impacts from similar objects or materials have some common intrinsic structures that listeners can identify. Our goal is not to develop a perceptual model but rather to construct a model that learns the common intrinsic structures of similar sounds as well as their variability.

This section presents two such models. The models discussed here represent the sounds in terms of a set of component signals, in other words, they represent them in a new coordinate system. The form of the basis functions in the new coordinate system depends on the initial representation of the

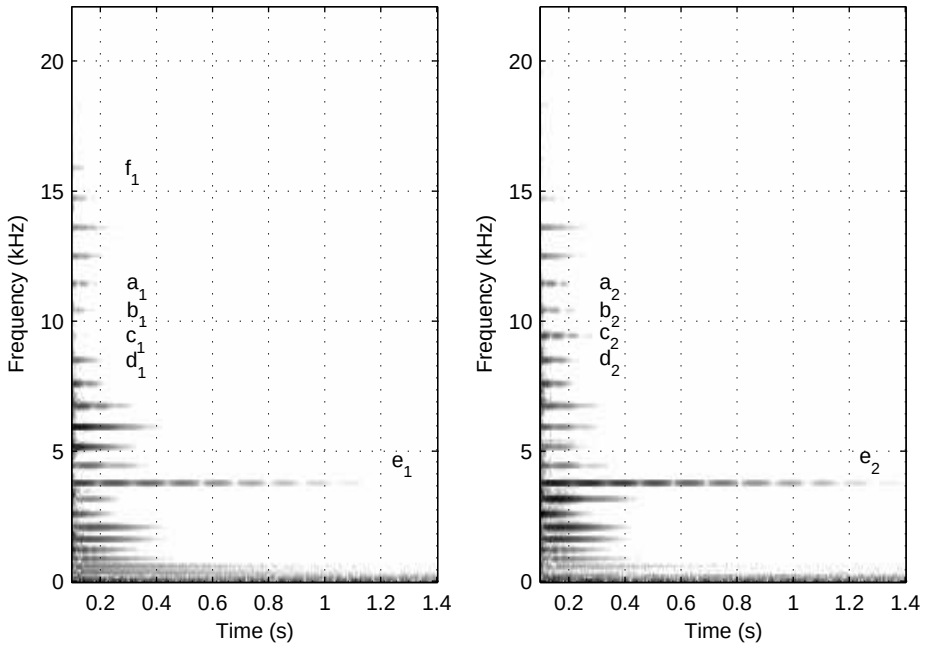


Figure 2.1: Two spectrograms **S** (in decibels) of sounds (Al1 on the left and Al3 on the right) from impacts on an aluminum rod at approximately the same location and with approximately the same force (these spectrograms have been normalized). The relative power and temporal behavior of the partials varies from one instance to the other. For instance, in the left spectrogram, partial $b_1$starts with a lower amplitude than partial a_1, while in the right spectrogram partial b_2 starts with a higher amplitude than a_2. The same happens with partials c and d: c_1 is weaker than d_1, while c_2 is stronger than d_2. Another example is the partial above 15 kHz. In the left spectrogram, this partial, f_1, is stronger than partial c_1, while in the right spectrogram c_2 is the strongest of the two. In fact, in the second spectrogram the partial above 15 kHz does not even appear.

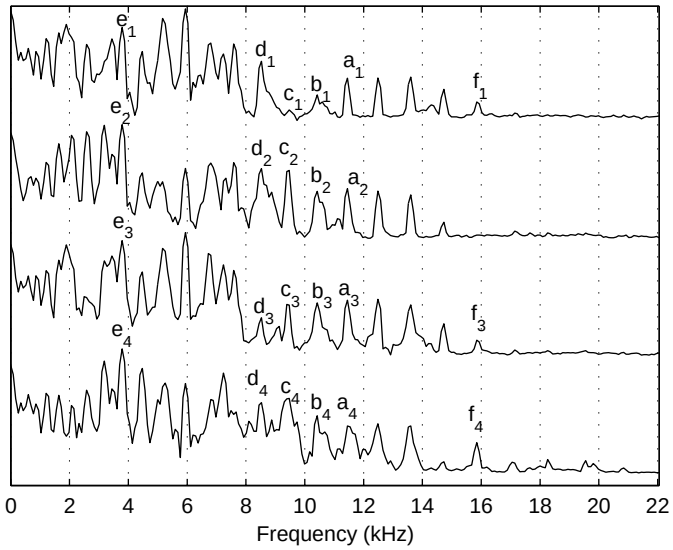


Figure 2.2: Power spectra of four different impacts on aluminum (Al1, Al3, Al10 and Al19 from top to bottom) at approximately the same location and with approximately the same force. The relative power of the partials varies from one impact to another. (The partials are marked with the same labels as in figure 2.1.) Again, it can be seen that the relative powers of partials a, b, c, and d vary in the four power spectra. Also note that partial f appears in the first, third and fourth lines (f_1, f_3, and f_4) but it is absent from the second line. Another interesting feature that can be observed in this figure is how the shape of the power spectrum changes from one sound to the next. For instance, note how partial a_1 is better defined than a_4.

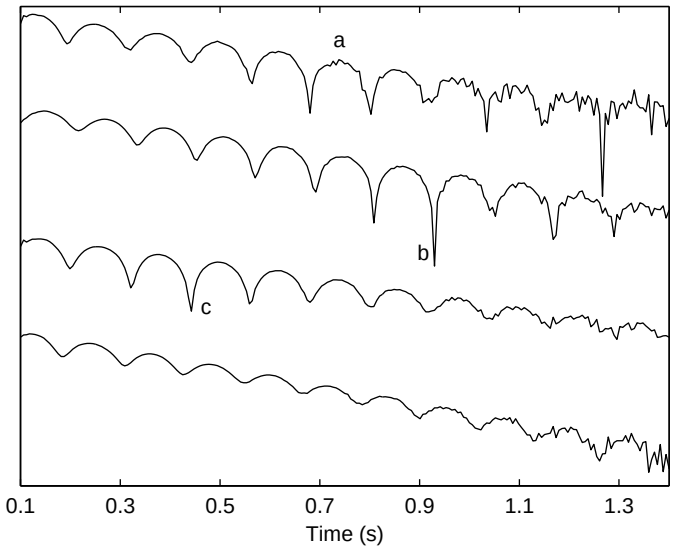


Figure 2.3: The decay shape of the partial at 3.95 kHz in dB (which is partial e in figures 2.1 and 2.2) for four different impacts on an aluminum rod (Al1, Al3, Al10 and Al19 from top to bottom) at approximately the same location and with approximately the same force. The temporal behavior of the partials varies from one impact to another; there is variability in the decay rate and beat pattern of this frequency bin. Note, for instance, the irregularities marked with an a in the first line, and notches b and c in the second and third lines. Also of interest is the consistency of the beating, which suggests that this rod has two close modes of vibration.

data, which here is the spectrogram. The frames are initially represented in an F-dimensional space with one dimension for each frequency bin f; let us call this space the *frequency space*. The bins are initially represented in a T-dimensional space, which we will call *time space*, with one dimension for each time frame t. A *spectral basis function* consists of a vector in the frequency space (that is, a spectra), while a *temporal basis function* consists of a vector in the time space (which can be thought of as a spectra's amplitude envelope). Using spectrograms as the initial representation allows us to model the sounds in spaces defined by spectral and temporal basis functions. (Section 2.5 contains graphical examples of these basis functions.)

Given that the spectrogram $\mathbf{S}$, of size $(F \times T)$, is defined over a discrete set of frequencies, $f \in \{f_1, \ldots, f_F\}$, and a discrete set of time instants, $t \in \{t_1, \ldots, t_T\}$, we can define $\mathbf{S}$ as an ordered set of bins or as a sequence of frames. (Here we use only the power spectrum, and we ignore the phase component.)

The frame model M_r

The first model, which we call the frame model, or M_r, represents the spectrogram $\mathbf{S}$ as a sequence of frames $\mathbf{r}_t$. Here, the frames are modeled as a linear combinations of spectral basis functions $\boldsymbol{\theta}_i$:

$$\mathbf{r}_t = \sum_{i=1}^{I} \boldsymbol{\theta}_i \, p_{i,t} \, . \tag{2.2}$$

where $\boldsymbol{\theta}_i$ is scaled at frame $\mathbf{r}_t$ by coefficient $p_{i,t}$.[1] The value of I depends on the technique used to learn the basis functions $\boldsymbol{\theta}_i$. Here $I \leqslant F$ (see section 2.3 for further details). The basis functions $\boldsymbol{\theta}_i$ describe the spectral regularities in the data set, that is, in $\mathbf{S}$. They can also be used to describe the spectral regularities of a set of related sounds simply by including the appropriate spectrogram frames in the data set. The vectors of coefficients are commonly called *source signals*. Since the vector that consists of the coefficients that are associated with basis function $\boldsymbol{\theta}_i$, that is $\mathbf{p}_i = (p_{i,t_1}, \ldots, p_{i,t_T})$, ranges over the time space, here we call it a *temporal source signal*. (For graphical examples of temporal source signals see section 2.5.3.) Temporal source signal $\mathbf{p}_i$ scales basis function $\boldsymbol{\theta}_i$ across frames.

In order to represent the spectrograms of different sounds with a fixed basis $\boldsymbol{\Theta}$ (where $\boldsymbol{\Theta}$ represents the set of spectral basis functions $\boldsymbol{\theta}_i$) the model requires different temporal source signals

[1] Here matrices are represented in bold upper case, vectors, which are column vectors unless the transpose is used, are represented in bold lower case, and scalars are represented in lower case. The horizontal concatenation of matrices $\mathbf{A}$ and $\mathbf{B}$ is $(\mathbf{A}, \mathbf{B})$.

to scale each basis function $\boldsymbol{\theta}_i$, i.e., there will be one set of temporal source signals for each sound. We distinguish these variables with an upper index k, that is, the set of temporal source signals associated with sound k, which is the set containing $\mathbf{p}_1^k, \ldots, \mathbf{p}_I^k$, is represented by $\mathbf{P}^k$.[2] We can thus rename some of the variables used above to take into account the sound they refer to. Equation 2.2 can thus be rewritten as:

$$\mathbf{r}_t^k = \sum_{i=1}^{I} \boldsymbol{\theta}_i \, p_{i,t}^k \,. \tag{2.3}$$

where $\mathbf{r}_t^k$ is the t^{th} frame of $\mathbf{S}^k$, i.e. the spectrogram of sound k, and the scalar $p_{i,t}^k$ is the t^{th} element of $\mathbf{p}_i^k$.

If we consider all T frames in $\mathbf{S}^k$, equation 2.3 can be rewritten as:

$$\mathbf{S}^k = \boldsymbol{\Theta} \, \mathbf{P}^k \,, \tag{2.4}$$

where the i^{th} column of matrix $\boldsymbol{\Theta}$ contains $\boldsymbol{\theta}_i$, and the i^{th} row of $\mathbf{P}^k$ contains $(\mathbf{p}_i^k)^T$. (See section A.1 in appendix A for figures of the matrices.)

The model, as presented thus far, describes the spectral structure, but not the temporal structure inherent in the temporal source signals $\mathbf{p}_i^k$. We can extend the current model to consider the regularities in the temporal source signals for an ensemble of related sounds. The form of this part of the model is similar to that given above, but now, instead of describing the spectrum of a given frame, it describes the temporal source signals $\mathbf{p}_i^k$. These signals are modeled as a linear combination of basis functions:

$$\mathbf{p}_i^k = \sum_{j=1}^{J} \boldsymbol{\lambda}_j^i \, h_{i,j}^k \,, \tag{2.5}$$

where $\boldsymbol{\lambda}_j^i$ is a temporal basis function and $h_{i,j}^k$ is a scaling coefficient. The temporal basis functions $\boldsymbol{\lambda}_j^i$ describe the temporal regularities in the temporal source signals. Again, the value of J depends on the technique used to learn the basis functions $\boldsymbol{\lambda}_j^i$. Here, $J \leqslant T$.

We can now consider the previous equation at a given time frame t and express $p_{i,t}^k$ as follows:

$$p_{i,t}^k = \sum_{j=1}^{J} \lambda_{j,t}^i \, h_{i,j}^k \,. \tag{2.6}$$

where the scalars $p_{i,t}^k$, and $\lambda_{j,t}^i$ are the values of $\mathbf{p}_i^k$, and $\boldsymbol{\lambda}_j^i$ at time frame t, respectively. (In other words, they are the t^{th} values of vectors $\mathbf{p}_i^k$ and $\boldsymbol{\lambda}_j^i$, respectively.)

[2] We use upper indexes to distinguish different variables of the same type, so for instance $\mathbf{X}^1$ and $\mathbf{X}^2$ are two different matrices of the same type. Lower indices are used to index values within a matrix or vector.

Combining equations 2.3 and 2.6 it follows that the frames of $\mathbf{S}^k$ can be expressed as:

$$\mathbf{r}_t^k = \sum_{i=1}^{I} \sum_{j=1}^{J} \boldsymbol{\theta}_i \, \lambda_{j,t}^i \, h_{i,j}^k \, . \tag{2.7}$$

Thus, $\mathbf{S}^k$ can be modeled by spectral bases $\boldsymbol{\Theta}$, temporal bases $\boldsymbol{\Lambda}$ (where $\boldsymbol{\Lambda}$ contains all temporal basis functions $\boldsymbol{\lambda}_j^i$), and a set of coefficients $\mathbf{H}^k$ (where $\mathbf{H}^k$ contains coefficients $h_{i,j}^k$), that is, $\mathbf{S}^k = M_r(\boldsymbol{\Theta}, \boldsymbol{\Lambda}, \mathbf{H}^k)$. (For further details, see the section A.1 in appendix A.)

The bin model M_b

The second model, which we call the *bin* model, or M_b, expresses the spectrogram $\mathbf{S}^k$ as an ordered set of bins. These are modeled as linear combinations of temporal basis functions $\boldsymbol{\phi}_i$:

$$\mathbf{b}_f^k = \sum_{i=1}^{I} \boldsymbol{\phi}_i \, c_{i,f}^k \, . \tag{2.8}$$

where $\mathbf{b}_f^k$ is the transpose of the f^{th} bin of $\mathbf{S}^k$. $\boldsymbol{\phi}_i$ is scaled at this bin by coefficient $c_{i,f}^k$. As above, the value of I depends on the technique used to learn the basis functions $\boldsymbol{\phi}_i$. Here $I \leqslant T$ (see section 2.3 and appendix B for further details). The basis functions $\boldsymbol{\phi}_i$ describe the temporal regularities in the bins in the data set, which can be composed of one or more spectrograms. (Section 2.5.1 shows how to learn $\boldsymbol{\phi}_i$.) Since the vectors that consists of the coefficients that are associated with basis function $\boldsymbol{\phi}_i$, that is, $\mathbf{c}_i^k = (c_{i,f_1}^k, \ldots, c_{i,f_F}^k)^T$, range over the frequency space, here we call them *spectral source signals*. There is one spectral source signal $\mathbf{c}_i^k$ for each sound k and basis function $\boldsymbol{\phi}_i$, and this spectral source signal scales basis function $\boldsymbol{\phi}_i$ across frequencies. (For graphical examples of spectral source signals see section 2.5.1.)

If we consider all F bins in $\mathbf{S}^k$, equation 2.8 can be rewritten as:

$$(\mathbf{S}^k)^T = \boldsymbol{\Phi} \, \mathbf{C}^k \, , \tag{2.9}$$

where the i^{th} column of matrix $\boldsymbol{\Phi}$ contains $\boldsymbol{\phi}_i$, and the i^{th} row of $\mathbf{C}^k$ contains $(\mathbf{c}_i^k)^T$. (See section A.2 in appendix A for figures of the matrices.)

Thus far, M_b describes the temporal structure, but not the spectral structure inherent in the spectral source signals $\mathbf{c}_i^k$. Just like we have done with model M_r, we can extend M_b to consider the regularities in the spectral source signals for an ensemble of related sounds. Instead of describing the temporal shape of a given bin, this part of the model describes the spectral source signals $\mathbf{c}_i^k$.

These signals are modeled as a linear combination of spectral basis functions $\boldsymbol{\psi}_j^i$:

$$\mathbf{c}_i^k = \sum_{j=1}^{J} \boldsymbol{\psi}_j^i \, v_{i,j}^k \,, \tag{2.10}$$

where the scalar $v_{i,j}^k$ is a scaling coefficient. The spectral basis functions $\boldsymbol{\psi}_j^i$ describe the spectral regularities in the spectral signals. Again, the value of J depends on the technique used to learn the basis functions $\boldsymbol{\psi}_j^i$. Here, $J \leqslant F$ (see appendix B for further details). (Section 2.5.2 shows how to learn $\boldsymbol{\psi}_j^i$.)

We can now consider the previous equation at a given frequency bin f and express $c_{i,f}^k$ as follows:

$$c_{i,f}^k = \sum_{j=1}^{J} \psi_{j,f}^i \, v_{i,j}^k \,. \tag{2.11}$$

where $c_{i,f}^k$, and $\psi_{j,f}^i$ are the values of $\mathbf{c}_i^k$, and $\boldsymbol{\psi}_j^i$ at frequency bin f, respectively. (In other words, they are the f^{th} values of vectors $\mathbf{c}_i^k$ and $\boldsymbol{\psi}_j^i$, respectively.)

Finally, combining equations 2.8 and 2.11 it follows that the bins of $\mathbf{S}^k$ can be expressed as:

$$\mathbf{b}_f^k = \sum_{i=1}^{I} \sum_{j=1}^{J} \boldsymbol{\phi}_i \, \psi_{j,f}^i \, v_{i,j}^k \,. \tag{2.12}$$

This shows that $\mathbf{S}^k$ can be modeled by temporal bases $\boldsymbol{\Phi}$, spectral bases $\boldsymbol{\Psi}$ (where $\boldsymbol{\Psi}$ contains all spectral basis functions $\boldsymbol{\psi}_j^i$), and a set of coefficients $\mathbf{V}^k$ (where $\mathbf{V}^k$ contains coefficients $v_{i,j}^k$), that is, $\mathbf{S}^k = M_b(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \mathbf{V}^k)$. (For more details and figures of the matrices used in this model, see section A.2 in appendix A.)

Summary

Each model is thus defined by two sets of basis functions, and the objective is to find the sets of basis functions with which the data can be better described: ideally only a few basis functions would be needed to accurately describe the data with less redundancy. Having two models to describe the same data raises questions regarding which one should be used and whether one is more advantageous than the other. In section 2.5.3 we will see that, depending on the techniques used to learn the basis functions, one of the models, namely M_b, is more appropriate than the other, in the sense that it may give a better description of the statistics of the data (see section 2.3 for a description of the data and section 2.4 for further details). In section 2.5, we show that the basis functions can be learned effectively by redundancy reduction techniques.

2.3 Methods and techniques

We used a set of impact sounds that were produced using four rods with the same length and diameter, but made of different materials. A wooden rod, with a much shorter length but the same diameter, was used as a mallet. Several impacts on each rod were recorded in an anechoic chamber. The location of the impacts and the impact force varied slightly from one instance to the next, since the rods were hit by hand. The sounds were digitized using a sampling frequency of 44100 Hz.

The spectrograms of the sounds were computed using a 11.6 ms sliding Hanning window. Successive frames overlapped by 5.8 ms. Like with any other system that uses spectrograms, there is a trade off between spectral and temporal resolution. Even though the type of structures obtained for different resolutions is the same, the choice of spectral versus temporal resolution affects the representation: the shapes of the structures obtained differ slightly; for instance a structure that includes a sharp onset can look more or less sharp depending on the resolution. Here, we only report the results obtained using an intermediate resolution of 512-point fast Fourier transform (FFT).

We use principal component analysis (PCA) and independent component analysis (ICA) to learn the sets of basis functions from section 2.2. PCA and ICA are redundancy reduction techniques that look for the axes that best describe the distribution of the data. These techniques are used to represent high dimensional data in a (usually lower dimensional) space with less redundancy. The data are expressed as a linear transformation of the basis functions (i.e., the axes that define the new space). Given a set of M *source signals* of size N (represented by a $(M \times N)$ matrix $\mathbf{Y}$ with one signal per row) mixed into a set of M *signal mixtures* (represented by a $(M \times N)$ matrix $\mathbf{X}$ with one signal mixture per row) PCA and ICA learn a $(M \times M)$ matrix $\mathbf{W}$ that allows extracting the source signals from matrix $\mathbf{X}$:

$$\mathbf{Y} = \mathbf{W}\,\mathbf{X}\,. \tag{2.13}$$

If $\mathbf{A} = \mathbf{W}^{-1}$ this equation can be rewritten as:

$$\mathbf{X} = \mathbf{A}\,\mathbf{Y}\,. \tag{2.14}$$

The two techniques differ importantly in the way they model the distribution of the data, and in their constraints. PCA is a second-order statistical method that assumes a Gaussian distribution and is restricted to orthogonal basis functions (that are the eigenvectors of the data covariance

matrix). This technique decomposes a set of signal mixtures into a set of decorrelated signals and can be used to reduce the dimensionality of the data by considering I basis functions, where $I < M$ (in which case only I source signals are obtained). ICA is a generative model that decomposes a set of signal mixtures into a set of maximally independent source signals. This higher-order statistical method models multivariate data with non-Gaussian distributions and is not restricted to orthogonal basis functions. ICA contains PCA as a special case when the marginal distributions of signals are assumed to be Gaussian and the bases are restricted to be orthogonal. (For more details on ICA and PCA, see [Hyvärinen, Karhunen, and Oja, 2001] or [Stone, 2004].)

For instance, in the case of the first part of M_b and when we consider K impact sounds, matrix $\mathbf{X}$ consists of the horizontal concatenation of transposed spectrograms $((\mathbf{S}^1)^T, (\mathbf{S}^2)^T, \ldots, (\mathbf{S}^K)^T)$, $\mathbf{A}$ is the spectral basis $\mathbf{\Phi}$, and $\mathbf{Y}$ is the horizontal concatenation of the matrices of spectral source signals $(\mathbf{C}^1, \mathbf{C}^2, \ldots, \mathbf{C}^K)$. A signal mixture is the concatenation of one transposed frame from each of the K spectrograms, and there are T signal mixtures. Therefore, $I \leqslant T$ in equations 2.8 to 2.12.

We used a built-in function from MATLAB to do PCA and the Fast ICA software package by Hyvärinen et al. [2001] to do ICA (for more details see appendix B). Because both PCA and ICA model the variation around the data mean, we used both the data matrix and its negative, i.e., we used the *extended matrix* $(-\mathbf{S}^T, \mathbf{S}^T)$, so that the mean would be zero. This was done so that the model describes the signal rising to and falling from zero, rather than the spectrogram mean.

2.4 Temporal and spectral analysis

ICA and PCA can be used to model the structure in the spectrogram frames or the structure in the spectrogram bins. In this section, we describe the differences between these two types of analysis.

Temporal analysis considers the frequency bins of $\mathbf{S}^k$ as temporal signal mixtures, i.e., the time varying signal in each frequency bin is considered to be a linear combination of independent (for ICA) or uncorrelated (for PCA) temporal source signals. The goal of temporal analysis is to decompose $\mathbf{S}^k$ into this set of temporal source signals. For instance, the method does a temporal analysis on matrix $(\mathbf{S}^1, \mathbf{S}^2, \ldots, \mathbf{S}^K)$ to learn the spectral basis functions $\mathbf{\Theta}$ of model M_r and find the sets of temporal source signals $\mathbf{P}^1, \mathbf{P}^2, \ldots, \mathbf{P}^K$. In other words, $\mathbf{X}$ from equation 2.14 is equal to $(\mathbf{S}^1, \mathbf{S}^2, \ldots, \mathbf{S}^K)$, $\mathbf{Y}$ is $(\mathbf{P}^1, \mathbf{P}^2, \ldots, \mathbf{P}^K)$, and $\mathbf{A}$ is $\mathbf{\Theta}$. Figure 2.4a illustrates this type of analysis.

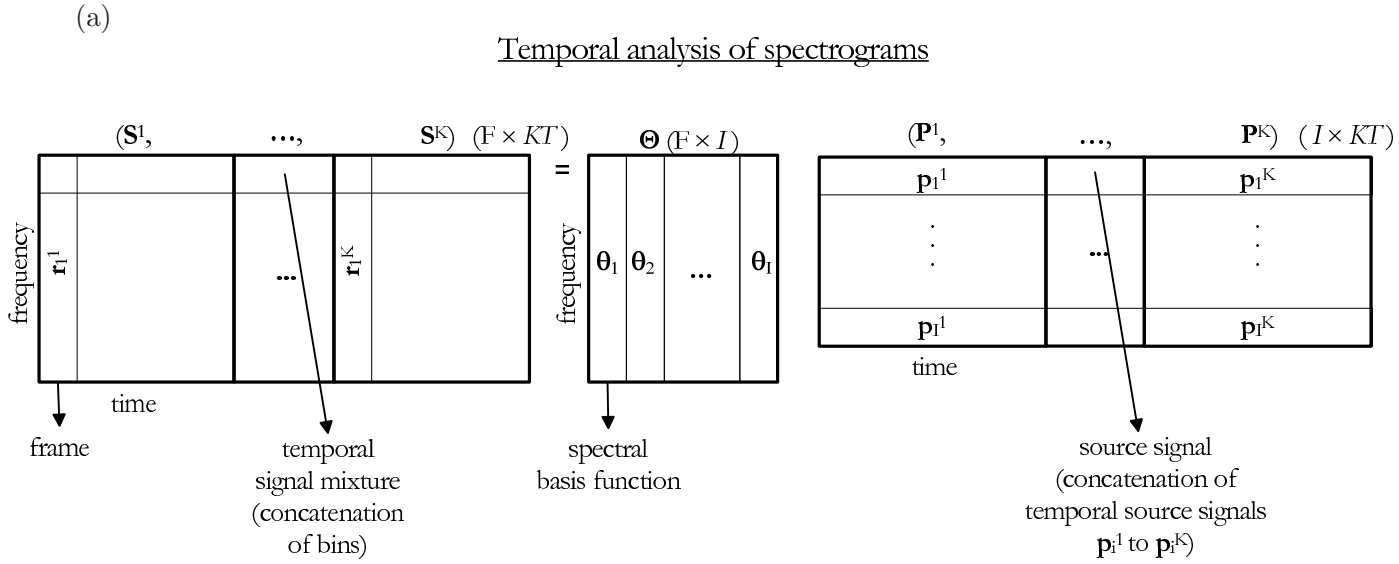
Spectral analysis considers the frames (or power spectra) of $\mathbf{S}^k$ as spectral signal mixtures, i.e., the spectral signal in each frame is considered to be a linear combination of independent or

uncorrelated spectral source signals (for ICA and PCA, respectively). Here the goal is to decompose $\mathbf{S}^k$ into this set of spectral source signals. For instance, in order to learn the temporal basis functions $\mathbf{\Phi}$ of model M_b and find the sets of spectral source signals $\mathbf{C}^1, \mathbf{C}^2, \ldots, \mathbf{C}^K$, the method does a spectral analysis on matrix $((\mathbf{S}^1)^T, (\mathbf{S}^2)^T, \ldots, (\mathbf{S}^K)^T)$, where $(\mathbf{S}^1)^T$ to $(\mathbf{S}^K)^T$ are time aligned, so that this data matrix has one row (i.e., transposed frame) that corresponds to the start of all K impacts. In other words, $\mathbf{X}$ from equation 2.14 is equal to $((\mathbf{S}^1)^T, (\mathbf{S}^2)^T, \ldots, (\mathbf{S}^K)^T)$, $\mathbf{Y}$ is $(\mathbf{C}^1, \mathbf{C}^2, \ldots, \mathbf{C}^K)$, and $\mathbf{A}$ is $\mathbf{\Phi}$. Figure 2.4b illustrates this type of analysis.

Considering the case of one sound, temporal PCA and spectral PCA give equivalent results, where the role of source signals and basis functions switch: when we consider a subset of I basis functions from the set of F spectral basis functions and a subset of I basis functions from the set of T temporal basis functions, learned by temporal or spectral PCA, respectively, the two types of analysis give equivalent results. Temporal PCA learns a set of spectral basis functions represented by an $(F \times I)$ matrix $\mathbf{A}^t$ (with $I \leqslant F$) and it finds the corresponding temporal source signals represented by an $(I \times T)$ matrix $\mathbf{Y}^t$, while spectral PCA learns a set of temporal basis functions represented by an $(T \times I)$ matrix $\mathbf{A}^s$ (with $I \leqslant T$) and it finds the corresponding spectral source signals represented by an $(I \times F)$ matrix $\mathbf{Y}^s$. The spectral basis functions learned by temporal PCA are the spectral source signals found by spectral PCA, that is, $\mathbf{A}^t = (\mathbf{Y}^s)^T$, and the temporal basis functions learned by spectral PCA are the temporal source signals found by temporal PCA, that is, $\mathbf{A}^s = (\mathbf{Y}^t)^T$.

Therefore, considering the case of one sound only and without extending the matrices (as discussed at the end of section 2.3), using temporal PCA to learn the spectral basis functions $\mathbf{\Theta}$ and find temporal source signals $\mathbf{P}^k$ is equivalent to using spectral PCA to learn the temporal basis functions $\mathbf{\Phi}$ and find spectral source signals $\mathbf{C}^k$. If the signal mixture matrices consist of the extended matrices $\mathbf{X}_1 = (-\mathbf{S}, \mathbf{S})$ for temporal analysis, and $\mathbf{X}_2 = (-\mathbf{S}^T, \mathbf{S}^T)$ for spectral analysis, in theory the results are different because since $\mathbf{X}_1$ and $\mathbf{X}_2$ have different sizes, $(F \times 2T)$ and $(T \times 2F)$, respectively, the results consist of matrices of different sizes. However, if we ignore the results due to the negative parts in the extended matrices, that is, if we ignore that the results include $-\mathbf{P}^k$ and $-\mathbf{C}^k$, we can consider them equivalent because $\mathbf{\Theta} = (\mathbf{C}^k)^T$ and $\mathbf{\Phi} = (\mathbf{P}^k)^T$. (With more than one sound the results would not be equivalent because the signal mixture matrices would have different sizes, due to the spectrograms being concatenated in different ways for temporal and spectral analysis.)

(a)



(b)

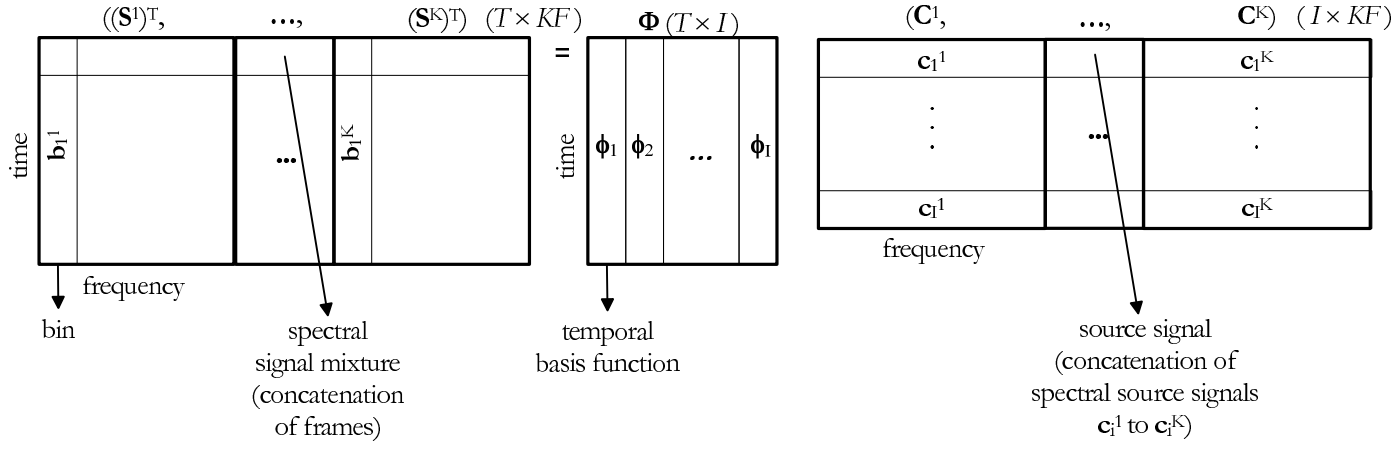


Figure 2.4: Analysis of spectrograms. (a) Temporal analysis considers the bins of the spectrograms as temporal signal mixtures and expresses the spectrograms as a sequence of frames as in equation 2.3. (b) Spectral analysis considers the frames of the spectrograms as spectral signal mixtures and expresses the spectrograms as an ordered set of bins as in equation 2.8.

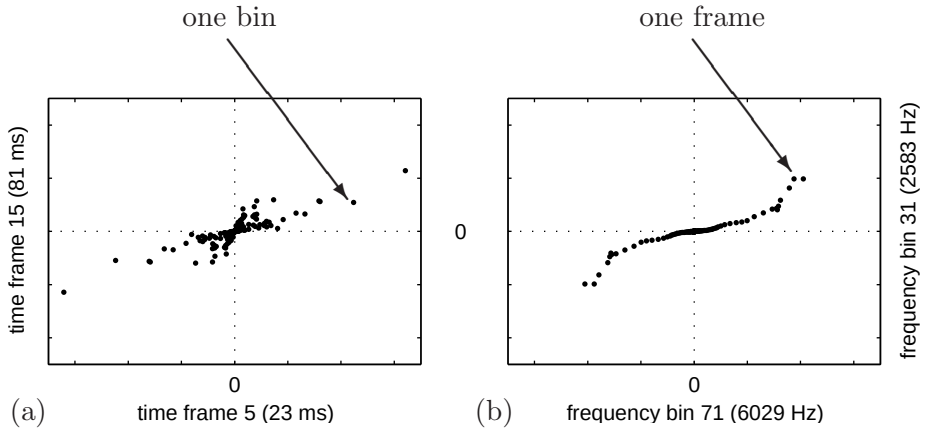


Figure 2.5: Scatter plots of frames and bins from a sound from an aluminum rod (Al1). Frames and bins with high energy were chosen in both plots to avoid having all the points falling into the neighborhood of zero. The points in the third quadrant of both plots are due to the negative part of the extended matrices (see the end of section 2.3 for details). (a) This scatter plot shows the values of time frame 5 (23 ms) plotted against the values of time frame 15 (81 ms), where each point in the plot corresponds to a different frequency bin. (These two frames have high energy in some frequency bins, which correspond to points far from zero in the plot, and low energy in other frequency bins, which correspond to points close to zero. Since frame 5 occurs earlier in the impact sound, it has more energy than frame 15.) (b) This scatter plot shows the values of frequency bin 71 (6 kHz) plotted against the values of frequency bin 31 (2.6 kHz), where each point in the plot corresponds to a different time frame. (These bins have high energy in the first time slots, corresponding to the points in the upper right or lower left, and then the energy decreases until they reach zero.)

The same does not happen with temporal and spectral ICA. ICA looks for correlations in the joint statistics of the data: spectral ICA looks for correlations (or structures) across the time frames, while temporal ICA looks for correlations across the frequency bins. Figure 2.5a shows the type of distributions that have to be modeled by spectral ICA. Each point in this figure corresponds to a different bin. A bin is a T-dimensional vector, with one dimension per time frame (in other words, a bin is a point in the time space). This figure shows the projection of the bins onto the plane defined by two dimensions (corresponding to time frames 5 and 15). Spectral ICA looks for axes that describe this data with less redundancy than the original axes (which correspond to the time

frames). Those axes are the temporal basis functions $\mathbf{\Phi}$, which represent the temporal structures in the data. Figure 2.5b shows the type of distributions that have to be modeled by temporal ICA. Each point in this figure corresponds to a different frame. A frame is an F-dimensional vector, with one dimension per frequency bin (in other words, a frame is a point in the frequency space). This figure shows the projection of the frames onto the plane defined by two dimensions (corresponding to frequency bins 31 and 71). Temporal ICA looks for axes that describe this data with less redundancy than the original axes (which correspond to the frequency bins). Those axes are the spectral basis functions $\mathbf{\Theta}$, that represent the spectral structures in the data. Due to ICA's underlying assumptions of the statistical model and to the differences in the joint distribution of frames and bins, the results obtained by temporal ICA and spectral ICA are not equivalent.

The type of distribution in figure 2.5a approximates the distribution assumed by ICA better than the type of distribution in figure 2.5b. Since spectral ICA looks for correlations across the time frames, it is able to match the statistics of the data better than temporal analysis. Therefore, one can expect spectral ICA to lead to better results than temporal ICA. Sections 2.5.1 and 2.5.3 explore these two types of analysis further, and show that the results obtained by spectral ICA are smother and more easily interpretable than the results of temporal ICA, which are noisier and not as easily interpretable.

2.5 Results

In this section, we show how to learn representations of the intrinsic structures of impact sounds. We show that the method developed in section 2.2 can be used to characterize the structures of a single sound or the structures of sets of related sounds. In the latter case, the method learns representations of the structures that are common to the set of sounds and models their natural variability.

Section 2.5.1 explores the first part of model M_b, which is characterized by the temporal basis functions $\mathbf{\Phi}$, while section 2.5.2 explores the second part of this model, which is characterized by the spectral basis functions $\mathbf{\Psi}$. The results from model M_r are shown in section 2.5.3, where the spectral basis functions $\mathbf{\Theta}$ are explored. It will be shown that the results obtained by model M_b are more appropriate to describe the structure of the spectrograms than the results from model M_r. Finally, section 2.5.4 illustrates how the natural variability of related sounds is represented by model M_b.

2.5.1 Temporal basis functions $\mathbf{\Phi}$

As seen in section 2.4, there are two ways of applying ICA and PCA to spectrograms: these techniques can be used to do a *spectral analysis* of $\mathbf{S}^k$, in which the signal mixtures and source signals are considered to be spectra, or a *temporal analysis* of $\mathbf{S}^k$, in which the signal mixtures and source signals are considered to be temporal signals. In order to learn the set of temporal basis functions $\mathbf{\Phi}$ and decompose the spectrograms into sets of spectral source signals, we apply *spectral* PCA and ICA to the spectrogram of a single impact or to the spectrograms of different impacts on the same rod. The temporal basis functions $\mathbf{\Phi}$ are time varying functions that represent temporal properties of different sub-spectra of the sounds. Each spectral source signal ($\mathbf{c}_i^k$) is associated with a particular temporal basis function ($\boldsymbol{\phi}_i$) that represents a component of the signal's temporal behavior.

For instance, in order to learn the temporal basis functions $\mathbf{\Phi}$ from model M_b and find the sets of spectral source signals $\mathbf{C}^1, \mathbf{C}^2, \ldots, \mathbf{C}^K$ for K sounds, the method does a spectral analysis on matrix $((\mathbf{S}^1)^T, (\mathbf{S}^2)^T, \ldots, (\mathbf{S}^K)^T)$, where $(\mathbf{S}^1)^T$ to $(\mathbf{S}^K)^T$ are time aligned, so that the matrix has one row (or transposed frame) that corresponds to the start of all K impacts. In order to align the spectrograms so that all impacts have the same amount of silence, or background noise, before they start, the method inspects the energy in the frames (that is, the overall power of all frequencies in each frame). More specifically, for each spectrogram the method looks for the first local maximum of the overall power that is above a certain threshold and that is preceded by silence or background noise, and aligns all spectrograms according to these local maxima.

One impact sound

We start with the spectrogram $\mathbf{S}$ of a single sound. Figures 2.6a and 2.6b show 6 out of the 10 most dominant basis functions (i.e., $\boldsymbol{\phi}_1$ to $\boldsymbol{\phi}_{10}$) learned by ICA.[3] As can be seen, ICA is able to isolate temporal properties of the sound: see for instance $\boldsymbol{\phi}_b$ in figure 2.6a, which represents a ringing property of the sound, $\boldsymbol{\phi}_d$ in the same figure, which represent a decay property of the sound, $\boldsymbol{\phi}_a$ in the same figure and $\boldsymbol{\phi}_e$ in figure 2.6b, which represent sustain properties, and the sharp basis functions like $\boldsymbol{\phi}_c$ in figure 2.6a and $\boldsymbol{\phi}_a$ and $\boldsymbol{\phi}_d$ in figure 2.6b, which are related to impact (i.e., attack) properties of the sounds.

[3] In order to make the graphs more readable, some of the basis functions $\boldsymbol{\phi}_i$ and corresponding spectral source signals $\mathbf{c}_i^k$ have been multiplied by -1.

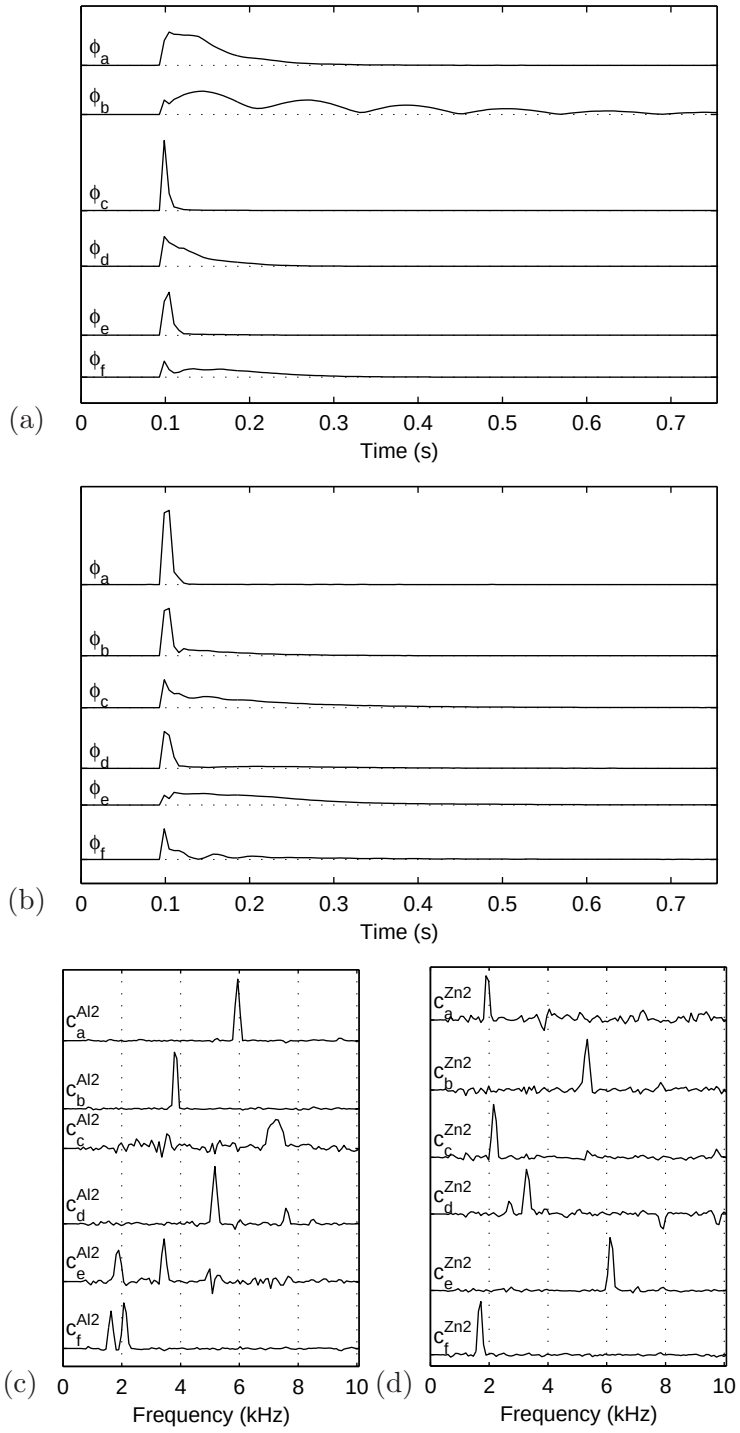


Figure 2.6: Temporal basis functions $\mathbf{\Phi}$ learned by ICA of the spectrogram of: (a) a sound (Al2) from an impact on an aluminum rod; (b) a sound (Zn2) from an impact on a zinc plated steel rod. In each case, 6 out of the 10 most dominant basis functions are shown in decreasing order of dominance from top to bottom. The corresponding spectral source signals for Al2 (c) and Zn2 (d) are shown also from top to bottom.

While ICA can model the data using non-orthogonal basis functions, PCA models the data with orthogonal bases. Consequently, the temporal basis functions learned by PCA can differ from those learned by ICA. Figure 2.7 illustrates the results obtained by PCA of the spectrogram of the sound of an impact on an aluminum rod. This figure shows that the dominant basis function, $\boldsymbol{\phi}_1$, has a much smoother shape than the other basis functions. This basis function shapes the overall decay of all partials. In fact, the results show that PCA extracts a dominant basis function $\boldsymbol{\phi}_1$ that represents most of the temporal structure of the sound (figure 2.8). On average this basis function accounts for more than 68% of the temporal variation in $\mathbf{S}$. This property of the dominant basis function is due to the lack of variation in the spectral structure of the sound over time. (As an example of this regularity, figure 2.1 shows that there is not much variation in which partials are active over time.) $\boldsymbol{\phi}_1$ has the ability to account for the temporal behavior of this spectral structure. To illustrate this point, figure 2.9 shows the average power spectrum of a sound from an impact on an aluminum rod and spectral source signal $\mathbf{c}_1^{Al2}$ obtained by PCA of the spectrogram of this sound. $\boldsymbol{\phi}_1$ describes the temporal behavior of spectra $\mathbf{c}_1^{Al2}$, which, as can be seen in this figure, is very similar to the sound's power spectrum, which represents the spectral structure of the sound over time.

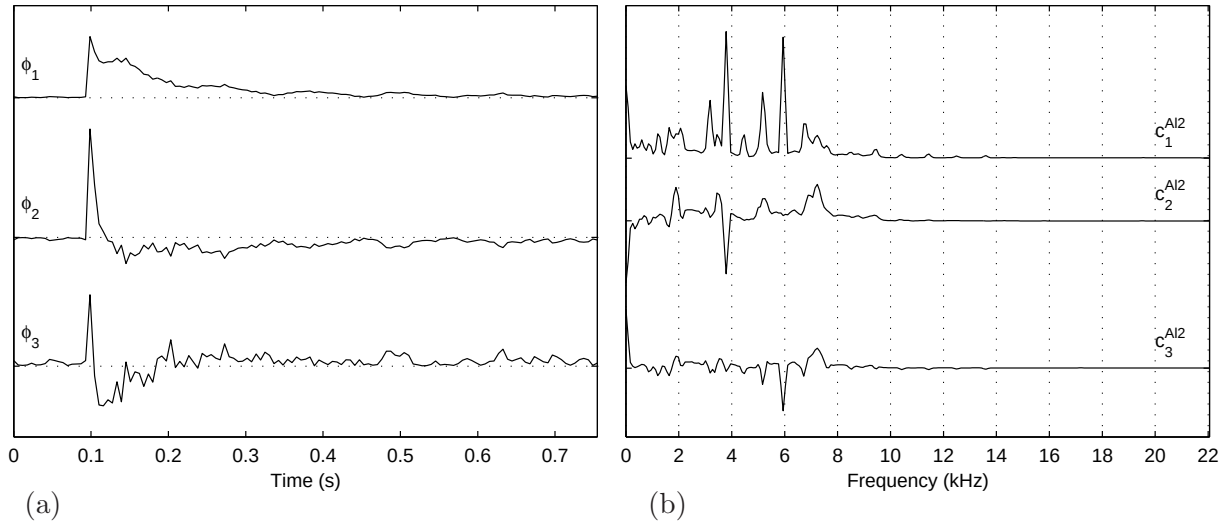


Figure 2.7: Temporal basis functions $\boldsymbol{\Phi}$ and spectral source signals ($\mathbf{C}^{Al2}$) obtained by PCA of the spectrogram of a single sound (Al2) from an impact on an aluminum rod. (a) The first three basis functions are shown from top to bottom. (b) The first three spectral source signals are shown from top to bottom.

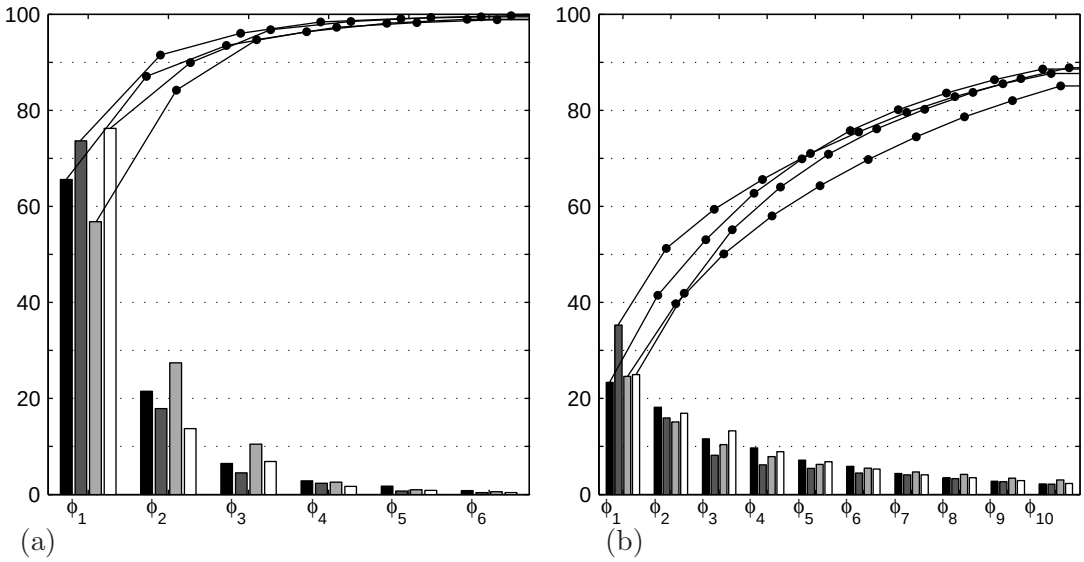


Figure 2.8: Percentage of variance explained by the basis functions in $\mathbf{\Phi}$. The spectrograms from ten impact sounds from each rod (aluminum in black, zinc plated steel in dark grey, steel in light grey, and wood in white) were used. $\mathbf{\Phi}$ was learned by spectral analysis on one spectrogram at a time. The 10 results obtained for each rod were averaged. Only the values for the first six or ten temporal basis functions are shown. The dots on the curves show the cumulative sums of the percentages. In (a) $\mathbf{\Phi}$ was learned by PCA. In (b) $\mathbf{\Phi}$ was learned by ICA.

Other less significant basis functions account for temporal behaviors that differ from the overall decay shape described by $\boldsymbol{\phi}_1$. For example, the temporal shapes of $\boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_3$ account for variations in the temporal behavior of sub-spectra $\mathbf{c}_2^{Al2}$ and $\mathbf{c}_3^{Al2}$ (figure 2.7). (Note also that these sub-spectra contain common partials with the spectral structure of the sound, but, as it can be easily seen in this figure, they account for much less of the spectral structure of $\mathbf{S}$ than $\mathbf{c}_1^{Al2}$ does. The same is true for other sounds. The less variance a basis function accounts for, the fewer partials its spectral source signal shares with $\mathbf{S}$.) In contrast to what was seen with ICA, these basis functions are not as directly related to temporal properties of the sounds. (Note that since the same sound, Al2, was used in both figure 2.6a and figure 2.7, these are directly comparable.)

As seen earlier, ICA obtains a greater variety of basis function shapes: some are similar to the most significant PCA basis functions, but ICA is also able to learn basis functions that capture structures besides decay. In fact, there seems to be a more direct relation between the shape of the basis functions learned by ICA and temporal properties like ringing, resonance, decay, impact

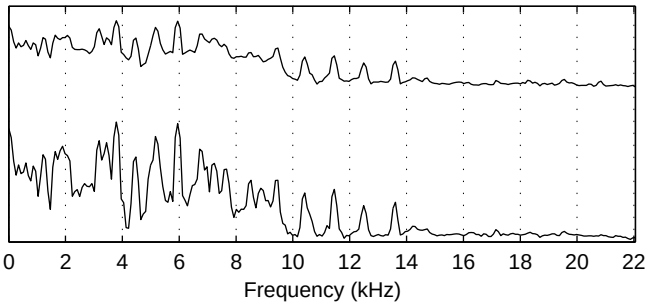


Figure 2.9: Power spectrum of a sound (Al2) from an impact on an aluminum rod. The bottom line shows the power spectrum. The line on the top shows spectral source signal $\mathbf{c}_1^{Al2}$ found by PCA of the spectrogram of this sound. Note how both lines show high energy on the same partials. (Here, the source signal $\mathbf{c}_1^{Al2}$ looks different than in figure 2.7 because it is plotted in a logarithmic scale.)

(or attack), etc. As a consequence, ICA needs more basis functions to explain the variance of $\mathbf{S}$ (figure 2.8). On average, the most significant basis function ($\boldsymbol{\phi}_1$) accounts only for about 27% of the temporal variation in $\mathbf{S}$ compared to 68% for PCA.

Up to this point we have considered the basis functions, now we will consider the spectral source signals. Because here we consider the spectrogram $\mathbf{S}$ of a single sound, there is only one spectral source signal $\mathbf{c}_i^k$ associated with each basis function $\boldsymbol{\phi}_i$. This source signal consists of the partials that have the time varying shape described by $\boldsymbol{\phi}_i$. In other words, the source signals consist of partials that have similar time varying shape. Unlike the source signal of the dominant basis function obtained by PCA, with ICA there is no source signal that accounts for most of the spectral structure in $\mathbf{S}$. ICA separates partials with different time varying shapes into different spectral source signals, which is better suited to represent the variability in the sounds. This point is illustrated by figures 2.6c and 2.6d, which show the spectral source signals for 6 out of the 10 most dominant basis functions obtained by spectral ICA. As can be seen, when ICA is used, the partials in one spectral source signal are typically not present in the remaining source signals. From another perspective, ICA learns basis functions that more directly relate to the underlying acoustic properties. This desirable effect allows ICA to extract more interesting temporal structures of the sounds than those seen with PCA.

Ensemble of impact sounds

We will now consider the more general case of an ensemble of impacts on the same rod. In this case, the data matrix is defined over a set of K sounds aligned at time zero. The result of applying spectral ICA or PCA to this data is a set of temporal basis functions $\mathbf{\Phi}$ and K sets of spectral source signals $\mathbf{C}^k$. The temporal basis functions $\mathbf{\Phi}$ model the common temporal properties of the sounds, and each set of spectral source signals $\mathbf{C}^k$ represents the spectra of sound k that have the temporal properties described by $\mathbf{\Phi}$. The spectral source signals (say $\mathbf{c}_i^{k_1}$ and $\mathbf{c}_i^{k_2}$) associated with the same basis function $\boldsymbol{\phi}_i$ are the sub-spectra (of sounds k_1 and k_2, respectively) that share the temporal property described by $\boldsymbol{\phi}_i$.

The results for multiple impacts resemble those for a single impact due to the similarity in the underlying acoustic structure across impacts. This is clear with the basis functions learned by ICA, for instance, compare $\boldsymbol{\phi}_a$ in figures 2.6a and 2.10a, and is particularly obvious with the most dominant basis function learned by PCA, for instance, compare the first line from figures 2.7a and 2.11a. Even though the temporal basis functions in these figures are not exactly the same, they have very similar shapes.

Because more impacts on the same rod imply more variability, some acoustic structures that were represented by a single basis function above (for the one sound case) are now represented by multiple basis functions. For example, the ringing structure represented by $\boldsymbol{\phi}_b$ in figure 2.6a is now represented by both $\boldsymbol{\phi}_c$ and $\boldsymbol{\phi}_e$ in figure 2.10a. In order to illustrate how the temporal variability is represented, we will examine these two basis functions more carefully. By inspecting the spectral source signals that correspond to $\boldsymbol{\phi}_c$ and $\boldsymbol{\phi}_e$ (see second and third plots in the middle column of figure 2.12) we can conclude that these two basis functions represent the temporal behavior of the partial at 3.95 kHz. In some impacts this partial has a temporal shape that is more closely described by $\boldsymbol{\phi}_c$ (observe that for Al1 there is a peak in $\mathbf{c}_c^{Al1}$ but not in $\mathbf{c}_e^{Al1}$), while in other impacts the partial's temporal shape is more closely described by $\boldsymbol{\phi}_e$ (for Al3 there is a peak in $\mathbf{c}_e^{Al3}$ but not in $\mathbf{c}_c^{Al3}$). Still in other impacts a mixture of both $\boldsymbol{\phi}_c$ and $\boldsymbol{\phi}_e$ is required to describe the partial's temporal shape (Al2 has a peak in both $\mathbf{c}_e^{Al2}$ and $\mathbf{c}_c^{Al2}$).

Even though on average the basis functions here account for a smaller percentage of variance than for the one sound case and more basis functions are needed to explain the same percentage of variance, the difference is not significant. For instance, the 10 most dominant basis functions learned by spectral ICA of a single sound account for at most 88% of the variance, while when a

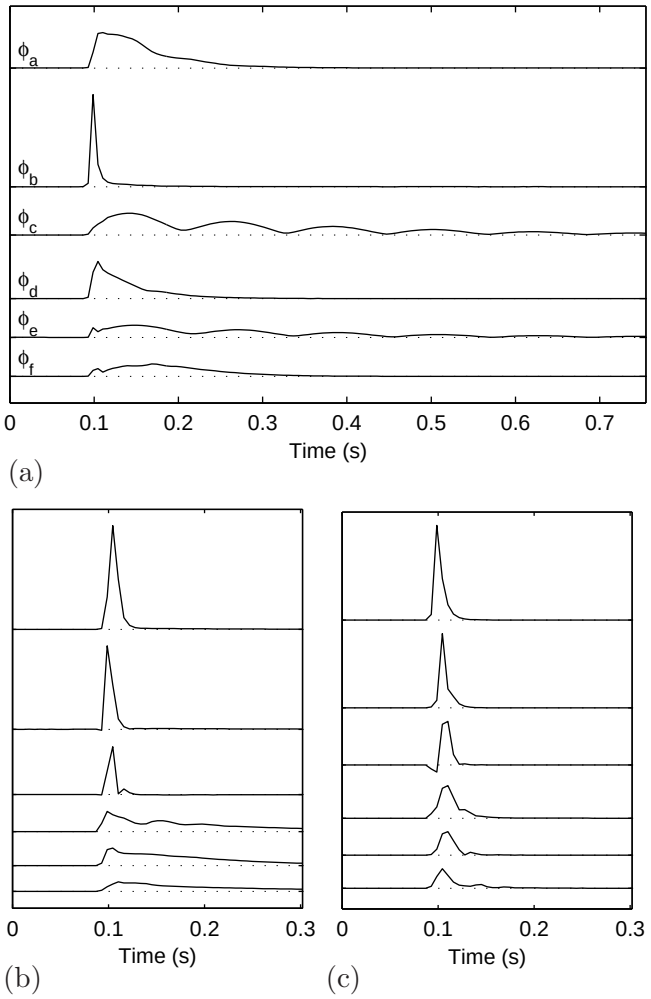


Figure 2.10: Temporal basis functions $\mathbf{\Phi}$ learned by ICA of the set of: (a) ten sounds from impacts on an aluminum rod (the spectral source signals, $\mathbf{C}^k$, for these basis functions are shown in figure C.1 in appendix C); (b) ten sounds from impacts on a zinc plated steel rod; (c) ten sounds from impacts on a wooden rod. In each figure, 6 out of the 10 most dominant basis functions are shown in decreasing order of dominance from top to bottom.

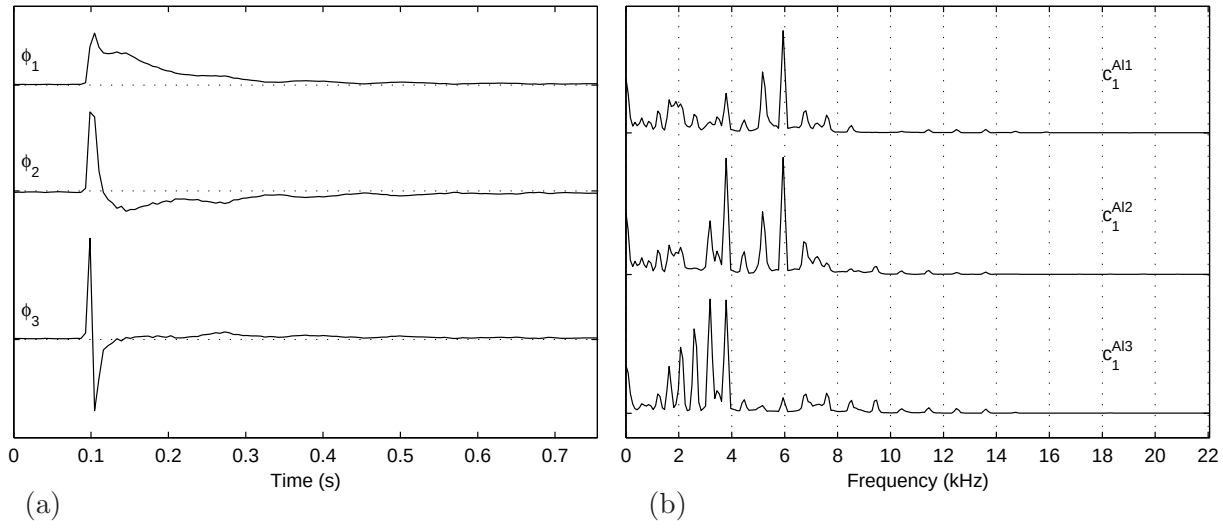


Figure 2.11: Temporal basis functions $\mathbf{\Phi}$ and spectral source signals $\mathbf{C}^k$ obtained by PCA of the set of 10 impacts on an aluminum rod. (a) The first three basis functions are shown from top to bottom. (b) The first spectral source signal for sounds Al1, Al2 and Al3. (Spectral source signals $\mathbf{c}_2^k$ and $\mathbf{c}_3^k$ are shown in figure C.2 in appendix C.)

set of 10 sounds is used, the same number of basis functions explains at most 84% of the variance of the data.[4] Spectral PCA shows similar results: 6 basis functions suffice to explain around 99% of the variance on a single sound, while for a set of 10 sounds, 6 basis functions can explain around 96% of the variance (figures 2.8 and 2.13).

The results shown here were obtained from sounds recorded in an anechoic chamber, however we also tested the model with sounds recorded in a normal room (with background noise and reverberation). In this case, $\mathbf{S}$ represented not only the structure of the sound, but also the structure of the background noise. Consequently, apart from the temporal basis functions that accounted for the temporal structure of the sound, spectral PCA and ICA also learned some basis functions that described the temporal structure of the background noise (data not shown).

The results are dependent on the sounds analyzed. Figure 2.10 shows that impacts on different rods are characterized by different basis functions. For instance, some of the basis functions that

[4]Since the basis functions given by PCA are orthogonal, the sum of the variances that they explain gives the total variance explained. However, the same is not true for the basis functions given by ICA, which are not restricted to being orthogonal. In this case, the sum of the variances may correspond to a quantity that is bigger than the actual variance explained by the basis functions.

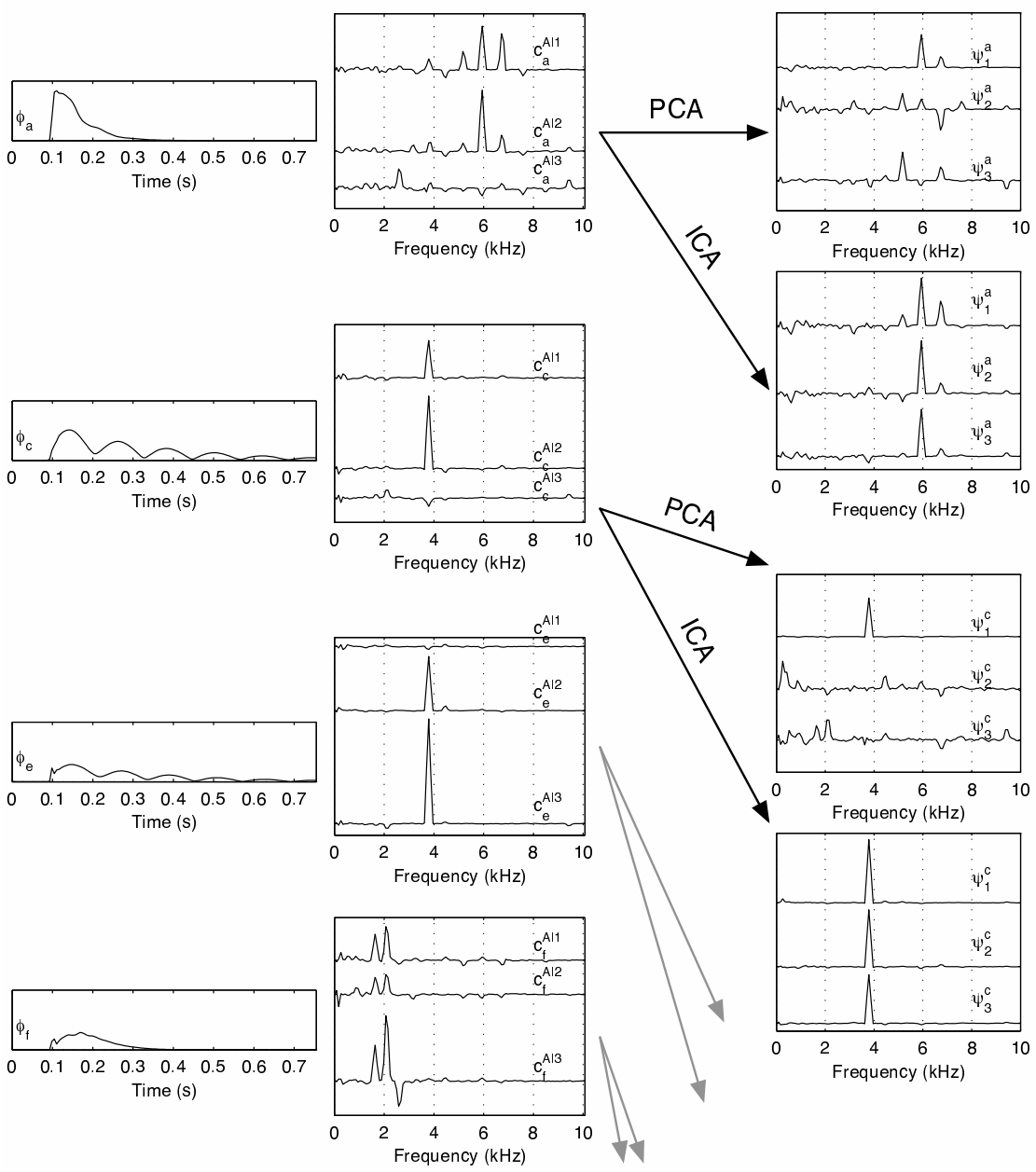


Figure 2.12: (Left column) Temporal basis functions ϕ_a, ϕ_c, ϕ_e and ϕ_f from figure 2.10a. These are learned by ICA of the set of 10 sounds from impacts on an aluminum rod. (Middle column) The corresponding spectral source signals for sounds Al1, Al2 and Al3. (Right column) Spectral basis functions $\mathbf{\Psi}$ obtained by analysis of the spectral source signals. The first and third figures in this column show the first three spectral basis functions from $\mathbf{\Psi}^a$ and $\mathbf{\Psi}^c$ learned by PCA. The second and fourth figures in this column show the first three spectral basis functions from $\mathbf{\Psi}^a$ and $\mathbf{\Psi}^c$ learned by ICA.

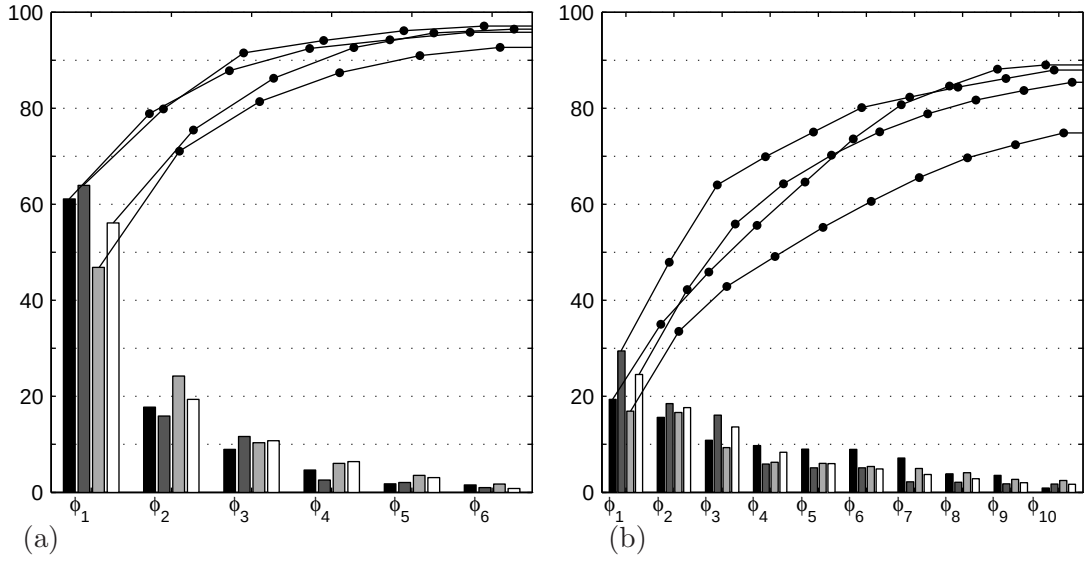


Figure 2.13: Percentage of variance explained by the basis functions in $\mathbf{\Phi}$ learned by spectral analysis on the set of 10 impacts on an aluminum rod (black), the set of 10 impacts on a zinc plated steel rod (dark grey), the set of 10 impacts on a steel rod (light grey), and the set of 10 impacts on a wooden rod (white). Only the values for the first six or ten temporal basis functions are shown. The dots on the curves show the cumulative sums of the percentages. In (a) $\mathbf{\Phi}$ was learned by PCA. In (b) $\mathbf{\Phi}$ was learned by ICA.

characterize impacts on aluminum have a longer duration than the basis functions that characterize impacts on wood. If sounds with different characteristics are used, the basis functions will reflect those characteristics.

2.5.2 Spectral basis functions Ψ

The sets of spectral source signals $\mathbf{C}^1,\ldots,\mathbf{C}^K$ represent the sub-spectra associated with the temporal basis functions in $\mathbf{\Phi}$. Even though each $\mathbf{C}^k$ is specific to an individual sound k, the sets of source signals do share common structures. This can be seen in figure 2.12. The middle column shows the spectral source signals obtained by spectral ICA of the set of 10 sounds from an aluminum rod. Although the source signals show considerable variability there is still much common structure. The same observations can be made on the results from spectral PCA (see figure 2.11b and figure 2.14b).

As explained in section 2.2, we can extend the approach to model the regularities in the spectral

source signals. In the extended model, these regularities are represented by the set of spectral basis functions $\mathbf{\Psi}$, which is learned by applying PCA or ICA to matrices of spectral source signals, such that $\mathbf{\Psi}^i$ consists of the spectral basis functions that represent the regularities of the source signals associated with the temporal basis function $\boldsymbol{\phi}_i$, that is, the regularities of source signals $\mathbf{c}_i^1$, ..., $\mathbf{c}_i^K$. ($\mathbf{\Psi}^i$ contains basis functions $\boldsymbol{\psi}_1^i$, ..., $\boldsymbol{\psi}_J^i$, and $\mathbf{\Psi}$ contains sets $\mathbf{\Psi}^1$, ..., $\mathbf{\Psi}^I$.)

Figure 2.14 shows the results obtained by PCA of the spectral source signals from PCA of the set of 10 impacts on a steel rod.[5] Since PCA models the data with orthogonal bases, all basis functions within each set $\mathbf{\Psi}^i$ are orthogonal. Comparing $\boldsymbol{\psi}_1^1$ with $\mathbf{c}_1^{St1}$, it can be seen that the energy found in spectrum $\mathbf{c}_1^{St1}$ is being represented by this spectral basis function. For instance, note the three peaks between 2 and 4 kHz in both lines. Even though $\mathbf{c}_1^{St2}$ and $\mathbf{c}_1^{St3}$ have peaks in the same region, they show less energy in these partials. This variability is accounted for in part by other spectral basis functions in $\mathbf{\Psi}^1$ and in part by $\mathbf{V}^k$. Note how $v_{1,1}^{St1}$ has a much higher value than $v_{1,1}^{St2}$ and $v_{1,1}^{St3}$.

PCA can also be applied to the spectral source signals that have been obtained by spectral ICA. The first and third graphs in the right column of figure 2.12 show the results obtained by PCA of the spectral source signals from ICA of the set of 10 impacts on the same aluminum rod. The set of spectral basis functions $\mathbf{\Psi}^a$ represent the regularities of the spectral source signals associated with $\boldsymbol{\phi}_a$. For instance, note how $\boldsymbol{\psi}_1^a$ represents the peaks close to 6 and 7 kHz, which can be seen in $\mathbf{c}_a^{Al1}$ and $\mathbf{c}_a^{Al2}$. Since these peaks are much lower (or negative) in $\mathbf{c}_a^{Al3}$, $v_{a,1}^{Al3}$ has a much lower value than $v_{a,1}^{Al1}$ and $v_{a,1}^{Al2}$ (these coefficients are not shown here).

The number of basis functions considered is arbitrary and depends on the application. It depends on how much of the structure of the sounds one needs to model. To completely represent the structure of the sounds, we need to be able to model all variability in all spectral source signals $\mathbf{c}_i^k$, and therefore we must consider all basis functions in $\mathbf{\Psi}$. However, the results show that when PCA is used there is a dominant component in each set $\mathbf{\Psi}^i$ that represents most of the structure in the spectral source signals $\mathbf{c}_i^k$. Thus, often a very good approximation of source signals $\mathbf{c}_i^k$ can be obtained by using a small subset of $\mathbf{\Psi}^i$.

Finally, we show some results from ICA of the spectral source signals that have been obtained by spectral ICA. Since the spectral basis functions learned by ICA are not restricted to be orthogonal, and not many sounds (and consequently not many spectral source signals) were used in this study,

[5] In order to make the graphs more readable, some of the basis functions $\boldsymbol{\psi}_j^i$ and corresponding coefficients $v_{i,j}^k$ have been multiplied by -1.

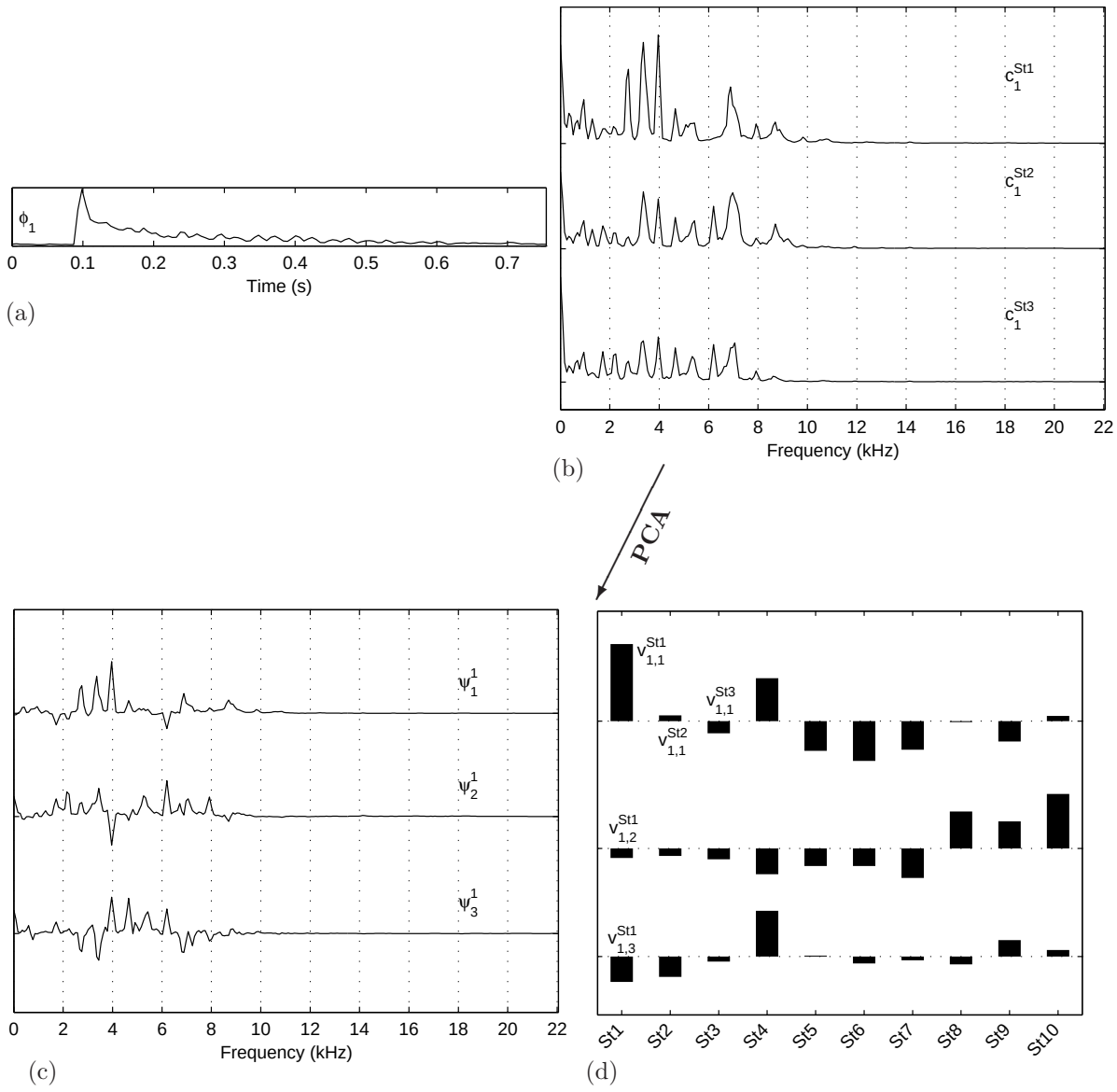


Figure 2.14: (Top row) Temporal basis functions $\mathbf{\Phi}$ and spectral source signals $\mathbf{C}^k$ obtained by spectral PCA of the set of 10 impacts on a steel rod. (a) First (most dominant) basis function. (b) First spectral source signal for sounds St1, St2 and St3. (Bottom row) Spectral basis functions $\mathbf{\Psi}$ and coefficients $\mathbf{V}^k$ (for $k \in \{\text{St1, St2, } \ldots, \text{St7}\}$) obtained by PCA of the source signals. (c) First three spectral basis functions from $\mathbf{\Psi}^1$. (d) Coefficients for spectral basis functions $\boldsymbol{\psi}_1^1$ to $\boldsymbol{\psi}_3^1$. The jth line, kth column shows $v_{1,j}^k$, that is, the coefficient for sound k and basis function $\boldsymbol{\psi}_j^1$.

the basis functions $\mathbf{\Psi}$ learned by ICA are more tuned to specific spectral source signals, that is, they resemble more closely the shape of specific spectral source signals. As a consequence, the representations obtained by ICA are less compact than the representations obtained by PCA. The second and fourth graphs in the right column of figure 2.12 show the results obtained by ICA of the spectral source signals from spectral ICA of the set of 10 impacts on an aluminum rod. It is interesting to note the similarities between individual spectral basis functions and spectral source signals. For instance, compare $\boldsymbol{\psi}_1^a$ to $\mathbf{c}_a^{Al1}$, and $\boldsymbol{\psi}_2^a$ and $\mathbf{c}_a^{Al2}$. See also how similar $\boldsymbol{\psi}_1^c$, $\boldsymbol{\psi}_2^c$ and $\boldsymbol{\psi}_3^c$ are.

2.5.3 Spectral basis functions $\mathbf{\Theta}$

While sections 2.5.1 and 2.5.2 referred to the results from model M_b (which is characterized by temporal basis functions $\mathbf{\Phi}$ and spectral basis functions $\mathbf{\Psi}$), this section refers to the results obtained by model M_r (which is characterized by spectral basis functions $\mathbf{\Theta}$ and temporal basis functions $\mathbf{\Lambda}$). In order to learn the set of spectral basis functions $\mathbf{\Theta}$ and decompose the spectrograms into sets of temporal source signals we use *temporal* PCA and ICA.

Again, we will start with the special case of a spectrogram $\mathbf{S}$ of a single sound. As explained in section 2.4, spectral PCA and temporal PCA give equivalent results where the roles of basis functions and source signals switch, which means that $\mathbf{\Theta}$ looks like $\mathbf{C}^k$ and $\mathbf{P}^k$ looks like $\mathbf{\Phi}$. For instance, the results of temporal PCA of sound Al2 look like the results in figure 2.7, where $\boldsymbol{\theta}_i$ looks like $\mathbf{c}_i^{Al2}$, and $\mathbf{p}_i^{Al2}$ looks like $\boldsymbol{\phi}_i$.

However, when temporal ICA is used, the results differ from the results obtained by spectral ICA. Figure 2.15 shows some results obtained by temporal ICA. Again, the roles of basis functions and source signals switch, but these results are not equivalent to those obtained by spectral ICA. Comparing these to figures 2.6a and 2.6c it can be seen that the spectral basis functions $\mathbf{\Theta}$ and temporal source signals $\mathbf{P}^k$ obtained by temporal ICA are noisier and not as easily interpretable as the results obtained by spectral ICA. (Note that the same sound was used in both cases and so these are directly comparable). For instance, the ringing property of aluminum sounds is easily identified in the results from spectral ICA (see $\boldsymbol{\phi}_b$ in figure 2.6a). The same is not true when temporal ICA is used. Instead, the ringing partial (3.95 kHz) is present in several spectral basis functions ($\boldsymbol{\theta}_a$ to $\boldsymbol{\theta}_e$ in figure 2.15), and each bump is reflected in one or more separate temporal source signals.

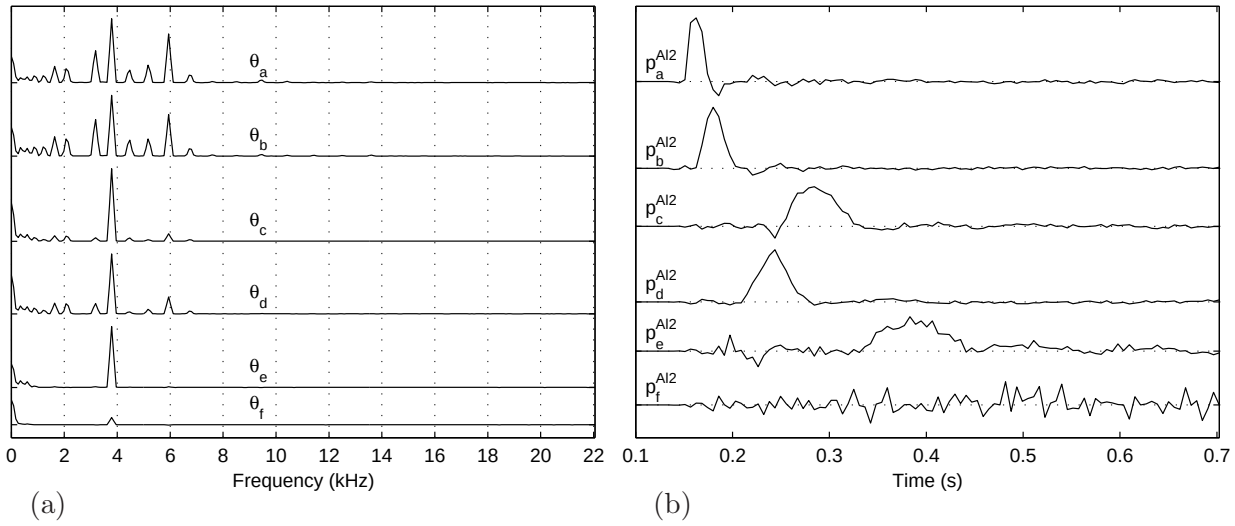
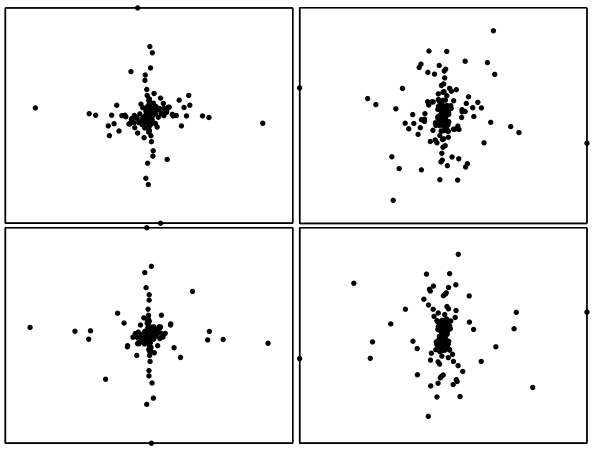


Figure 2.15: Spectral basis functions $\mathbf{\Theta}$ and temporal source signals $\mathbf{P}^{Al2}$ obtained by temporal ICA of the spectrogram of a sound (Al2) from an impact on an aluminum rod. (a) 6 out of the 22 most dominant basis functions are shown in decreasing order of dominance from top to bottom. (b) The corresponding temporal source signals are shown also from top to bottom.

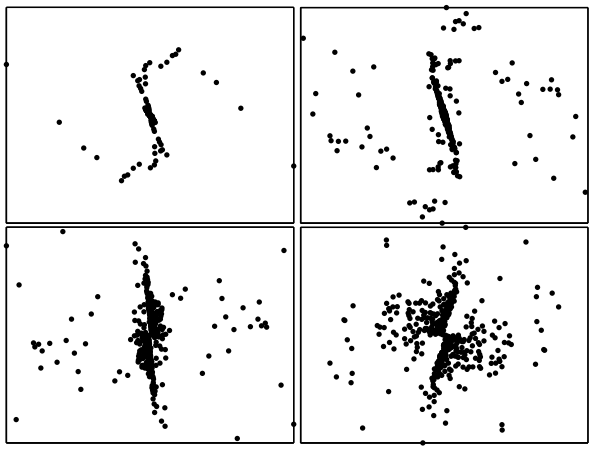
The reason for having worse results with temporal ICA is that there is structure that is not represented by the spectral basis functions learned by temporal ICA. Figure 2.16 shows that the source signals found by spectral and temporal ICA have different types of distributions. While most scatter plots obtained by spectral ICA (figure 2.16a) have clear directions that define the data, the same does not happen in the plots obtained by temporal ICA (figure 2.16b).

The lack of ability of the spectral basis functions (from temporal ICA) to explain the whole structure in the spectrogram results from the difference between ICA's underlying assumptions of the statistical model and the joint distribution of the bins (figure 2.5). As explained in section 2.4, the distribution of the frames approximates the distribution assumed by ICA better than the distribution of the bins. Since spectral ICA looks for correlations across the frames and temporal ICA looks for correlations across the bins, spectral analysis matches the statistics of the data better than temporal analysis. This property applies to all sounds that have the same type of bin and frame joint distribution as in figure 2.5.

Since temporal ICA is not well suited to explain the structure of the spectrograms, we will not analyze model M_r any further and all results presented hence forth are obtained using model M_b.



(a)



(b)

Figure 2.16: Scatter plots of source signals. Each plot shows one source signal plotted against another source signal. The source signals have been found by either spectral or temporal analysis of a spectrogram from a sound from an aluminum rod (Al1). (a) Results obtained by spectral ICA. (b) Results obtained by temporal ICA.

2.5.4 Variability

Natural sounds have significant variability as was illustrated in figures 2.1 to 2.3. Because model M_b is adapted to represent the distribution of the ensemble of impact sounds, it also captures this variability. The variability is represented by different basis functions (like $\boldsymbol{\phi}_c$ and $\boldsymbol{\phi}_e$ in figure 2.12) and by the distribution of the coefficients $v_{i,j}^k$. To illustrate this, figure 2.17 shows that by giving different values to the coefficients $v_{i,j}^k$, one can use different combinations of the temporal and spectral structures represented by the basis functions in $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ to simulate the variability present in the sounds. By randomly sampling the coefficients $v_{i,j}^k$ we can generate different instances from the model distribution. Figure 2.17a shows the variation in the partial at 3.95 kHz, and figure 2.17b shows four different synthesis instances of the same partial (3.95 kHz), each extracted from a different synthesized spectrogram. To synthesize the spectrograms we used the temporal basis functions ($\boldsymbol{\Phi}$) learned by spectral ICA of the spectrograms of 10 aluminum rod impacts, the spectral basis functions ($\boldsymbol{\Psi}$) learned by PCA of the corresponding spectral source signals and the coefficients obtained for one of the sounds ($\mathbf{V}^{Al4}$). To simulate the variability caused by $\boldsymbol{\phi}_c$ and $\boldsymbol{\phi}_e$

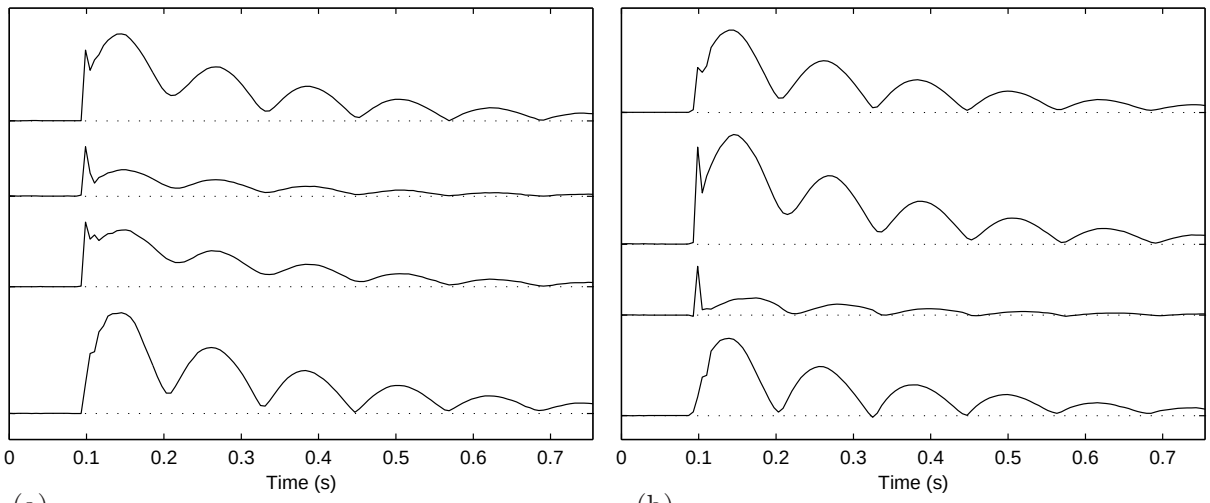


Figure 2.17: The decay shape of the partial at 3.95 kHz (which is partial e in figures 2.1 and 2.2) from different spectrograms of impacts on an aluminum rod. (a) The original partials show considerable variability. The partials (from top to bottom) were extracted from the spectrograms of sounds Al2, Al4, Al9 and Al10. (The partial from Al10 looks different in figure 2.3 because there it was plotted in a logarithmic scale.) (b) The synthesized partials have a similar range of variability. The partials were extracted from four synthesized spectrograms. See text for details.

(from figure 2.12) we varied the weightings of these two basis functions. That was done by varying the values of $\mathbf{v}_c^{Al4}$ and $\mathbf{v}_e^{Al4}$ for each synthesized spectrogram. The values were randomly sampled from the coefficients' distribution. (Note that in this way we are also varying the weightings of $\mathbf{\Psi}^c$ and $\mathbf{\Psi}^e$.) This figure confirms that model M_b is suited to represent the natural variability of the sounds. The variations obtained by the model are similar to those in the ensemble of impact sounds (compare the variations in figures 2.17a to those in 2.17b).

2.6 Summary and discussion

This chapter discusses the ISA method, a data-driven method for learning a representation of the intrinsic structures of impact sounds. We showed that, by using redundancy reduction techniques, it is possible to build a model that uses temporal and spectral basis functions that represent the intrinsic temporal and spectral structures of the sounds. While the techniques used here are ICA and PCA, other redundancy reduction techniques can also be considered. The ISA method can be used to characterize the structures of a single sound or the structures common to a set of impact sounds, in which case it also captures the natural variability in the structures. The method does not require any prior knowledge of the physics, acoustics or dynamics of the objects and events, and is able to represent the underlying acoustical structures in the sounds, which could offer advantages over previous knowledge-based models.

The ISA method can use one of two models: a model M_b that describes the spectrograms as an ordered set of bins and uses spectral analysis of the spectrograms to decompose them into spectral source signals, or a model M_r that describes the spectrograms as a sequence of frames and uses temporal analysis of spectrograms to decompose them into temporal source signals. Since spectral analysis is better suited to explain the structures of the spectrograms than temporal analysis, for the remaining of this dissertation the ISA method uses spectral analysis of the spectrograms (that is, it uses model M_b) unless stated otherwise. With model M_b, the temporal structures of the sounds are represented by the temporal basis functions $\mathbf{\Phi}$, which are learned by spectral analysis of the spectrograms. The spectral structures of the sounds are represented by the spectral basis functions $\mathbf{\Psi}$, which are learned in a second step by PCA or ICA of the spectral source signals associated with the temporal basis functions $\mathbf{\Phi}$.

Spectral ICA is able to decompose spectrograms into a small number of underlying features that characterize acoustic properties such as ringing, resonance, sustain, decay, and onsets. Since the

method is not restricted to learn explicit features (or structures) of the sounds, the representations obtained include new information that was not represented by previous physical models. For instance, features that are more abstract than simple decay rate or average spectra, like features that characterize ringing, or decay shapes that are not exponential, can now be modeled and easily extracted from the sounds. Spectral PCA gives compact representations of the temporal structures in the spectrograms. For instance, 6 basis functions can explain 96% or more of the variance of the data (section 2.5.1). Such low dimensional characterizations of the data can present advantages over previous physical models. For example, since impact sounds can have hundreds of partials [van den Doel, Pai, Adam, Kortchmar, and Pichora-Fuller, 2002], modeling them with equation 2.1 would mean using a very big N. When the objective is to model only the perceptually relevant portions of the sound, much less partials can be used (that is, N can be substantially smaller), yet determining which partials should be used is also a difficult question [van den Doel et al., 2002].

Brown and Smaragdis [2004] have used ICA to separate different notes from two-note musical trills. In another study the same authors have used non-negative matrix factorization (NMF), which is another redundancy reduction technique, to analyze polyphonic musical passages [Smaragdis and Brown, 2003]. Although these approaches are related to those presented here, their goal was to separate notes from musical segments with more than two notes. Even though the analyses used in both these studies resemble our analysis method, there are some fundamental differences. The main difference is that we are partitioning individual sounds according to the temporal behavior of the partials, whereas in their studies the sounds are being segmented according to events; we are interested in representing the structure of sounds of the same type efficiently, whereas they are interested in segmenting sound events. Also, while their analyses are appropriate for highly harmonic sounds, transient sounds with high structure variability are better described by our method, given that here individual sounds are represented by more than one temporal and spectral basis functions.

Most work employing redundancy reduction techniques (like ICA, PCA, NMF, singular value decomposition and sparse coding) and spectrograms or other time-frequency representations (like constant Q-transforms and wavelets), focuses on the source separation problem, and, as with the above two studies, it segments sounds according to events [e.g. Virtanen, 2004, Smaragdis, 2004, Barros, Rutkowski, Itakura, and Ohnishi, 2002, Casey and Westner, 2000]. Some MPEG-7 audio features are obtained using similar techniques and there has been work on sound classification,

recognition, and event detection using these features [e.g. Xiong, Radhakrishnan, Divakaran, and Huang, 2003, Kim, Berdahl, and Sikora, 2004]. All these studies use techniques similar to those used by the ISA method, but their goals are very different and, to the best of our knowledge, our method is the first to partition individual sounds according to the temporal behavior of the partials.

The ISA method has many applications, like in the study of the intrinsic structures of impact sounds, in sound recognition, in sound clustering, sound perception, etc. The method considers only the spectral (and temporal) content of the signals. Nonetheless, there is also complex structure in the phase of the signals, which is important for synthesizing sound waveforms from the model. As a result, even though the ISA method as it is, is useful in many areas, it has some limitations when we consider the modification and synthesis of the modeled sounds. Chapter 3 discusses these issues further and extends the method so that its limitations regarding synthesis are overcome.

Chapter 3

Synthesis of the Intrinsic Structures of Impact Sounds

An important part of this research is the synthesis of impact sounds using the structures learned by the method described in chapter 2. The reconstruction of sounds using (or excluding) a set of structures (features) will allow determining their relevance in sound perception. Furthermore, the possibility of manipulating the structures in order to modify the original sound is also of relevance to sound synthesis, virtual reality and multimedia.

Since the ISA method from chapter 2 uses magnitude spectrograms as the initial representation of the sounds, it loses phase information, and this may affect the resynthesis of the sounds. Here we start by looking into which portions of the signals are affected by the loss of phase information, and later we propose some extensions to the method so that the affected portions can be synthesized without deficiencies. We will also see that the structures learned by the extended method have essentially the same properties as the structures seen in chapter 2, but that these structures also have some new properties that reflect the way the extended method treats the signals. In addition, we also discuss how these underlying structures can be manipulated to modify the original signal and obtain new sounds.

3.1 The importance of phase

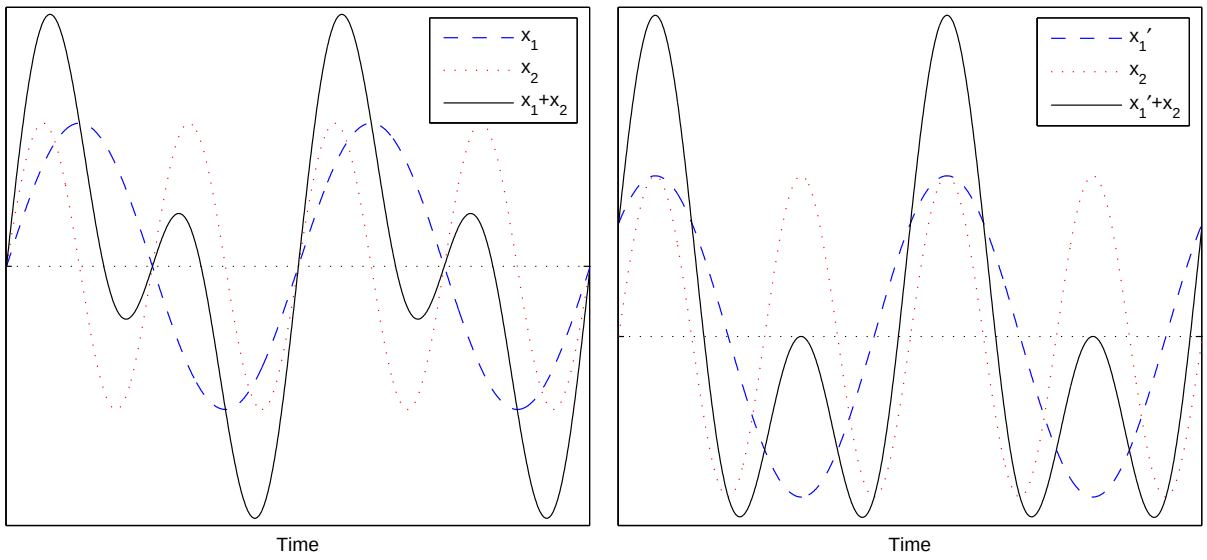
Phase information is not perceptually relevant in the steady state (or periodic) regions of the sounds [Roads, 1996, Ladefoged, 1996]. Periodic sounds can be combined with different phases without

any effect on the perceived sound. In other words, two sounds that have the same sinusoidal components with different phases are perceived as identical or are very difficult to distinguish. As an example, figure 3.1 shows two ways of combining a sine wave of 100 Hz with a sine wave of 200 Hz. In the first case (figure 3.1a) the two pure tones are added with initial phase 0. The result is shown as the solid line called $x_1 + x_2$. In the second case (figure 3.1b) the sine wave of 100 Hz is shifted by 45° and added to the other pure tone. The result is shown as the solid line named $x_1' + x_2$. In figure 3.1c the waveforms $x_1 + x_2$ and $x_1' + x_2$ were smoothed with a Hanning window to avoid the perceived clicks caused by sharp onsets and endings. Even though the waveforms $x_1 + x_2$ and $x_1' + x_2$ are different, their sinusoidal components and power spectra are the same. Moreover, when we hear them, we perceive them as the same sound.

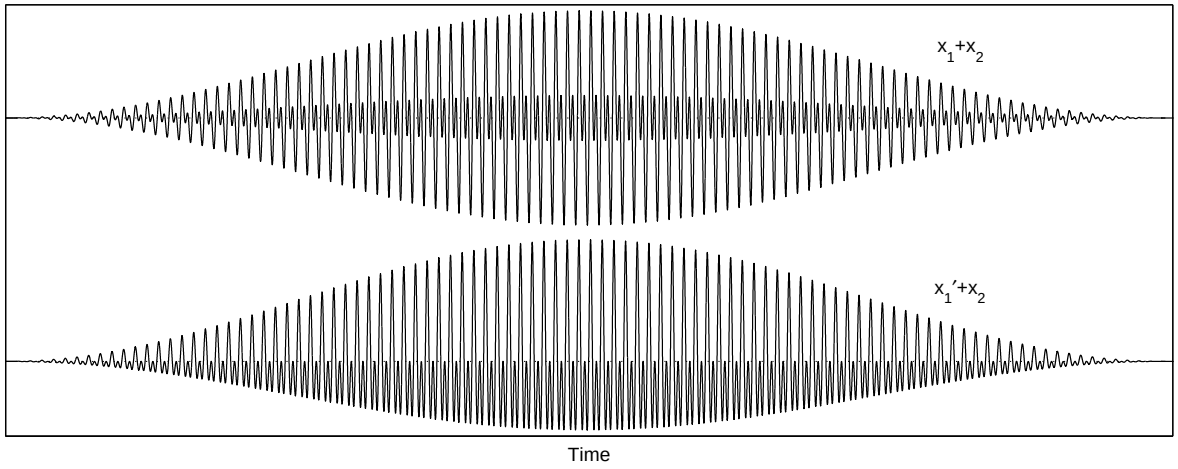
While phase information is not perceptually significant in the periodic regions of the sounds, it is vital for the perception of the attack transients, which are perceptually important portions of the sounds [Roads, 1996]. For instance, the successful recognition of musical instruments is mainly dependent on the information provided by the attack transients and their initial decay [Yost, 2000]. However, without correct phase alignment of the components in a signal, the resynthesis of attack transients is not successful.

Although the results from the ISA method can be used to (optionally modify and) resynthesize the periodic portions of the signals, for instance with sinusoidal modeling (which is covered in section 3.2.2), the same is not true for the transient portions of the sound. If our objective were simply to use all (unmodified) basis functions, we could reconstruct the whole magnitude spectrogram from the model and combine it with the phase spectrogram of the original sound to resynthesize the sound (for instance with the inverse Fourier transform). Yet our objective also includes synthesizing sounds associated with only a part of the structures, and reconstructing sounds after modifying structures or coefficients. Due to the loss of phase information, the modification and resynthesis of the attack transients are not very convincing (as the attacks are perceived as noisier and less sharp).

Later in this chapter, we propose an extension of the ISA method that includes representing transient sounds in a more suitable way to make the synthesis of transients possible. In addition, we also discuss how the underlying features can be manipulated to obtain new sounds. Having a good representation of all the components in the sound will allow the underlying features to be manipulated. For example, if we have recordings of impacts on a certain rod, ideally it will be



(a) (b)



(c)

Figure 3.1: Two pure tones (a sine wave of 100 Hz, x_1, and a sine wave of 200 Hz, x_2) are combined in two different ways. (a) Both sinusoids are added with initial phase 0. (b) x_1 is shifted by 45° and added to x_2. (c) The whole waveforms are windowed by a Hanning window.

possible to interpolate them so as to synthesize a new impact on the same rod.

3.2 Previous work on synthesis and signal decomposition

This section reviews some modeling and synthesis techniques that can be combined with the ISA method, and which include sinusoidal and transient synthesis methods. Even though many synthesis methods exist, here we focus only on those that are relevant to this work (see, for instance, [Roads, 1996, Borin, De Poli, and Sarti, 1997] for other synthesis methods). While section 2.1 focus mainly on physical modeling techniques (that is, models derived from the physical and dynamic properties of the object that produces the sound), here the focus is mainly on signal modeling techniques (that is, techniques that describe the acoustic structure of the sound, independently of the properties of the object).

3.2.1 Phase vocoder

The phase vocoder is one of the best known signal modeling techniques [Flanagan and Golden, 1966, Portnoff, 1976, 1981, Moorer, 1978]. This technique successfully models and synthesizes harmonic signals with static pitch characteristics. However, natural sounds can be inharmonic, and are typically not purely periodic because their sinusoidal components can have slowly time-varying frequencies. The phase vocoder has problems when dealing with pitch changing sounds, as it does not model the slow frequency variations in the partials. It has a bank of fixed frequency bandpass filters, with a sinusoid passing through each filter. Since the bandwidths of the filters are fixed, the frequency of each sinusoid is limited to within the bandwidth of its filter, that is, it cannot vary to outside that bandwidth. This limits the ability of the phase vocoder to model sounds with time-frequency variations in the partials. Also the sinusoids are assumed to be harmonic, which limits the phase vocoder's ability to model inharmonic sounds.

Over the years, many extensions and alternatives to the phase vocoder have been proposed [Almeida and Silva, 1984, Marques and Almeida, 1988, Griffin and Lim, 1988, Puckette, 1995, Laroche and Dolson, 1999a,b]. In the next section, we will cover two of these alternatives, namely the techniques proposed by McAulay and Quatieri [1986, 1995], and Smith and Serra [1987], which deal with inharmonic and pitch-changing sounds.

3.2.2 Sinusoidal modeling and synthesis

Sinusoidal models represent the sounds as a sum of sinusoids. Even though there are a few sinusoidal modeling techniques (for instance, see [George and Smith, 1992, Ding and Qian, 1997]), here we will focus on only two of them: the technique of McAulay and Quatieri [1986, 1995], also know as MQ modeling, and the technique of Smith and Serra [1987], called PARSHL (read as *partial*).

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 and they are both able to deal with inharmonic and pitch-changing sounds. The two methods have an analysis module, which creates a representation of the signal, and a synthesis module, which uses that representation to synthesize the sounds. Optionally, the representation created by the analysis module can be modified before it is used by the synthesis module, so that the latter synthesizes a sound that differs from the original.

Modeling sinusoids

Sinusoidal modeling represents signals as a sum of sinusoids with slowly varying amplitude and frequency,

$$y(t) = \sum_{r=1}^{R} A_r(t) \cos \theta_r(t), \tag{3.1}$$

where y is the representation of the signal, R is the number of modeled sinusoids, A_r is the instantaneous amplitude of sinusoid r, and θ_r is the instantaneous phase of that same sinusoid, which is given by the integral of the instantaneous frequency $\omega_r(t)$,

$$\theta_r(t) = \int_0^t \omega_r(\tau)\, d\tau. \tag{3.2}$$

Since the model uses instantaneous frequencies (instead of a constant frequency for each sinusoid) it is able to characterize pitch changing sounds, i.e., sounds that have partials with time-varying frequencies.

In order to determine the model parameters, the method considers that each slowly varying sinusoidal component of the signal is represented by a horizontal ridge of energy in the signal's spectrogram, and it uses 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 pa-

rameters that determine the track's instantaneous amplitudes, $A_r(t)$, and frequencies, $\omega_r(t)$. To avoid discontinuities between frames, the amplitude values in neighboring frames are linearly interpolated. Since the spectrogram is a discrete representation of the signal, the peaks in a frame (i.e., spectrum) correspond to approximations of the real frequencies in the signal. In order to have better estimations of the actual frequencies in the signal, the algorithm uses a cubic interpolation of the values around the peaks to estimate the instantaneous frequencies.

Synthesis of sinusoids

In order to transform the sound, its representation (that is, the sequences of parameters) can be modified, after which the synthesis module uses it to synthesize the sound represented by equation 3.1. The synthesis can be done in the time domain with an oscillator bank and additive synthesis: 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, or more precisely with an inverse short time Fourier transform (STFT). (Since the frames of the STFT correspond to overlapping windows of the waveform, the inverse Fourier transform must be combined with an overlap-add process so that the results obtained by each call to the inverse Fourier transform are adequately combined.)

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 [Serra, 1989, Serra and Smith, 1990, Serra, 1997]. This method is covered in the next section.

3.2.3 Spectral modeling synthesis

Spectral modeling synthesis (SMS) combines sinusoidal modeling and noise modeling to represent and synthesize sounds with both sinusoidal and noise components [Serra, 1989, Serra and Smith, 1990, Serra, 1997]. This technique is an extension of PARSHL (described in section 3.2.2) that was

primarily developed to overcome the problems of sinusoidal synthesis when dealing with signals with a broad band spectrum and stochastic characteristics. The method considers that signals can have sinusoidal components as well as stochastic components (from which noise is an example) and treats the two types of components in different ways: the method has a sinusoidal modeling module and a sinusoidal synthesis module that deal with the sinusoidal components, and it has a noise modeling module and a noise synthesis module that deal with the stochastic components of the sound. This section gives details about each of these modules and how they are integrated.

Modeling sinusoids and noise

The sinusoidal modeling part of SMS is similar to what was described in section 3.2.2. This module represents the sinusoidal components of the signal, which usually correspond to the main modes of vibration of the sound source, as a sum of sinusoids with slowly varying amplitudes and frequencies. (There are some slight differences between this module and the methods described in section 3.2.2. For instance, the peak tracking algorithm is slightly different. We are not covering those differences here. For more details see [Serra, 1997].)

The noise modeling part of SMS, which is the main improvement of SMS in relation to sinusoidal modeling, represents the non-sinusoidal portions of the signal, which include the excitation energy that is not transformed into stationary vibrations of the sound source (like the sound of the bow sliding against a string in a string instrument). 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. The module 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.

$$e(t) = \int_0^t h(t,\tau)u(\tau)\,d\tau, \tag{3.3}$$

where e represents the stochastic components (i.e., the time-varying filtered white noise), h is the time-varying filter and u is white noise. This filter is created with a linear approximation of the stochastic components' magnitude spectrum (see [Serra, 1997, 1989] for more details).

The temporal signal is modeled by combining the representation of the sinusoidal and stochastic components,

$$y_{sms}(t) = \sum_{r=1}^{R} A_r(t)\cos\theta_r(t) + e(t), \tag{3.4}$$

where y_{sms} is the representation of the signal (here we use the subscript $_{sms}$ to distinguish this variable from y in equation 3.1), A_r and θ_r are the instantaneous amplitude and phase of sinusoid r, and R is the number of modeled sinusoids. This model is an extension of the model given by equation 3.1 that includes the new term for the representation of the stochastic components.

While we have given details on how the stochastic components of the signal are represented, we have not seen how to extract, or distinguish, these components from the sinusoidal portion of the signal. To extract the stochastic portion of the signal, SMS first models the original signal, which here we call x, with sinusoidal modeling, that is, with equation 3.1. It then computes the residual r that is not represented by this equation, by removing y from x,

$$r(t) = x(t) - y(t), \tag{3.5}$$

where y is the portion of the original signal that is effectively represented by equation 3.1.

SMS can compute r by subtracting the spectrogram of y from the spectrogram of x, or by subtracting the temporal signal $y(t)$ from $x(t)$. Both the spectrogram of y and the temporal signal $y(t)$ can be built using the information provided by the sinusoidal model. If the latter approach is chosen, i.e., if SMS subtracts the temporal signals, $y(t)$ is obtained by sinusoidal synthesis, as explained in section 3.2.2. (In this case, the instantaneous phases of x, which are used to estimate the instantaneous frequencies ω, have to be preserved and used in the synthesis of y so that the shape of the waveforms match and can be successfully subtracted.)

Synthesis of sinusoids and noise

The sinusoidal synthesis part of SMS is similar to what was described in section 3.2.2. This module synthesizes the modeled sinusoids, after optional transformations to their representation, with sinusoidal synthesis as in the MQ method and PARSHL. The noise synthesis part of SMS can synthesize the stochastic components by filtering white noise in the time domain or by doing an inverse Fourier transform of the spectral envelopes of the stochastic components. The latter technique uses a complex spectrum for each frame. It creates the magnitude spectrum using the magnitude envelopes computed in the noise modeling module, and it fills the phase spectrum with random values.

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. Using SMS it is

possible to preserve these noises while reducing or removing other less relevant types of noises, like background noise. However, this technique fails 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 [Masri, 1996, Ali, 1996, Verma, Levine, and Meng, 1997, Verma, 1999, Verma and Meng, 2000]. In section 3.2.4 we describe one of these methods, namely the Transient Modeling Synthesis.

3.2.4 Transient modeling synthesis

Transient modeling synthesis (TMS) is an analysis/synthesis method proposed by Verma and colleagues to model and synthesize transient sounds [Verma et al., 1997, Verma, 1999, Verma and Meng, 2000]. As it will be explained in section 3.2.5, TMS can be combined with SMS to model the sinusoidal, attack transients and noise portions in the sounds.

Modeling transients

Sinusoidal modeling (as in PARSHL and the MQ method) uses a peak tracking algorithm to track the spectrogram's energy ridges that correspond to the sinusoids from the time domain of periodic sounds. As explained in section 3.2.2, this algorithm builds the time tracks by identifying the peaks in the magnitude spectra. The idea behind TMS is to use this same algorithm but to track peaks in a different domain, the frequency domain. 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. The peak tracking technique described in section 3.2.2 for sinusoidal modeling can be used to track these slowly varying curves, that is, it can be used to build frequency tracks.

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 was chosen because it represents the transients as sinusoids in the frequency domain.

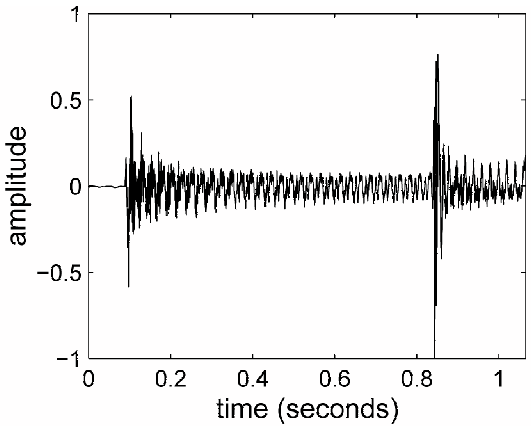
This transform maps the signal into a frequency by amplitude space and retains the phase information. (For more details on the DCT, please refer to appendix D.) Figure 3.2 illustrates this space transformation. In figure 3.2a the waveform $x(t)$ is represented in a time by amplitude space. Figure 3.2b shows the DCT of $x(t)$, here the signal is represented in a space of cosine basis functions (represented by a DCT bin number) by amplitude, i.e. a space of frequency by amplitude.

In the second step, TMS computes the magnitude spectrogram (i.e., magnitude STFT) of the DCT of x, to represent the signal in a frequency by time space. Figure 3.2c shows the STFT of the DCT of x. The signal is represented in a space of STFT frame number (in the horizontal axis) by discrete Fourier transform (DFT) bin number (in the vertical axis). The STFT frame represents a window of DCT bins, and in turn, as was seen above, a DCT bin corresponds to a cosine basis function, or frequency. Thus, the STFT frame stands for frequency. As shown in appendix D, 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. Therefore, low frequency sinusoids represent signals towards the left side of the window, while high frequency sinusoids represent signals towards the right side of the window. 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. Since the STFT frame correspond to frequency and the DFT bins correspond to time, the STFT of the DCT represents the signal in a space of frequency by time.

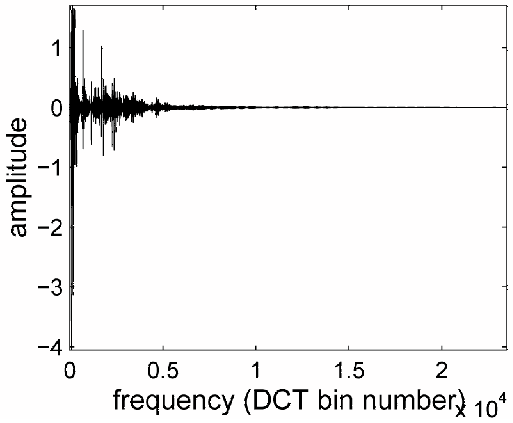
The vertical lines in figure 3.2c represent the sinusoids of the sinusoidal portion of the signal, which are easily modeled by SMS and sinusoidal modeling. The horizontal lines in the figure are the transients of x, which in this space are represented by slowly varying sinusoids. In order to model the slowly varying sinusoids on the spectrogram of the DCT, TMS uses a peak tracking algorithm, which is similar to the one used in PARSHL and the MQ method. 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).

Synthesis of transients

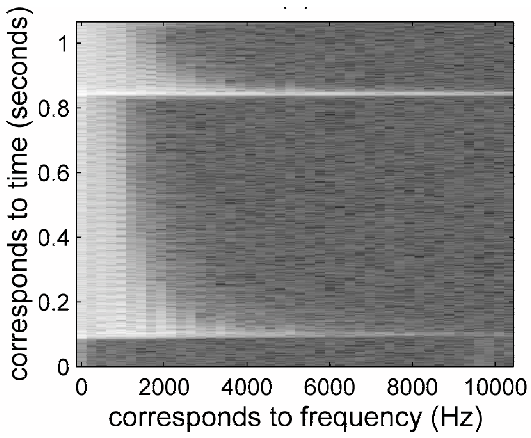
Once the transients are modeled by these sequences of parameters, and after optional modifications to the parameters, the transients can be synthesized by first using a process similar to the synthesis process described in section 3.2.2 (for sinusoidal synthesis), followed by an Inverse DCT.



(a)



(b)



(c)

Figure 3.2: Reproduced from [Verma, 1999]. TMS signal transformation. (a) The waveform $x(t)$. (b) Using the DCT of $x(t)$, the signal is transformed into a frequency by amplitude space. (c) The STFT of the DCT of $x(t)$. Here the signal is represented in a frequency by time space.

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.

3.2.5 Sines, transients and noise modeling and synthesis

The three-part sines+transients+noise model (which we will refer to as the S+T+N model) is an extension of Serra and Smith's SMS, proposed by Verma and colleagues to model the sinusoidal, transients and noise portions in the sounds [Verma et al., 1997, Verma, 1999, Verma and Meng, 2000]. This model combines SMS and TMS to this end (see sections 3.2.3 and 3.2.4). It separates the signal into three parts: sinusoids, noise and transients. The sinusoids and noise are modeled and synthesized by SMS, and the transients are modeled and synthesized by TMS.

Modeling of sinusoids, transients and noise

Like in SMS, the S+T+N model first uses sinusoidal modeling and synthesis to model and synthesize the sinusoidal portion of the signal (which is y in equation 3.1, and which here we call s for sinusoidal portion). It then removes the sinusoidal portion s from the original signal x and produces a first residual r that consists of transients and noise (for more details see section 3.2.3),

$$r(t) = x(t) - s(t).$$

Afterwards, it uses the modeling and synthesis components of TMS, to model the attack transients, a, in the residual r, which are then removed from r. As a result, a second residual n consisting of noise is produced:

$$n(t) = r(t) - a(t).$$

This second residual is then analyzed by the noise modeling component of SMS (as explained in section 3.2.3). (The order of this analysis does not necessarily have to be this. It is possible to first analyze and extract the transients, then the sinusoidal portion, and finally the noise). As a result, the S+T+N model obtains a separate representation for the sinusoids, transients and noise.

Synthesis of sinusoids, transients and noise

These representations can be (optionally) modified and synthesized by the synthesis modules of SMS and TMS. The three synthesized signals (synthesized sinusoids, synthesized transients, and synthesized noise) can be combined by addition to obtain a final synthesized signal.

3.2.6 Other methods for modeling and synthesis of sinusoids and transients

The previous sections discuss some techniques to model and synthesize the sinusoids, transients and noise 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 [1997] proposed a parametric method to extract the sinusoids from the spectrogram. George and Smith [1992] proposed the analysis-by-synthesis overlap-add method, which extracts sinusoids in an iterative way. Other sinusoidal modeling techniques and extensions to the methods discussed above include [Ding and Qian, 1997, Marchand, 1998, Masri, 1996, Ellis and Vercoe, 1992]. Also, some other solutions have been proposed to estimate the sinusoidal parameters. For example, Depalle, Garcia, and Rodet [1993] use hidden Markov models for that purpose, while Lagrange, Marchand, Raspaud, and Rault [2003] use linear prediction of the frequency evolutions of the partials. Some algorithms to estimate the sinusoidal parameters are discussed and compared in [Hainsworth and Macleod, 2003].

There are also other ways to treat the noise in the signal. For example, Desainte-Catherine and Hanna [2000] use a statistical approach to model noise. Ding and Qian [1997] use linear predictive coding (LPC) to synthesize the residual noise, and Goodwin [1996] uses equivalent rectangular bands.

Also other methods have been proposed to model transients. For instance, Christensen and van de Par [2006] model transients with a sum of sinusoids, whose amplitudes are modulated by gamma envelopes. Nsabimana and Zölzer [2006] proposed improvements to TMS (which was discussed in section 3.2.4). Still other methods to estimate and model transients have been proposed in the literature [Rodet and Jaillet, 2001, Fitz, Haken, and Christensen, 2000, Thornburg and Gouyon, 2000, Masri, 1996, Masri and Bateman, 1996, Ali, 1996].

3.3 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. However, the same is not true when it comes to synthesizing the attack transients of the sounds. As explained before, this is due to the lack of phase information in the magnitude spectrograms

(other possible causes of the problem are dependent on the synthesis method, for instance if we use sinusoidal modeling with a limited number of synthesized pure tones, this leads to poor attack synthesis). In this section, the ISA method is extended 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 part*, or *analysis method*, that decomposes the sounds into basis functions that represent their structures, and a *synthesis part*, or *synthesis method*, that uses those basis functions to produce waveforms (figure 3.3). The analysis part contains a module that divides the signal into two parts: the *sinusoidal sub-signal*, which we call s, and the *transient sub-signal*, which we call a (from attack). Ideally, the first sub-signal should consist of the pure tones in the original signal. The second sub-signal should be composed of the transients in the original signal, which are the very brief component signals characterized by a sudden increase and decrease of energy and which usually have a noisy broad band spectrum [Yost, 2000, Picket, 1999, McAdams and Bigand, 1993]. For now, we will ignore the details about the procedure used to obtain the sub-signals s and a. The pre-process signal box of figure 3.3 is discussed in more detail in section 3.4.2. Apart from this module, the analysis part also contains modules that learn the basis functions that represent the structures in the sinusoidal sub-signals (ISA method box in figure 3.3) and in the transients sub-signals (transients method box in the same figure). More details about these two modules will be given below.

The synthesis part receives the basis functions learned by the ISA method and transients method, and uses them to synthesize new sounds. In order to synthesize a sinusoidal signal s', the modified sinusoidal synthesis module uses the basis functions that represent the structures of the sinusoidal sub-signals s. Similarly, the modified TMS module uses the basis functions learned for the transients to synthesize a transients signal a'. In order to obtain a final synthesized sound y, the synthesis method combines waveforms s' and a'. More details about the modified sinusoidal synthesis and modified TMS modules are given below. Optionally, the synthesis method can include modifications to the basis functions and coefficients to obtain sounds that differ from the originals. These modifications are done by the modification modules (see the modifications boxes in figure 3.3), which will be discussed in more detail in section 3.6. For now we will only discuss synthesis without modifications.

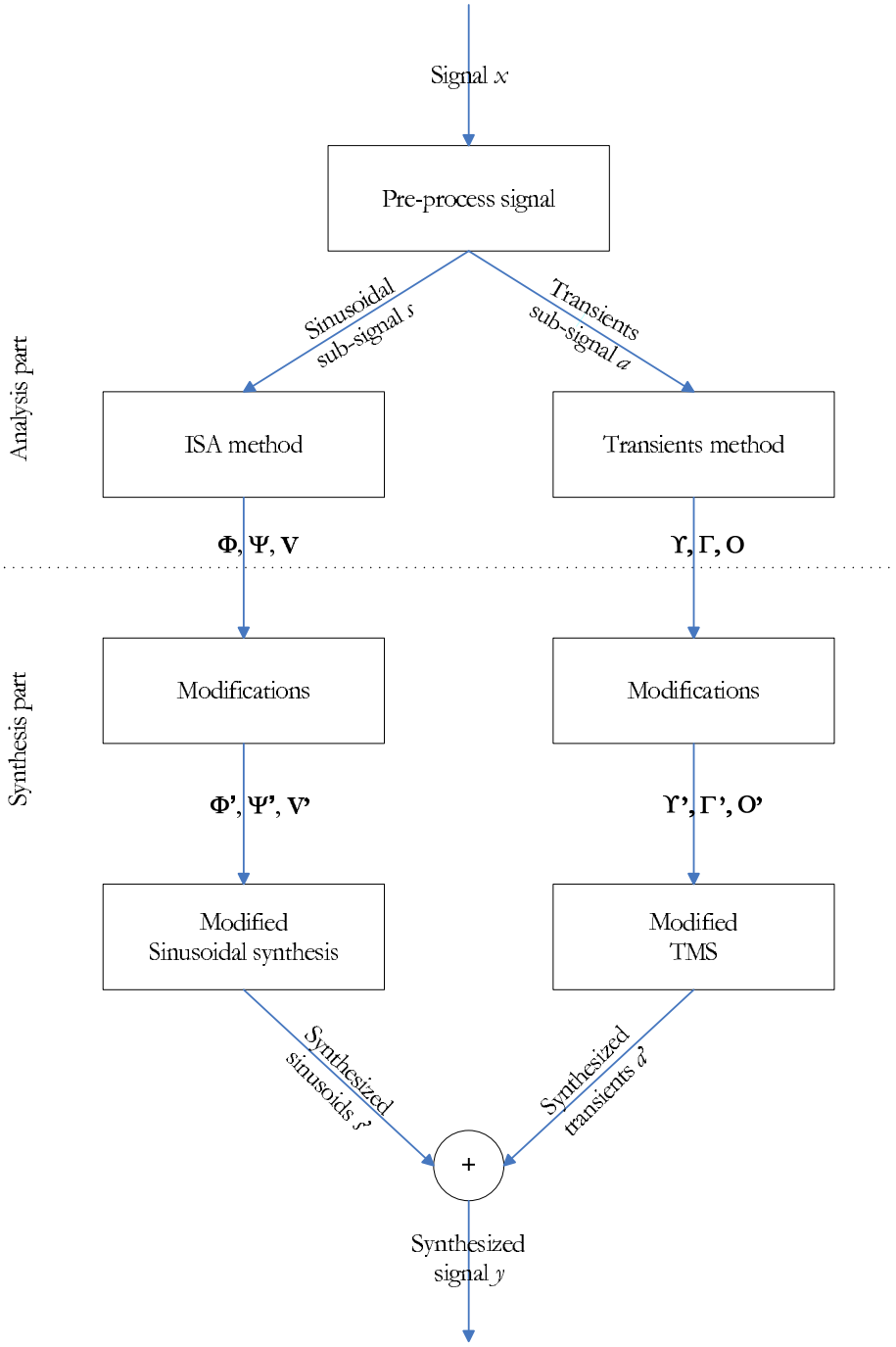


Figure 3.3: ISAS Method. The analysis part divides the signal into two sub-signals: the *sinusoidal sub-signal* s, which is processed by the ISA method, and the *transients sub-signal* a, which is processed by the transients method. The synthesis part uses the output of the analysis part to generate a final waveform y. After some optional modifications to the basis functions and coefficients from the ISA method, these are used by the synthesis part to generate a sinusoidal waveform s'. Similarly, after some optional modifications to the outputs of the transients method, the synthesis part produces a transients waveform a'. The two synthesized waveforms are combined to obtain the final waveform y.

Modeling and synthesis of the sinusoidal sub-signal

The analysis of the sinusoidal sub-signal s (ISA method box in figure 3.3) is done as before by the ISA method, which starts by representing s with a spectrogram that is then used to learn the temporal and spectral basis functions, $\mathbf{\Phi}$ and $\mathbf{\Psi}$, respectively (see details in chapter 2). Since the input to the ISA method here is different from that used in chapter 2 (i.e., sinusoidal sub-signal s instead of the original waveform x) one can wonder if the learned structures will differ. Section 3.5.1 shows the type of structures learned when the inputs to the ISA method are sinusoidal sub-signals, and compares them to those discussed in chapter 2 when the inputs were the whole signals.

Since 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, as seen before, phase information is not perceptually significant in the periodic regions of the sounds. To synthesize the structures represented by $\mathbf{\Phi}$ and $\mathbf{\Psi}$, the synthesis part can use an algorithm similar to sinusoidal modeling and synthesis, as in the MQ method and PARSHL (this corresponds to the modified sinusoidal synthesis box in figure 3.3). The method builds a spectrogram from the information in the sets of basis functions $\mathbf{\Phi}$ and $\mathbf{\Psi}$, and coefficients $\mathbf{V}$, where $\mathbf{V}$ represents the set of sets of coefficients $\mathbf{V}^k$ (see equation 2.12), and use 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 and additive synthesis) to obtain a sinusoidal waveform s' .

While that is the approach we use for the modified sinusoidal synthesis module, an alternative would be to build the sequences of parameters that represent the peak tracks directly from the information in the basis functions and coefficients. Just as the peak tracking algorithm uses an interpolation of the neighboring values in the frames to better estimate the peaks' frequencies, here, the algorithm must interpolate the values of neighboring values not only within a spectrum, but also among the different spectra (defined by the basis functions).

Modeling and synthesis of the transients sub-signal

In order to avoid the problems introduced by the lack of phase information, the analysis part treats the transients sub-signal a in a different way. To begin with, the analysis part, or more specifically, the transients method module in figure 3.3, 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 3.2.4). 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 transients method then decomposes this periodic signal into its underlying structures, represented by a set of spectral basis functions $\mathbf{\Upsilon}$, a set of temporal basis functions $\mathbf{\Gamma}$ and a set of coefficients $\mathbf{O}$. In order to learn these structures, this module uses a process similar to the ISA method. Section 3.5.2 gives a more detailed description of the transients method, and shows the types of structures that it learns.

The synthesis part (in the modified TMS box in figure 3.3) uses the representation of these structures (that is, the basis functions and the coefficients obtained by the transients method) and a process similar to TMS, to produce a transients waveform a': First, it builds a spectrogram from the information in the basis functions in $\mathbf{\Upsilon}$ and $\mathbf{\Gamma}$, and coefficients $\mathbf{O}$. It then 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.

Residual noise

The residual noise, which consists of the noisy components in the original signal, can also be considered. Depending on the type of signals used, the residual noise can contain information that is more useful than just background or recording noise. For example, the sounds of musical instruments may contain types of noises that one may not wish to discard.

The residual noise can be obtained by removing the sinusoidal sub-signal s and the transients sub-signal a from the original signal,

$$n(t) = x(t) - s(t) - a(t). \tag{3.6}$$

n can be considered as a third sub-signal (extracted by the pre-process signal box in figure 3.3), and it can be treated by a third series of modules in the ISAS method, parallel to the modules that treat the sinusoidal and transient sub-signals. These can include a module that analyses the residual noise, one that modifies it, and one that synthesizes it.

The tasks performed by these modules will depend on the goal of the application. For example, if the ISAS method is used to analyze a wind instrument sound that contains breath noise, which should be kept, and background noise, which should be discarded, the modules that analyze, modify and synthesize noise can be similar to the noise analysis, and synthesis of SMS: the residual noise can be modeled as filtered white noise, where the filter will keep the breath noise but will attenuate or eliminate the background noise. If the goal is to analyze the structure of noise that

is common to an ensemble of sounds, the ISA method can be used to analyze the residual noises of the ensemble. As mentioned in chapter 2 we have also used the ISA method to analyze sounds recorded in a normal room (with background noise and reverberation). In this case, some of the learned basis functions accounted for the structure of noise. The temporal structures learned looked random, which agrees with SMS assumption that the noise components in the signal do not require a precise description of the time-varying magnitude shape of each frequency bin. Having said that, this analysis method can be combined with SMS to synthesize a noise signal: the learned noise structures can be represented by time-varying frequency shaping-filters that describe the expected magnitude of each frequency bin over time and the filters can be applied to white noise.

As it will be seen in section 4, we obtained very good results by considering only the sub-signals s and a. Therefore we will not analyze the residual noise any further.

Summary

To summarize, the ISAS method consists of two parts: the analysis and the synthesis methods. The analysis method divides the original signal into sub-signals s and a, which are analyzed in different ways. The initial representation used for each sub-signal is different: a spectrogram of the waveform for the sinusoidal sub-signal s, and a spectrogram of the DCT for the transients sub-signal a. As a consequence, the type of structures learned to represent the sinusoids and transients are also different (section 3.5 shows the type of structures learned in each case).

Those structures are used by the synthesis method to obtain the synthesized waveforms s' and a', which are then combined in order to produce the final synthesized signal,

$$y(t) = s'(t) + a'(t).$$

Even though we have been explaining the ISAS method in terms of the representation of one signal, this method can be used to analyze ensembles of sounds, so that the structures common to sounds of the same type, their variability and the structures that differ in sounds of different types are identified. This is a major advantage over most methods discussed in section 3.2, which were developed to analyze and modify one sound at a time. In short, instead of decomposing only one signal into its sinusoidal and transients sub-signals, the method will decompose several signals into their sub-signals. These sub-signals are fed to the appropriate modules and are analyzed as a set. (For instance, all the sinusoidal sub-signals are passed together to the ISA method box in figure 3.3. The basis functions $\mathbf{\Phi}$ and $\mathbf{\Psi}$ will reflect the properties of the ensemble of sounds.) The

synthesis method can synthesize one or several waveforms y, where the different waveforms are obtained through the use of different coefficients.

3.4 Sub-signal extraction

As seen in section 3.3, the ISAS method starts by dividing the original signal x into a sinusoidal sub-signal s, and a transients sub-signal a. This operation is performed by the first module of the analysis part in figure 3.3, the pre-process signal module, and will be explored in detail in section 3.4.2, where a few possible approaches to divide the signal into the sub-signals s and a are discussed. Before we explore those approaches, section 3.4.1 introduces matching pursuit, a technique that is used by some of these approaches. Afterwards, sections 3.4.3 and 3.4.4 explore some implementation options for detecting if the signal contains transients, and sinusoidal components. These are needed by some of the approaches explored in section 3.4.2 that use matching pursuit, but the reader can safely skip sections 3.4.1, 3.4.3 and 3.4.4, if he/she is not interested in these approaches.

3.4.1 Matching pursuit

Matching pursuit (MP) is a non-linear greedy algorithm that decomposes the signal into a linear combination of functions, also called *atoms*, from an overcomplete basis, that is, a basis that has more basis functions than dimensions, and which is also called a *dictionary* in the MP context [Mallat and Zhang, 1993, Davis, Mallat, and Avellaneda, 1997]. MP is an iterative algorithm that at each step chooses the atom that best represents (a part of) the signal, and removes the structure represented by that atom from the signal. As a result, at each iteration the algorithm obtains a residual, which will be further decomposed in the next iteration. The algorithm runs until a stopping criterion is met. This is often when the signal to noise ratio falls below a given threshold, but other criteria can also be used (sections 3.4.3 and 3.4.4 give some examples). Since MP uses an overcomplete basis, its dictionary can contain more than the necessary basis functions to span a given space. Therefore, it can have functions (or atoms) that are appropriate to describe different types of signals (for instance, some atoms can be specialized in describing impulses while other can be specialized in describing sinusoidal waves). In this way, MP can obtain representations that are more suitable to represent the inherent structures in the signal. For instance, while when other methods are used, the representation of the structures of certain types of signals may be spread across the whole basis, however, with MP it is possible to obtain a sparse representation using the

atoms that more closely represent the structure in the signal.

Many different dictionaries have been proposed. Mallat and Zhang [1993] have used a dictionary of Gabor functions. Having a dictionary that contains a wide range of scaled and translated Gabor functions is appropriate to give good time and frequency representations of the signals. Variations of Gabor dictionaries have also been proposed, such as the dictionary of harmonic atoms, which are combinations of Gabor atoms with the same time localization but centered at different frequencies and which are suitable to represent harmonic sounds [Gribonval and Bacry, 2003]. However, due to their symmetry, Gabor atoms are not the best option to represent signals with asymmetric characteristics, like transients. Using these atoms to represent asymmetric signals introduces pre-echo artifacts in the residual. One solution to avoid introducing pre-echo artifacts is to use a correlation function that takes into account whether the atom represents energy from silent parts of the signal [Gribonval, Depalle, Rodet, Bacry, and Mallat, 1996].

Goodwin [1997] has used a dictionary of damped sinusoids, with varying damping factors, modulation frequencies and start times. This dictionary avoids the pre-echo artifacts and models the signal as a sum of exponentially decaying sinusoids. Even though this representation bears similarities to Prony's representation (see section 2.1), an advantage here is that the start time of the damped sinusoidal waves can vary. A combination of Gabor atoms and damped sinusoidal atoms has also been proposed to combine the advantages of both dictionaries [Goodwin and Vetterli, 1999].

Smith and Lewicki have used a dictionary of gammatone atoms [Lewicki, 2002b, Smith and Lewicki, 2005, Smith, 2006]. This dictionary is motivated by natural sound statistics and it has been shown to give efficient codes of the signals. It is also possible to use dictionaries that fit the statistics of the waveforms. For example, Smith and Lewicki [2006] propose a dictionary that is adapted to fit the statistics of the waveforms of natural sounds. They show that both the gammatone dictionary and the fitted dictionary give efficient representations, also called *spike codes*, of stationary and non-stationary sounds, in particular speech and music (where efficiency is measured by the sparseness of the code and the representation of the underlying structure of the sounds). (While the spike codes can be learned with MP, other implementations are also possible [Lewicki and Sejnowski, 1999, Lewicki, 2002a, Smith and Lewicki, 2005, Smith, 2006].)

3.4.2 Sub-signal extraction algorithms

In this section, we explore the pre-process signal module of the ISAS method (figure 3.3). More specifically, here we discuss a few possible approaches to divide the signal into the sinusoidal sub-signal s, which consists of the pure tones in x, and transients sub-signal a, which consists of very brief component signals with a sudden increase and decrease of energy, and usually with a broad band spectrum.

Approach *ST*. One possibility to extract the sub-signals from the original sound x, is to use a variant of the S+T+N model (section 3.2.5). In order to extract the sinusoidal sub-signal s, the steady sinusoids in the spectrogram of x are identified with a peak tracking algorithm (for more details on this algorithm, see in section 3.2.2). Then these sinusoids are synthesized (with a bank of oscillators and additive synthesis) to produce the sub-signal s. A residual r_s that consists of transients and noise, is then obtained by removing s from the original signal x (figure 3.4):

$$r_s(t) = x(t) - s(t).$$

This process is similar to that used in SMS (section 3.2.3).

The next step consists of extracting the transients sub-signal a, from the residual r_s. Since our goal is to have the ability of synthesizing all parts of impact sounds, including the transients, and since representing transients with spectrograms does not permit their accurate synthesis (due to the loss of phase information inherent in the spectrograms), the process described in the previous paragraph is not appropriate to extract the transients sub-signal a from r_s. Instead, the transients sub-signal can be extracted from r_s by a process similar to that used in TMS: the steady sinusoids in the spectrogram of the DCT of r_s are identified with a peak tracking algorithm, and then synthesized with a bank of oscillators, additive synthesis and an inverse DCT to produce the sub-signal a. (See section 3.2.4 for more details.) Even though the idea behind the S+T+N model is similar to this algorithm, the process used to extract the sinusoids from x is different because the S+T+N model uses MP to implement sinusoidal modeling [Verma, 1999].

Approach *TS*. A variant of this process inverts the order by which the transients and the sinusoids are extracted. Similarly to what was explained above, the transients sub-signal a can be extracted with TMS, but now it is extracted from the original signal x rather than from r_s. a is

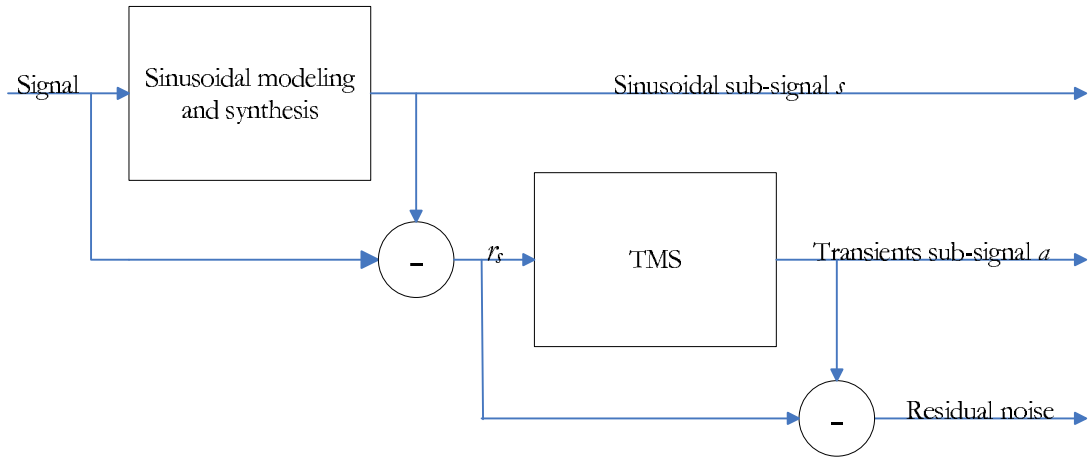


Figure 3.4: Approach *ST*. The sinusoidal sub-signal s is extracted from the spectrogram of the original signal with a peak tracking algorithm (as in sinusoidal modeling and synthesis). The transients sub-signal a is extracted with a process similar to that used in TMS.

then removed from x to obtain a residual r_a that contains the sinusoids and noise in the signal:

$$r_a(t) = x(t) - a(t).$$

The next step consists of extracting the sinusoids s from r_a: the steady sinusoids in the spectrogram of r_a are identified with a peak tracking algorithm, and then are synthesized to produce the sub-signal s.

This second variant may present a disadvantage. Since the transients sub-signal a is the first to be extracted, it is possible that some of the energy of the sinusoidal part of x ends up being represented in a (instead of s).

In section 4.1 we will compare the results obtained by these two approaches, and in the rest of this thesis we use approach *ST*. However, there are many alternative ways of extracting the sub-signals. Below we will discuss some of these alternative approaches.

Approach *SinMod+MP*. Another solution to decompose signal x into s and a would be to use the same process as in approaches *ST* and *TS* to extract the sinusoidal sub-signal, but MP to extract the transients (section 3.4.1). We can either first extract the sinusoidal sub-signal s from the spectrogram of x and then use MP on the residual r_s to extract the transients a, or we can first use MP on the original signal x to extract the transients, and then extract the sinusoids s from the

spectrogram of the residual r_a, which consists of sinusoids and noise.

In either case, we need to use a stopping criterion, which in the first case can be signal to noise ratio (as ideally, when MP stops, the residual will consist of only noise), and in the second case it must be a *transient measure*, that is, some measure that indicates whether the residual still contains transients. Possible implementations of this measure are discussed in more detail in section 3.4.3. When MP stops running, we can build the transient sub-signal a by adding the (weighted) chosen atoms in the appropriate time frames.

In order to extract the transients using MP, we need a dictionary suitable to represent transients. Some possibilities are Smith and Lewicki's spikes dictionary [Smith and Lewicki, 2005, 2006, Smith, 2006], and Goodwin's damped sinusoidal waves dictionary [Goodwin, 1997]. Both of these dictionaries have fast decaying asymmetric atoms, which are suitable to represent transient signals.

Approach *MP*. MP can also be used to extract both the transients and the sinusoids from the original signal. It can be used once with a dictionary of amplitude modulated sinusoidal waves to extract the sinusoidal sub-signal s and once with a dictionary suitable to represent transients (as discussed in approach *SinMod+MP*) to extract the transients sub-signal a. Alternatively, we can use only one dictionary that contains amplitude modulated sinusoids as well as other atoms appropriate to representing transients.

In the former case and depending on which sub-signal is extracted first, we need to use a *sinusoidal measure* (that is, a measure that indicates if the residual still contains sinusoids, more details about this measure are given in section 3.4.4) or a transient measure as the stopping criterion. The stopping criterion for the second stage of this process can be signal to noise ratio.

In the later case, in each iteration MP will choose either a sinusoid or another atom from the dictionary. The residual at each step will consist of the residual from the previous step less the chosen (weighted) atom. Since, ideally, after this process terminates the residual will consist of only noise, the stopping criterion can be signal to noise ratio. In the end of this process, we can build the sinusoidal sub-signal by adding the (weighted) sinusoidal atoms chosen by MP (in the appropriate time frames), and similarly we can build the transients sub-signal by using the (weighted) transient atoms chosen by MP (also in the appropriate time frames).

Ideally, when any of these different algorithms terminates, the final residual, which is the residual noise in the figures above, will consist of the noisy part of the original signal (like background noise).

The algorithms considered here use (or are variations of) the ideas behind sinusoidal modeling and synthesis, its peak tracking algorithm, SMS, TMS, the S+T+N model and MP. Yet, these are not the only options that may be considered to extract the sub-signals. As mentioned in section 3.2.6, there are other methods to estimate and to extract the sinusoids and transients in the signal. Algorithms that use those methodologies also may be considered.

3.4.3 Transient measure

In order to extract the transients from the signal using MP, some of the sub-signal extraction methods proposed in section 3.4.2 must use a *transient measure* to identify whether a signal contains transients. These methods will keep trying to extract transient structures form the signal while the measure indicates that the signal still contains transients. Since transients are characterized by abrupt changes in energy, their detection is usually done by inspecting the energy variations in the signal. The basic idea behind transient detection is to look for rapid and strong changes in energy. This section reviews some possible solutions to measure the energy changes in the signal and to implement the transient measure.

We have implemented an attack transient detection method that uses the energy in the spectrogram frames (that is, the overall power of all frequencies in each frame). To identify the attacks of the impacts, this method looks for the first local maximum that is above a certain threshold and that is preceded by silence or background noise. This method was used to align the impact sounds for the analysis reported in chapter 2 (so that all impacts have the same amount of silence, or background noise, before they start). Even though this method can identify attack transients in the sounds used in chapter 2, it may fail to detect transients in more complex sounds, such as when the loudness of the background noise changes. Also, this method will not work if more than one sound is combined because, in that case, transients are not necessarily preceded by silence or noise.

In the context of the S+T+N model, Verma et al. propose a transient detection method that relies on the rapid energy increase in the signal caused by transients [Verma, 1999, Verma et al., 1997]. This method compares the energy of the sinusoidal part of the signal to the energy of the residual. After decomposing the original signal into sinusoids and residual, the ratio of the normalized estimations of local energy from the sinusoids and from the residual is computed (local and global measures of energy are extracted from the power spectra of the signals). Transients are

identified as the regions where this ratio is higher than a threshold.

The same process can be done without dividing the signal into sinusoids and residual [Verma, 1999, Verma et al., 1997]. In this approach, they use only the higher frequencies in the signal. Again, the local and global energies of the signal are measured to obtain a normalized measure of local energy. Transients are identified by thresholding the normalized local energy. This latter solution is more appropriate for our purposes, because the methods proposed in section 3.4.2 need the transient measure to analyze sounds from which the sinusoidal components have not been extracted.

Röbel [2003, 2005] has proposed a slightly different approach, which operates on individual frequency regions while taking into account individual spectral peaks. Instead of using the time evolution of the energy of the signal, this approach uses the center of gravity (COG) of the energy, which is measured from the power and phase spectra of the signal. The COG uses a measure of phase variation (more specifically, the group delay, which is the negative phase derivative) to estimate if the peaks consist of transient peaks. If the COG of a peak is in the right region of the window, this is a potential attack transient peak. To distinguish attack transient peaks from noise peaks, the method relies on the synchronization between different spectral peaks.

Another solution that uses the power spectra of the sound and looks into individual regions of frequency was proposed by Rodet and Jaillet [2001]. This method looks for triangular shapes in the power spectra of different frequency bins. Large peaks are considered to correspond to attack transients if synchronized with large peaks from other frequency bins.

3.4.4 Sinusoidal measure

In order to decide when to stop MP from trying to extract sinusoidal components from the signal, some of the sub-signal extraction methods proposed in section 3.4.2 must identify whether a signal contains sinusoids. For that, these methods can use a *sinusoidal measure* that indicates whether a signal contains pure tones or only transients and noise. The methods will keep trying to extract sinusoidal structures from the signal while the sinusoidal measure indicates that the signal contains pure tones.

A simple way to identify if a signal has pure tones is to look for high peaks in its Fourier spectrum. Since transients and noise are mainly characterized by broad band spectra, if the signal's spectrum has some peaks higher than the average peaks' height, these may indicate the presence

of pure tones. This measure has a very simple implementation and may prove efficient for some signals. Nonetheless, it also may fail, for instance if the signal contains only pure tones of lower average energy than that of the transients.

More robust approaches are also possible. For instance, we can consider the time evolution of the spectrum (that is, the STFT). Since pure tones usually have a longer duration than transients, we can identify pure tones by looking for bins that show energy in more than a certain number of continuous frames. The number of frames to be considered should depend on the typical duration of transients.

This measure can work in ideal cases, for which the discrete Fourier transform (DFT) of a pure tone always corresponds to the same bin, independently of the frame. However, some variations on the representation of a pure tone may occur, because real signals are very complex and do not correspond exactly to the type of signals that the Fourier transform describes (which are pure periodic signals) and because this is a discrete representation of the signal with inherent quantization and leaking. Thus, one effect that must be taken into account is that pure tone descriptions may present some bin fluctuations from frame to frame. Also, we have to consider leaking, as a neighborhood of active bins may be describing one pure tone only. One solution that takes these effects into account is to identify the pure tones as the ridges of energy in the spectrogram, where a ridge corresponds to a bin or to a neighborhood of bins active in a sequence of frames, and it can account for bin fluctuations from frame to frame. This is equivalent to the peak tracking algorithm described in section 3.2.2.

Another DFT effect that must be taken into account is the potential interference between pure tones very close in frequency. Depending on the parameters used in the DFT computation, two close pure tones may be represented by one bin only that turns on and off continuously. We should consider this effect when analyzing the time evolution of the STFT to determine the number of continuous frames in which the bin (or the neighborhood of bins) is active.

3.5 Results

In this section we show the results obtained by the ISAS method. As seen above, the ISAS method is divided into the analysis and the synthesis parts, and even though this chapter focuses mainly on synthesis, here we explore the results of both these parts and not just the synthesis part. Since the analyzed signals (sub-signals s and a) now are different from those of chapter 2 (which were

the original signals x) one may ask if the properties of the learned structures are different from those seen in that chapter. In sections 3.5.1 and 3.5.2 we explore the similarities and differences between the basis functions obtained by the analysis part of the ISAS method and those obtained in chapter 2.

The ISAS method starts by subdividing each of the original waveforms into two sub-signals: the sinusoidal and transients sub-signals, s and a, respectively. As seen in section 3.4.2, this subdivision can be performed in several ways. In order to obtain sub-signals s and a, here, we used the first approach described in that section, that is, approach ST. The sub-signals s and a are then sent to the ISA method module and the transients method module, respectively (figure 3.3). Section 3.5.1 shows the results obtained by the ISA method module, that is, the analysis of the sinusoidal sub-signals, while section 3.5.2 shows the results obtained by the transients method module, that is, the analysis of transients sub-signals. Section 3.5.2 also gives more details on the implementation of the transients method.

Section 3.5.3 discussed the results from the synthesis part. There we show that the ISAS method is able to synthesized waveforms that contain both transients and slower decay portions. The synthesis results will be further explored in chapter 4, where we discuss tests and user studies done to validate the plausibility of the synthesized sounds.

3.5.1 Sinusoidal sub-signal modeling

The analysis of the sinusoidal sub-signals (ISA method box in figure 3.3) is equivalent to the analysis process introduced in chapter 2. Both methods represent the signals in terms of a set of temporal basis functions $\mathbf{\Phi}$ and a set of spectral basis functions $\mathbf{\Psi}$ (see model M_b in section 2.2). However, the ISAS method uses the spectrograms of the sinusoidal sub-signals, while the ISA method in chapter 2 starts with the spectrograms of the original sounds.

Temporal basis functions $\mathbf{\Phi}$

Like the ISA method, the ISAS method models the bins of the spectrograms as linear combinations of temporal basis functions $\boldsymbol{\phi}_i$ (equation 2.8). In order to learn the set of these temporal basis functions $\mathbf{\Phi}$ and to decompose the spectrograms into sets of spectral source signals, the ISAS method applies spectral PCA or spectral ICA to the spectrogram of a single sinusoidal sub-signal or to the spectrograms of different sinusoidal sub-signals. (For more details, please refer to model M_b in chapter 2.)

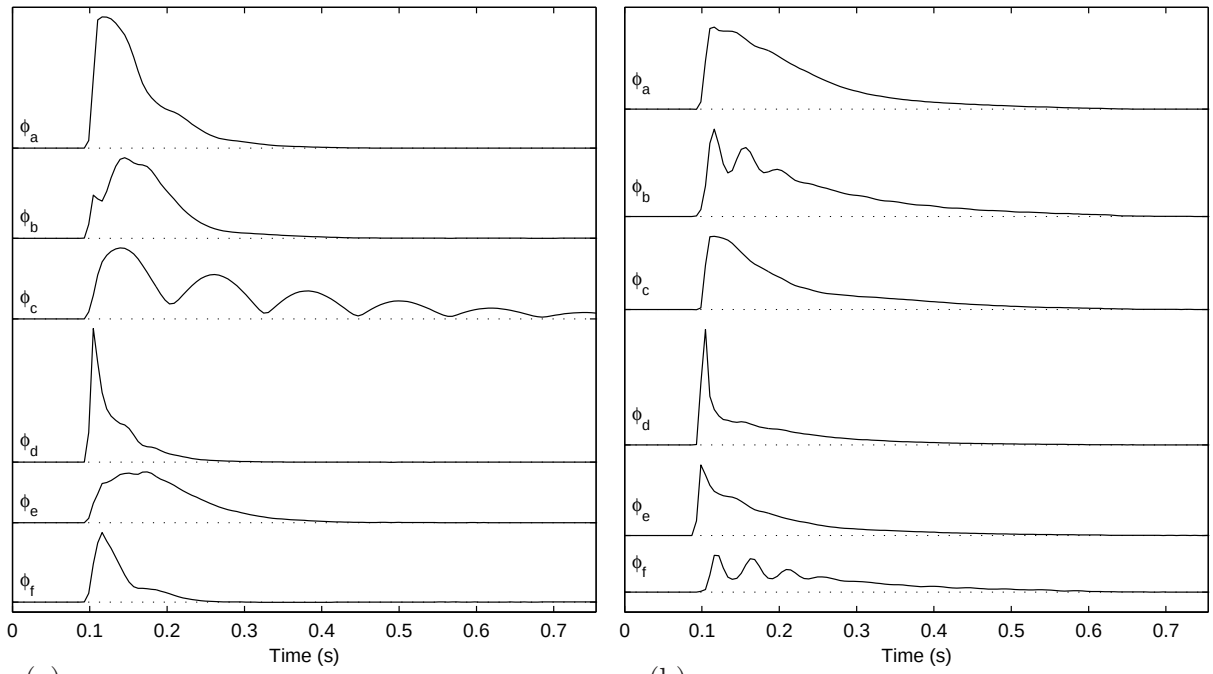


Figure 3.5: Temporal basis functions $\mathbf{\Phi}$ learned by ICA of the set of ten sinusoidal sub-signals from impacts on (a) an aluminum rod, and (b) a zinc plated steel rod. In each figure, 6 out of the 11 most dominant basis functions are shown in decreasing order of dominance from top to bottom.

Figure 3.5 shows the basis functions learned by spectral ICA of sets of sinusoidal sub-signals. The types of shapes obtained here are similar to the shapes obtained before with the ISA method in chapter 2: there are basis functions that represent ringing (like $\boldsymbol{\phi}_c$ in figure 3.5a, and $\boldsymbol{\phi}_b$ and $\boldsymbol{\phi}_f$ in figure 3.5b), sustain (like $\boldsymbol{\phi}_e$ in figure 3.5a), decay (like $\boldsymbol{\phi}_a$ in figures 3.5a and 3.5b), and onsets followed by very sharp decays (like $\boldsymbol{\phi}_d$ in figure 3.5b).

Since the transient parts of the original signals, and consequently most of the onset structure, now are represented in the transients sub-signals, and are not present in the sinusoidal sub-signals, there are two main differences between the results obtained here and those from chapter 2. First, there are fewer basis functions that represent the onsets followed by very sharp decays. Second, the spectral source signals associated with these basis functions were usually characterized by a broad band spectrum in chapter 2, but here this characteristic is much less pronounced. For example, compare the spectral source signals in figure C.1b (in appendix C), which are mostly broad band, with those in figure 3.6, which have better defined partials and are less noisy. (The first set of spectral source signals is associated with the temporal basis function $\boldsymbol{\phi}_b$ from figure 2.10a and

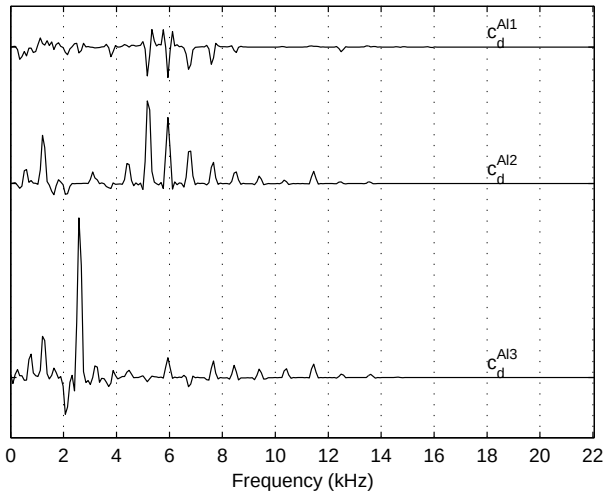


Figure 3.6: Spectral source signals ($\mathbf{C}^k$) obtained by spectral ICA of ten sinusoidal sub-signals from impacts on an aluminum rod. The spectral source signals corresponding to the basis function $\boldsymbol{\phi}_d$ in figure 3.5a and sounds Al1, Al2 and Al3, are shown from top to bottom.

the second is associated with the temporal basis function $\boldsymbol{\phi}_d$ from figure 3.5a, where both basis functions have a sharp decay).

The set of temporal basis functions $\boldsymbol{\Phi}$ learned by spectral PCA of the spectrograms of the sinusoidal sub-signals has similarities to the set of basis functions learned by spectral PCA of the spectrograms of the original signals. In both sets, the dominant basis function $\boldsymbol{\phi}_1$ shapes the overall decay of all partials, while other less significant basis functions account for temporal behaviors that differ from this overall decay shape.

On the other hand, there are also interesting differences. The sinusoidal sub-signals contain only part of the structure present in the original signals; they lack most of the structure of the attack and noise. Consequently, the shape of the basis functions learned here may differ from the shape of the basis functions reported in chapter 2. In other words, since much of the power in the original signals is absent from the sinusoidal sub-signals, the basis functions are now free to point to directions that differ from those seen in chapter 2. For instance, $\boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_3$ in figure 3.7a have a ringing shape that is not present in $\boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_3$ from figure 2.11a. The cause of the differences is the sinusoidal characteristic of the signals used to learn the basis function in figure 3.7a. Since these signals contain no or less noise and transient structures, the basis functions do not need to account for the temporal shape of noise or for broad band structure. This is very obvious when we observe the spectral source signals associated with the basis functions: Figure 3.8 shows the spectral source

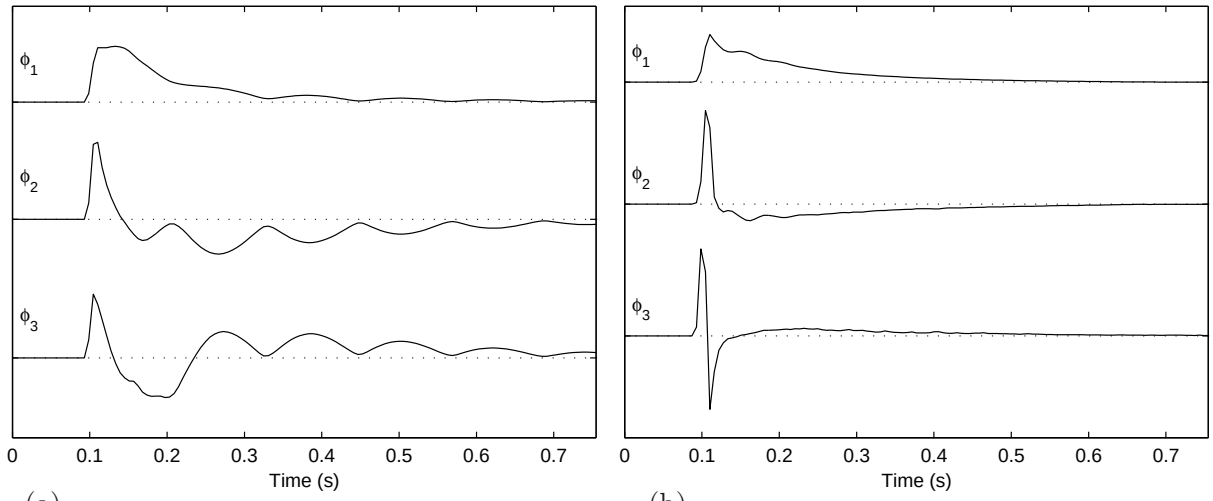


Figure 3.7: Temporal basis functions $\mathbf{\Phi}$ learned by PCA of the set of ten sinusoidal sub-signals from impacts on (a) an aluminum rod, and (b) zinc plated steel rod. In each figure, the first three basis functions are shown from top to bottom.

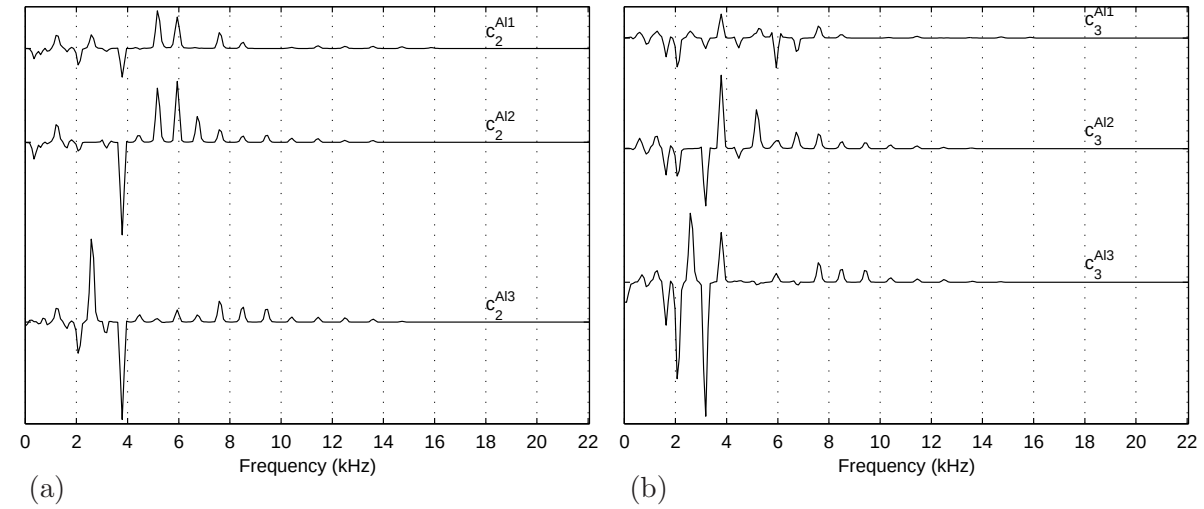


Figure 3.8: Spectral source signals ($\mathbf{C}^k$) obtained by spectral PCA of ten sinusoidal sub-signals from impacts on an aluminum rod. Second and third spectral source signals for to sounds Al1, Al2 and Al3 are shown from top to bottom: (a) shows $\mathbf{c}_2^{Al1}$ to $\mathbf{c}_2^{Al3}$, and (b) shows $\mathbf{c}_3^{Al1}$ to $\mathbf{c}_3^{Al3}$.

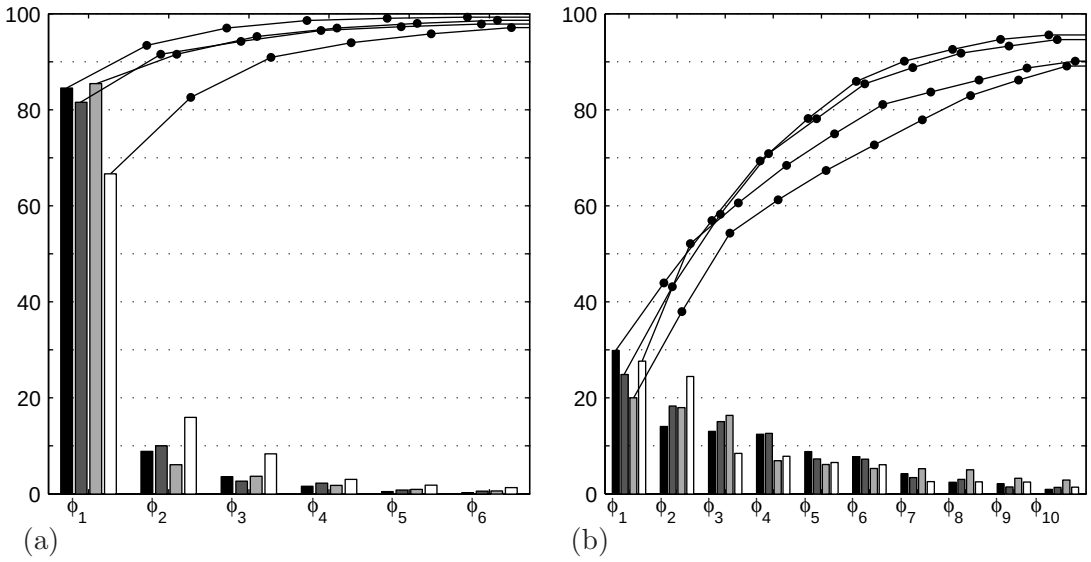


Figure 3.9: Percentage of variance explained by the basis functions in $\mathbf{\Phi}$ learned by spectral analysis of the set of 10 sinusoidal sub-signals from impacts on an aluminum rod (black), the set of 10 sinusoidal sub-signals from impacts on a zinc plated steel rod (dark grey), the set of 10 sinusoidal sub-signals from impacts on a steel rod (light grey), and the set of 10 sinusoidal sub-signals from impacts on a wooden rod (white). Only the values for the first six or ten temporal basis functions are shown. The dots on the curves show the cumulative sums of the percentages. In (a) $\mathbf{\Phi}$ was learned by PCA. In (b) $\mathbf{\Phi}$ was learned by ICA.

signals associated with $\boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_3$ from figure 3.7a, and figure C.2 shows those associated with $\boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_3$ from figure 2.11a. It is easily observed that the spectra in figure C.2 have noisier and more broad band characteristics.

Finally, we will look into the percentage of variance explained by the temporal basis functions $\mathbf{\Phi}$ learned by spectral PCA and spectral ICA (figure 3.9). Since the sinusoidal sub-signals lack most of the noise and transient structures, these signals are more regular than the original impacts. In other words, different sinusoidal sub-signals from impacts on the same rod are more similar to each other, than are the original impacts. Consequently, the method requires fewer basis functions to describe the sinusoidal sub-signals, or in other words, it is possible to explain the same percentage of variance in the sinusoidal sub-signals with fewer basis functions than those needed to explain the variance in the original signals. For instance, the 10 most dominant basis functions learned by spectral ICA on a set of 10 impacts, explains at most 84% of the variance of the data, while when

the data consists of sinusoidal sub-signals, the same number of basis functions explains at most 92% of the variance. The same observation is made with spectral PCA: 6 basis functions explain around 96% of the variance of 10 impacts, while the same number of basis functions can explain around 98% of the variance of 10 sinusoidal sub-signals (figures 2.13 and 3.9).

Spectral basis functions $\mathbf{\Psi}$

Similar to the ISA method, the ISAS method describes the spectral source signals as a linear combination of spectral basis functions $\boldsymbol{\psi}_j^i$ (equation 2.10). In order to learn the set of these spectral basis functions $\mathbf{\Psi}$, which represent the regularities in the spectral source signals $\mathbf{C}^k$, the ISAS method uses the same process as descibed in chapter 2: it applies PCA or ICA to matrices of spectral source signals. (For more details on this process, refer to model M_b in chapter 2.)

Here, the spectral basis functions in $\mathbf{\Psi}$ are very similar to the spectral basis functions from chapter 2 that are associated to the temporal basis functions in $\mathbf{\Phi}$ that characterize the periodic (i.e. sinusoidal) structures in the sounds (like ringing, sustain and slow decay). For instance, compare the right columns of figures 3.10 and 2.12, and the bottom lines in figures 3.11 and 2.14. In some cases, these similarities are remarkable. For example, the most dominant basis functions $\boldsymbol{\psi}_1^a$ and $\boldsymbol{\psi}_1^c$ in figure 3.10 are very similar to the same basis functions in figure 2.12. This suggests that the same structures are being described here and there. It means that the most dominant basis function learned by the ISA method in chapter 2 describes sinusoidal structures.

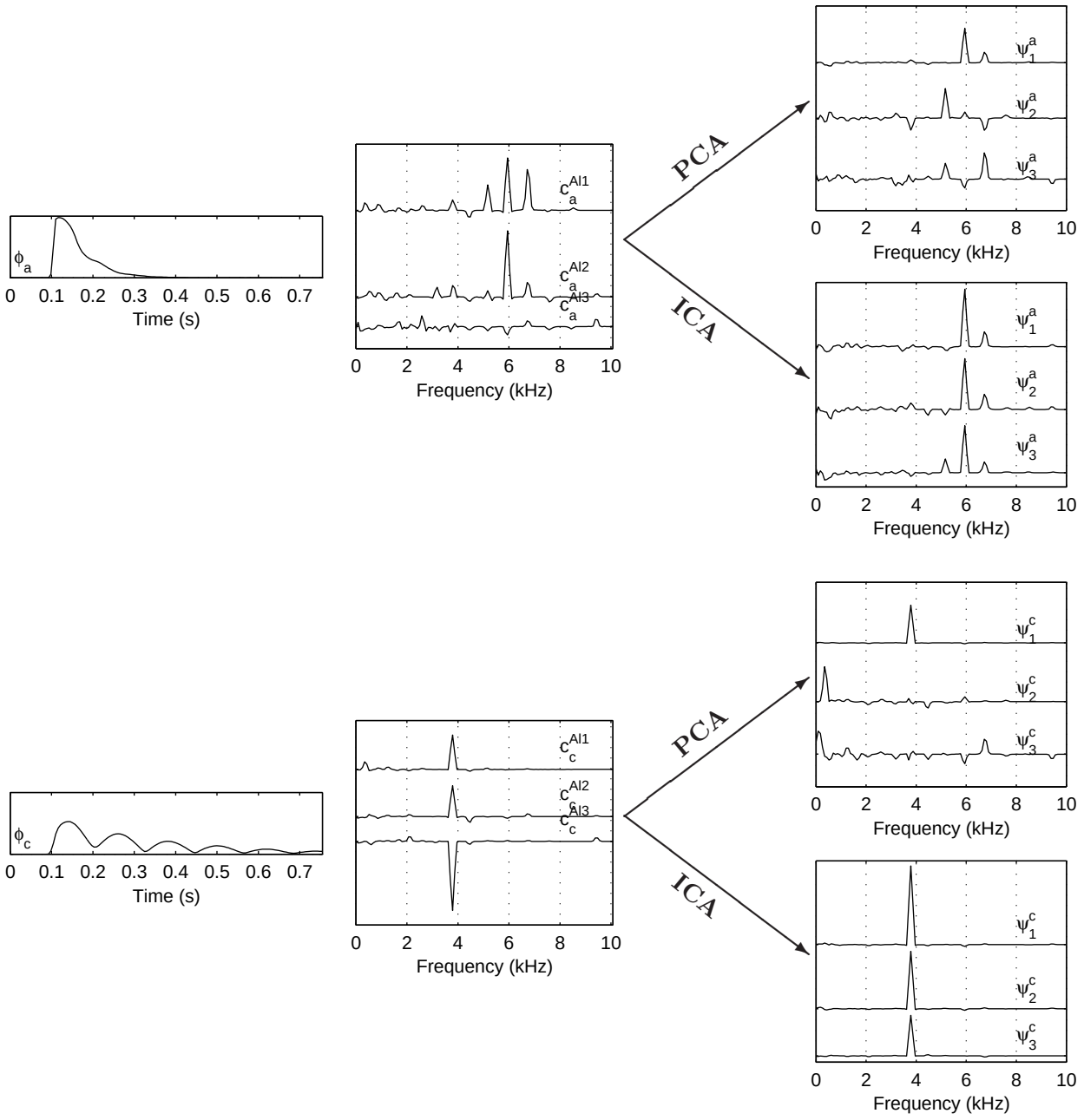


Figure 3.10: (Left column) Temporal basis functions $\boldsymbol{\phi}_a$, and $\boldsymbol{\phi}_c$ from figure 3.5a. These are learned by ICA of the set of 10 sinusoidal sub-signals from impacts on an aluminum rod. (Middle column) The corresponding spectral source signals for sounds Al1, Al2 and Al3. (Right column) Spectral basis functions $\boldsymbol{\Psi}$ obtained by analysis of the spectral source signals. The first and third figures in this column show the first three spectral basis functions from $\boldsymbol{\Psi}^a$ and $\boldsymbol{\Psi}^c$ learned by PCA. The second and fourth figures in this column show the first three spectral basis functions from $\boldsymbol{\Psi}^a$ and $\boldsymbol{\Psi}^c$ learned by ICA.

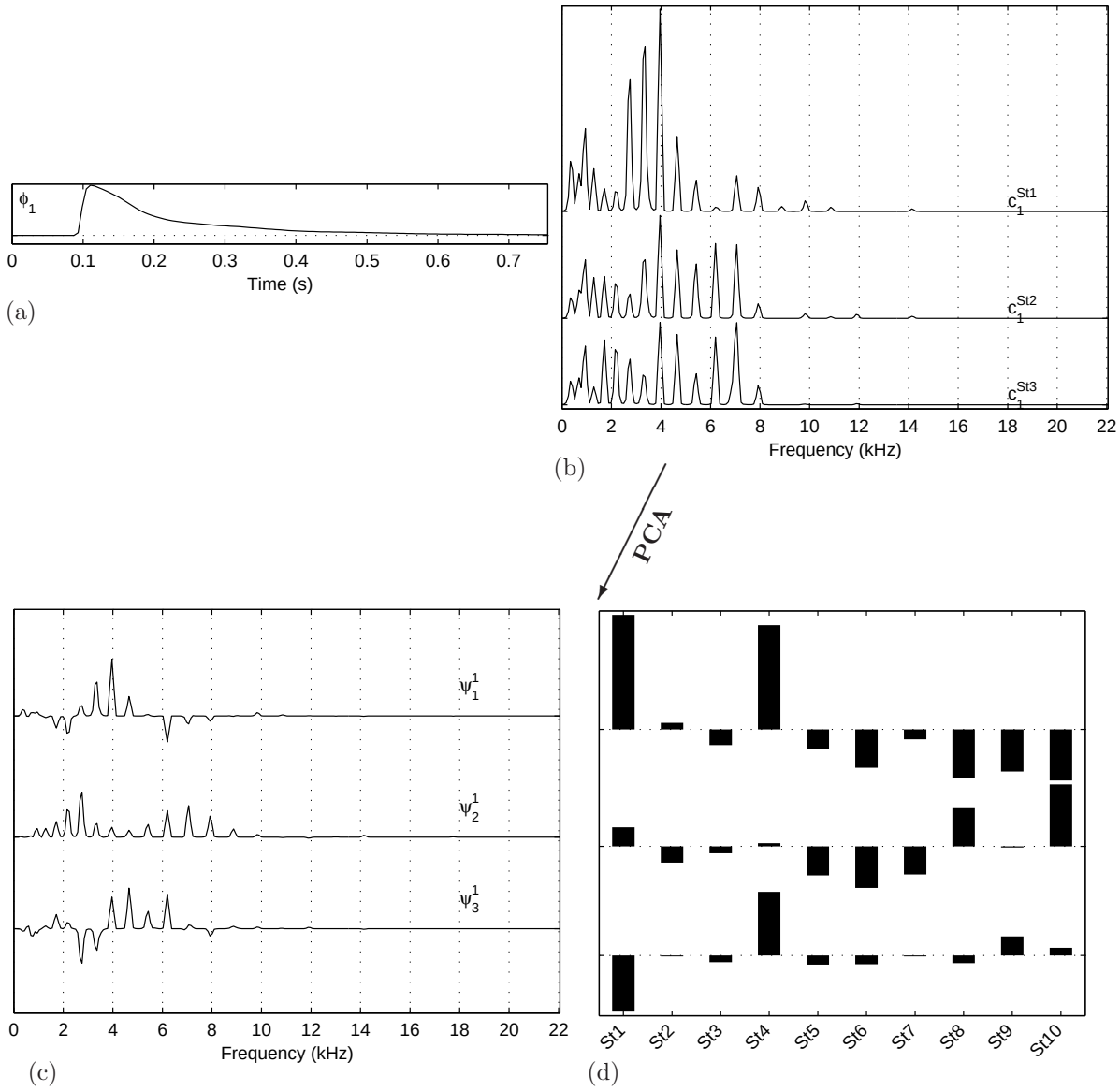


Figure 3.11: (Top row) Temporal basis functions $\mathbf{\Phi}$ and spectral source signals $\mathbf{C}^k$ obtained by spectral PCA of the set of 10 sinusoidal sub-signals from impacts on a steel rod. (a) First (most dominant) basis function. (b) First spectral source signal for sounds St1, St2 and St3. (Bottom row) Spectral basis functions $\mathbf{\Psi}$ and coefficients $\mathbf{V}^k$ (for $k \in \{\text{St1, St2, } \ldots, \text{St10}\}$) obtained by PCA of the source signals. (c) First three spectral basis functions from $\mathbf{\Psi}^1$. (d) Coefficients for spectral basis functions $\boldsymbol{\psi}_1^1$ to $\boldsymbol{\psi}_3^1$. The jth line, kth column shows $v_{1,j}^k$, that is, the coefficient for sound k and basis function $\boldsymbol{\psi}_j^1$.

3.5.2 Transients sub-signal modeling

The analysis of the transients sub-signal (transient method box in figure 3.3) starts by representing ensembles of transients signals (or a single transients signal) with the spectrogram of the DCT of the signals' waveforms. As seen in section 3.2.4, this represents the transients as a periodic signal that can easily be modified and synthesized while preserving the precise timing synchronization between the various frequency components of the signal. To distinguish the spectrogram of the DCT of the waveform from the spectrogram of the waveform ($\mathbf{S}^k$), we represent the former by $\mathbf{Z}^k$.

This module then decomposes the ensemble of spectrograms into its underlying structures; it uses a method similar to the ISA method (with model M_b) to decompose the signal into temporal and spectral basis functions. However, there are some differences between this and the ISA method, namelly the initial representation of the data (the spectrogram of the DCT of the waveform in the former case, and spectrogram of the waveform in the later case) and the meaning of the learned basis functions. (The basis functions now have different meanings because the frames of $\mathbf{Z}^k$ correspond to frequencies of the original signal, and the bins correspond to time, while in $\mathbf{S}^k$, the frames correspond to time, and the bins to frequencies).

Like the ISA method with model M_b, the transients method uses spectral analysis of the ensemble of spectrograms: it concatenates the spectrograms in the same way as the ISA method, such that the data matrix $\mathbf{X}$ is the concatenation of transposed spectrograms, $\mathbf{X} = ((\mathbf{Z}^1)^T, (\mathbf{Z}^2)^T, ...(\mathbf{Z}^K)^T)$, and it treats the (concatenated) frames, i.e. the rows of $\mathbf{X}$, as signal mixtures. As a result, the transients method learns a set of basis functions and a set of source signals. 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 vectors in the frequency space because the frames of $\mathbf{Z}^k$ correspond to frequencies of the original signal.

Figure 3.12 shows the most dominant basis functions learned by spectral PCA of transients from impacts on an aluminum rod and on a steel rod. As one could expect, these basis functions are very broad band, which is consistent with the broad band characteristics of the transients. Spectral PCA gives a very compact and efficient description of the data, in which the first basis function accounts for most of the variance of the data. On average, the most dominant basis function accounts for 78% of variance on $\mathbf{X}$, and the 6 most dominant basis functions account for 96% of this variance.

While model M_b obtains spectral source signals $\mathbf{C}^k$, the transients method obtains temporal source signals. The source signals obtained at this step are vectors in the time space because the

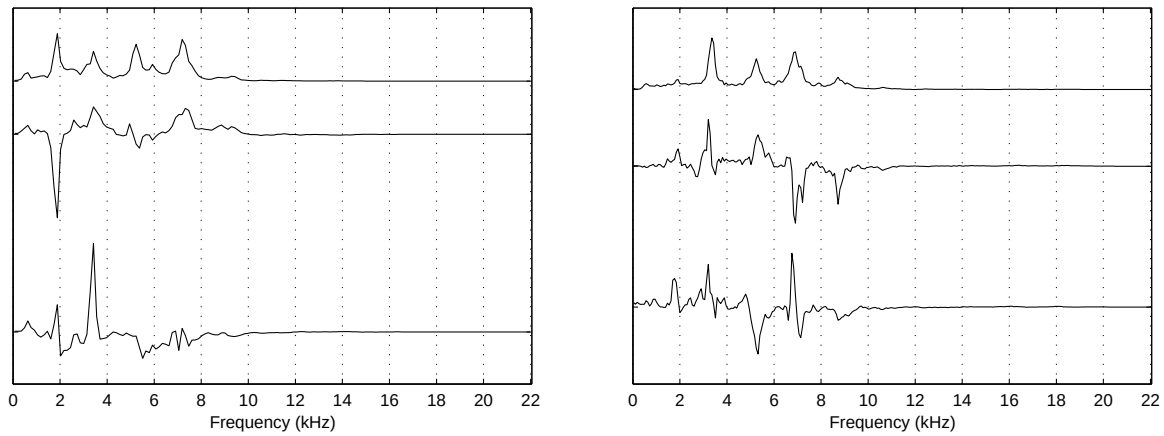


Figure 3.12: Spectral basis functions $\boldsymbol{\Upsilon}$ learned by PCA of the set of ten transients sub-signals from impacts on (a) an aluminum rod, and (b) a steel rod. Each figure shows the first three (most dominant) basis functions from top to bottom.

bins of $\mathbf{Z}^k$ correspond to time of the original signal. Similarly to the ISA method, the transients method next learns the structures in the source signals. The process is the same as explained in section 2.2 for model M_b and in appendix A.2, 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 step are vectors in the time space, that is, they are temporal basis functions, which we call $\boldsymbol{\Gamma}$. As an example, figure 3.13 shows temporal basis functions learned by PCA. Because they characterize the temporal structure of transients, these basis functions are always very sharp; they have very sudden increases of energy and fast decays. These characteristics are also found in the temporal basis functions $\boldsymbol{\Phi}$ from section 2.5 that characterized the attack structures in the sounds. Associated with each temporal basis function and each transients sub-signal, there is a coefficient that scales the basis function. Here, $\mathbf{O}$ is the set of coefficients.

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 ISAS method (with model M_b). Since the frames of $\mathbf{Z}^k$ correspond to frequencies of the original signal, and the bins correspond to time, the transient

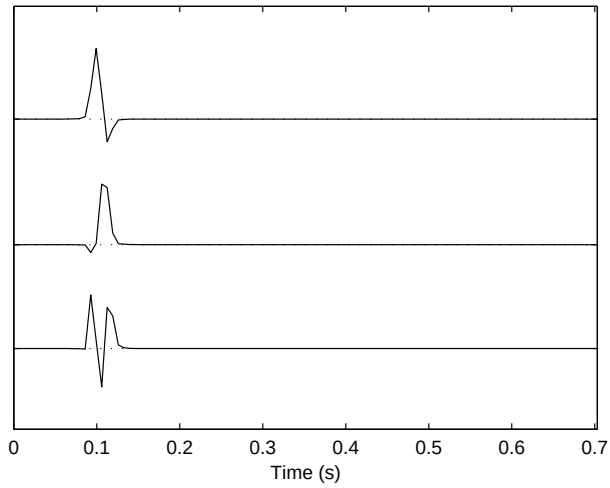


Figure 3.13: Temporal basis functions $\mathbf{\Gamma}$ learned by PCA of the temporal source signals, which in turn were obtained by PCA of the set of ten transients sub-signals from impacts on a steel rod. The first three (most dominant) basis functions are shown from top to bottom.

method first obtains a set of spectral basis functions $\mathbf{\Upsilon}$, and a set of temporal source signals (with spectral analysis of the ensemble of spectrograms). Afterwards, it learns a set of temporal basis functions $\mathbf{\Gamma}$, and a set of coefficients $\mathbf{O}$ (by analyzing the temporal source signals obtained in the previous step).

3.5.3 Signal synthesis

Once the analysis part of the ISAS method learns the basis functions that represent the structures in the signals, it passes them to the synthesis part (figure 3.3). The modified sinusoidal synthesis module uses the basis functions and coefficients that represent the sinusoidal sub-signals (that is $\mathbf{\Phi}$, $\mathbf{\Psi}$ and $\mathbf{V}$) to produce a sinusoidal waveform s', while the modified TMS module uses the basis functions and coefficients that represent the transients sub-signals (that is $\mathbf{\Upsilon}$, $\mathbf{\Gamma}$ and $\mathbf{O}$) to produce a transients waveform a'. These waveforms then are combined to produce the final synthesized waveform. For more details on the synthesis process, refer to section 3.3.

By combining waveforms s' and a', the ISAS method generates sounds that contain both transients and sinusoidal portions. The bottom line in figure 3.14 shows such a waveform, y_{ISAS}. 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

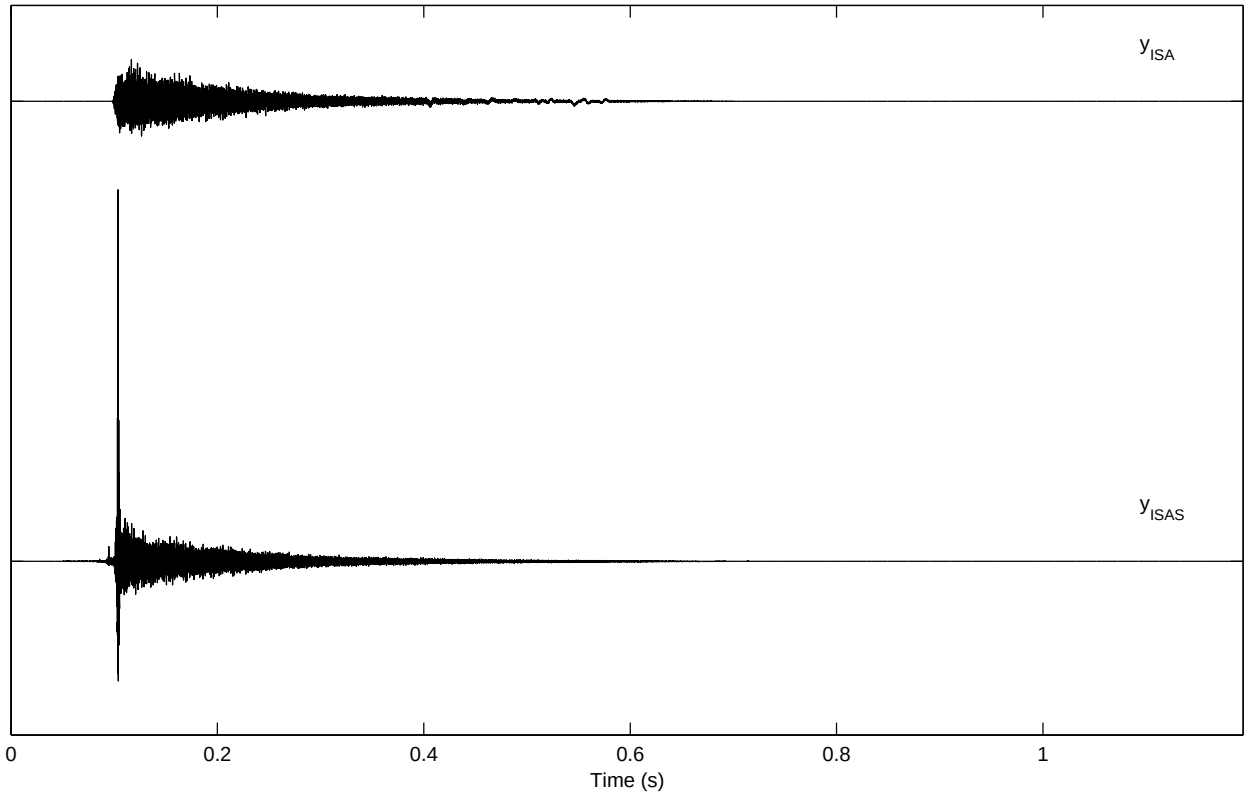


Figure 3.14: Synthesized signals. (Top) y_{ISA} was synthesized by sinusoidal modeling and synthesis and by the ISA method. Here, the sets of basis functions $\mathbf{\Phi}$, $\mathbf{\Psi}$ and coefficients $\mathbf{V}$, were learned by the ISA method (with spectral ICA of the spectrograms of ten sounds from impacts on a zinc plated steel rod, and with PCA of the spectral source signals). These basis functions and the coefficients in $\mathbf{V}^{Zn1}$ were used to synthesize a spectrogram, which was modeled and turned into a waveform by sinusoidal modeling and synthesis. (Bottom) y_{ISAS} was synthesized by the ISAS method. The synthesis modules used the basis functions learned by the analysis part (see figure 3.3) and the coefficients obtained for sound $Zn1$, that is $\mathbf{V}^{Zn1}$ and $\mathbf{O}^{Zn1}$. The same techniques (that is, spectral ICA of the spectrograms and PCA of the source signals) and sounds were used as above.

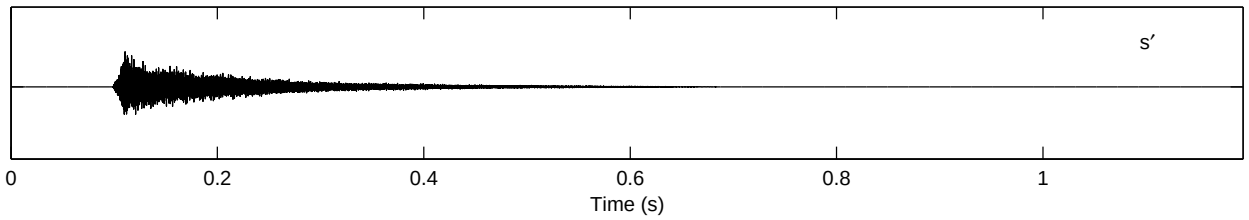


Figure 3.15: Synthesized sinusoids s'. Here, the sets of basis functions $\mathbf{\Phi}$, $\mathbf{\Psi}$ and coefficients $\mathbf{V}$, were learned by the ISA method module of the ISAS method (with spectral ICA of the spectrograms of ten sinusoidal sub-signals from impacts on a zinc plated steel rod, and PCA of the spectral source signals). These basis functions and the coefficients in $\mathbf{V}^{Zn1}$ were used by the modified sinusoidal synthesis module to produce waveform s'.

attack transient is followed by a slower decaying portion, which corresponds to the sinusoidal part of the sound.

As mentioned before in this chapter, the purpose of the ISAS method is to overcome the limitations that the ISA method has when dealing with the synthesis of the transient portions of the sound. 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 3.14 shows a waveform, y_{ISA}, obtained by the ISA method and by sinusoidal modeling and synthesis. This waveform contains the steady, slower decaying, portion of the sound, but lacks the initial sharp and big increase of energy that can be seen in waveform y_{ISAS}. In fact, y_{ISA} is quite similar to s', which is the synthesized sinusoids produced by the ISAS method and which corresponds to the steady portion of y_{ISAS} (figure 3.15). In contrast to the ISA method, and as discussed in the previous paragraphs, the ISAS method can deal with the synthesis of transients.

3.6 Modifying impact sounds

In order to synthesize sounds, the synthesis part of the ISAS method uses the information in the basis functions and coefficients: it uses $\mathbf{\Phi}$, $\mathbf{\Psi}$, and $\mathbf{V}$ to produce a sinusoidal signal s', and $\mathbf{\Upsilon}$, $\mathbf{\Gamma}$, and $\mathbf{O}$ to produce a transients signal a' (figure 3.3). For more details on the synthesis, please refer to section 3.3. Optionally, the synthesis part can include modifications to the basis functions and

to the coefficients in order to obtain new sounds. In figure 3.3, we represented the modified sets of basis functions and coefficients by $\mathbf{\Phi}'$, $\mathbf{\Psi}'$, $\mathbf{V}'$, $\mathbf{\Upsilon}'$, $\mathbf{\Gamma}'$, and $\mathbf{O}'$.

Here, we are particularly interested in manipulating the coefficients to obtain new sounds. We will focus on manipulations to the coefficients in $\mathbf{V}$, which refer to the sinusoidal sub-signal, but the same type of manipulations and reasoning can be applied for the coefficients in $\mathbf{O}$, which are associated to the transients sub-signal. By changing the values of the coefficients in $\mathbf{V}$, we can use different combinations of the temporal and spectral structures represented by the basis functions in $\mathbf{\Phi}$ and $\mathbf{\Psi}$ to obtain different synthesized sinusoidal signals s', and, consequently, different synthesized signals y (see figure 3.3).

An example was given in section 2.5.4. There, we used the (original) coefficients obtained for sound *Al*4, that is, $\mathbf{V}^{Al4}$, except for those associated with the basis functions in $\mathbf{\Psi}^c$ and $\mathbf{\Psi}^e$. New values for these coefficients were sampled randomly from the original coefficients' distribution, that is, from the distribution of original coefficients $\mathbf{v}_c^k$ and $\mathbf{v}_c^k$, respectively. As a result the weightings of the spectral basis functions in $\mathbf{\Psi}^c$ and $\mathbf{\Psi}^e$ were varied, and consequently the weightings of the temporal basis functions $\boldsymbol{\phi}_c$ and $\boldsymbol{\phi}_e$ were also varied (see equation 2.12). In this way, the temporal behavior of the frequency partials represented by $\mathbf{\Psi}^c$ and $\mathbf{\Psi}^e$ was altered (figure 2.17b).

The values of the coefficients can be changed with or without constraints. If the goal is just to obtain interesting, but not necessarily realistic, sounds, one can change these values arbitrarily. However, when the goal is to obtain realistic sounds, one should be careful to choose values that fall within the distribution of the original coefficients. When the basis functions are not orthogonal (which can happen when ICA is used), choosing appropriate values for the coefficients can prove difficult. A simple and interesting way of choosing these values, is to interpolate the coefficients from different sounds. For example, given sound k_1 and sound k_2, with coefficients $\mathbf{V}^{k_1}$ and $\mathbf{V}^{k_2}$, respectively, we can generate a new set of coefficients $\mathbf{V}^y$ by interpolating the values in $\mathbf{V}^{k_1}$ and $\mathbf{V}^{k_2}$. Then, by combining $\mathbf{V}^y$ with $\mathbf{\Phi}$ and $\mathbf{\Psi}$, we can generate a new spectrogram $\mathbf{S}$ and synthesize the sound it represents, s' (see section 3.3 for more details). In this way, the synthesized sound s' is an interpolation of (the sinusoidal parts of) sounds k_1 and k_2. In section 4.2, we will see that this simple algorithm can generate realistic sounds. That section describes a user study that tests the validity of sounds obtained in this way, that is, it tests if these signals sound realistic.

3.7 Summary

In this chapter, the method that was discussed in chapter 2 was extended into the ISAS method, such that the learned representations of the structures can be used to synthesize both the sinusoidal and transient parts in the sounds. This extended method, divides the signals into sinusoidal and transients sub-signals, and analyses and synthesizes these sub-signals in different ways. The results obtained for the analysis of the sub-signals were shown in section 3.5. That section also discussed the synthesis results, showing that the sounds synthesized by the ISAS method contain both transients and steady portions. Yet, in order to validate the method we still need to verify that these sounds are plausible. In chapter 4 we report tests and user studies that show that these sounds actually are perceived as real.

Chapter 4

Listening Tests

In previous sections we saw how to decompose the sounds into basis functions that represent their spectral and temporal structure, and how to use those basis functions to generate synthesized sounds. One important aspect of this research is the validity of these synthesized sounds. Do the synthesized sounds sound real or do they lack some important properties (present in real sounds) that allow us (humans) to distinguish them as synthesized?

This chapter describes some tests that evaluate the realism of the synthesized sounds. The first is a small informal test that validates the sub-signal extraction module of the ISAS method (section 4.1). This was designed to analyze whether the subdivision of the original signals into the sinusoidal and transients sub-signals causes loss of important acoustic information. The second test is a more thorough user study that tests if the sounds synthesized by the ISAS method are perceived as real (section 4.2). The last test is a user study designed to determine the number of temporal basis functions in $\mathbf{\Phi}$ needed to synthesize the sounds (section 4.3).

4.1 Validation of sub-signal extraction

This section describes a small test that validates the sub-signal extraction (performed by the pre-process signal module of the ISAS method, see figure 3.3). Here we want to ensure that when the original signals x are decomposed into the sinusoidal and transients sub-signals, important properties of the sounds that make them *sound real*, are not lost. In order to isolate the effects of this module, here we did not use the remaining modules of the method.

There were 5 sounds presented to 17 participants, who heard them through headphones in their

personal computers. They could hear the sounds as often as they wished and in any order they wanted. Also, they could hear them in sequence or go back and forth between sounds and compare them. Their task was to classify each sound either as *real* or *synthesized*. Below we give more details about these sounds and describe the results.

Stimuli

The sounds used in this test are real and synthesized impacts on a zinc plated steel rod. Even though the real sounds were recorded under controlled conditions (see section 2.3 for more details), there was some very low level background noise present in the recordings because there was a person inside the chamber to perform the impacts and there was some low level noise due to the air circulation system. This background noise was filtered out for this test.

In order to synthesize the sounds we used the first two approaches described in section 3.4.2, that is approach *ST* and approach *TS*. These two approaches use the same techniques to extract the sinusoidal sub-signal s and the transients sub-signal a from the original signal x.[1] They differ only in the order by which s and a are extracted: the first approach extracts s first (figure 3.4), while the second approach extracts a first. (Please refer to section 3.4.2 for more details)

When the complete ISAS method is used, the final synthesized signal y results from the combination of the synthesized sinusoids s' and the synthesized transients a' (see figure 3.3 and section 3.3). In contrast, here the synthesized signal y results from the combinations of sub-signals s and a: $y(t) = s(t) + a(t)$. Both sub-signal extraction approaches produce residual noise, which is the part of x that is not represented by s and a (see figure 3.4 and equation 3.6). Even though it is possible to model the noise (see section 3.3 for more details) we did not model it because the synthesized signals obtained by additive synthesis of only s and a have already very good quality.

To produce the different sounds used in the test, we used the two sub-signal extraction approaches mentioned above, and varied the FFT size used in the approaches' computation of spectrograms. For some sounds we used a 512-n point FFT, while for other sounds we used a 1024-n point FFT. (The overlap was always half of the FFT size.) When a higher FFT size is used, the sinusoidal sub-signal s has higher spectral resolution (because it results from the synthesis of the spectrogram of the original waveform x) while the transients sub-signal a has a higher temporal resolution (because it results from the synthesis of the spectrogram of the DCT of x, and as was

[1] One of these techniques is sinusoidal modeling and synthesis. Our implementation of this technique was based on the implementation of SMS available in [Ellis, 2003].

seen in section 3.2.4, the bins of the STFT of the DCT corresponds to time).

Even though the sinusoidal sub-signals s obtained with both a 512-n point FFT and a 1024-n point FFT have good quality, the same does not happen with the transients sub-signals a. The sub-signals a obtained using a 512-n point FFT may show some artifacts due to poor temporal resolution. These artifacts are less pronounced or even not present in the sub-signals a obtained with the higher FFT size (and consequently higher temporal resolution). Since the final synthesized signals y inherit the artifacts from the sub-signals a, a 1024-n point FFT seems more appropriate, yet we also used the less convincing sounds obtained with a 512-n point FFT, in order to have some worse examples in the set of tested sounds.

As mentioned before, the test used 5 sounds. There was an original denoised waveform x, which was a real impact on the zinc plated steel rod, and 4 synthesized sounds: $y_{ST,1024}$ was obtained by the sub-signal extraction approach ST, and a 1024-n point FFT; $y_{ST,512}$ was obtained by the same approach and a 512-n point FFT; $y_{TS,1024}$ was obtained by approach TS, and a 1024-n point FFT; and $y_{TS,512}$ was also obtained by approach TS but a 512-n point FFT. For all synthesized signals, s and a were extracted from x.

Results

The results show that many people classify the sounds as real (figure 4.1). Even the sounds that have poorer quality were often classified as real. As mentioned above, we opted to use some less convincing sounds, $y_{ST,512}$ and $y_{TS,512}$ obtained with a 512-n point FFT. Since their attacks, a, have artifacts due to the lower temporal resolution used in the computation of the spectrograms, we expected a poor outcome for these sounds. Yet, even though these two sounds were classified as synthesized more often, the results contradicted our expectations as many people still classified these sounds as real (41% for $y_{ST,512}$ and 47% for $y_{TS,512}$).

Although sound $y_{ST,1024}$ is synthesized, it is classified as real by 71% of the participants. These results are actually very similar to those obtained for the real sound x. As expected, the real sound was classified as real most of the time, but even this sound was classified as synthesized sometimes (more specifically 35% of the times). The results obtained for sound $y_{ST,1024}$ show that with appropriate FFT sizes, and approach ST to extract the sub-signals (which computes s before a), the pre-process signal module is able to process the sounds while allowing them to keep the properties that make them sound real.

The results obtained with sub-signal extraction approach ST (results for sounds $y_{ST,1024}$ and

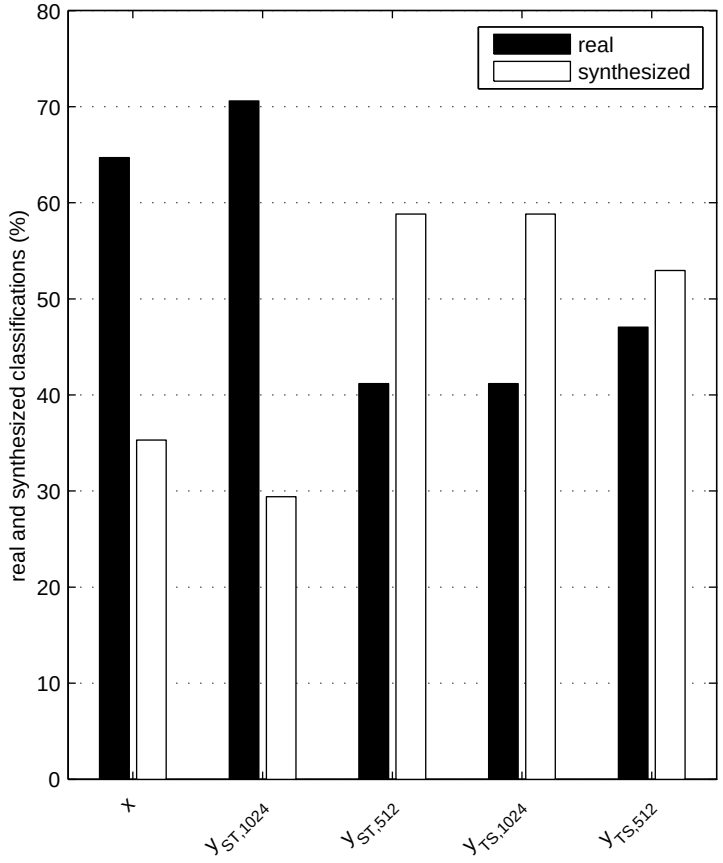


Figure 4.1: Classification of the sounds. The vertical axis shows the percentage of times sounds are classified as real or synthesized.

$y_{ST,512}$) were better than the results obtained with approach TS (results for sounds $y_{TS,1024}$ and $y_{TS,512}$). Since the difference between these two approaches is the order by which the sub-signals s and a are extracted, the results lead us to conclude that for this type of sounds, it is more appropriate to extract the sinusoidal sub-signal s before the transients sub-signal a. Possibly, when a is extracted first (by approach TS) some of the energy of the sinusoidal part of the original signal ends up represented in a, which may affect the quality of the final synthesized signal y. Also when approach ST is used, better results were obtained when a higher temporal resolution was used in the spectrograms from which the transients sub-signal is extracted.

To conclude we would like to mention that many people reported that it was very difficult to classify the sounds because they were very similar. Their first impression was that the sounds were the same, and they reported having listened to the sounds many times. The results of this test

show that the original signals x can be decomposed into the sinusoidal and transients sub-signals without losing important properties of the sounds that make them sound real.

4.2 Validation of the ISAS method: Do the synthesized sounds sound real?

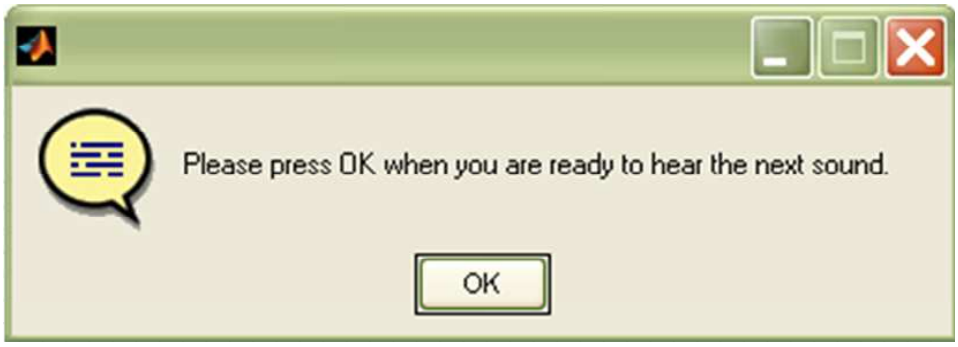
While the previous test validates the sub-signal extraction method, it does not validate the complete ISAS method. For that we did a more thorough user study that uses sounds generated by the ISAS method, that is, both the analysis and synthesis parts of the ISAS method were used here (figure 3.3). The goal of this user study is to determine if the sounds produced by the ISAS method are realistic (that is, if the synthesized sounds are perceived as real sounds). Briefly, in this user study, subjects were presented several real impact sounds as well as impacts synthesized by the ISAS method. For each sound they were asked if the sound was real or synthesized. Below we give more details on the stimuli, protocol and results of this user study.

Protocol

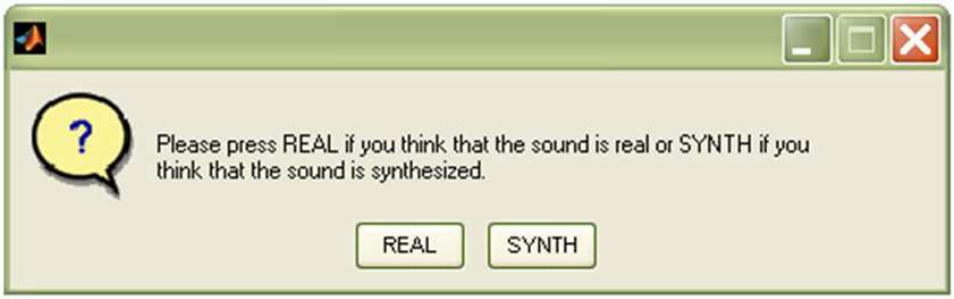
The same sounds were used for all subjects, and each sound was presented only once per participant. The order of presentation varied: to avoid having presentation order effects, every time the study was run a different random order was assigned to the sounds. In this way, each subject heard the sounds in a different order.

The test was performed on a computer and the sounds were heard through headphones. Subjects were explained that they would be presented real and synthesized sounds from impacts on rods, and that they had to classify each sound either as *real* or *synthesized*. In order to familiarize the subjects with the process and the type of sounds used, there were 6 training trials. Subjects could adjust the sound level to a comfortable setting at that time. No information was given as whether the sounds in the training trials were real or synthesized.

In order to avoid having sounds played when the subjects were distracted or not expecting them, before each sound was presented, a dialog box popped up asking the subject to press a button when he/she was ready to hear the next sound (figure 4.2a). Once the sound had been played, another dialog box popped up asking the subject to classify the sound as real or synthesized (figure 4.2b). Since this is a forced choice study, subjects were instructed to make their best guess whenever in doubt about the type of the sound.



(a)



(b)

Figure 4.2: Main dialog boxes used in the user study. (a) Sounds were played only when subjects were ready to hear them. (b) Subjects classified the sounds either as real or synthesized. They did not have a neutral option.

The subjects heard 210 sounds, each lasting less than 2 seconds. Hearing all the sounds takes less than 7 minutes. However, since subjects take a few seconds to decide what button to press (REAL or SYNTH), they would spend around 20 minutes doing the study (of course this time was subject dependent). In order to avoid subject tiredness effects, participants were able to take rest intervals in between any 2 sounds. They could rest whenever they wanted and they were reminded twice during the study (after 1/3 and 2/3 of the sounds had been played) that they could take a rest interval. The intervals were optional, and we observed that some subjects took the rest intervals to relax while others did not did not take any intervals.

Stimuli

The sounds used in this study are organized in a few different sets. The first set, which we call X, consists of 30 real sounds, that is, 30 original waveforms, from which background noise was removed. This set contains 10 impacts on an aluminum rod, 10 impacts on a steel rod, and 10

impacts on a zinc plated steel rod. All rods have the same length and diameter.

The following four sets, which we will refer to as sets $Y_{\mathbf{\Phi}-,\mathbf{V}-}$, contain sounds synthesized by the (complete) ISAS method. To create the sounds for each set, the ISAS method used either spectral ICA or spectral PCA to learn the set of temporal basis functions $\mathbf{\Phi}$, but always used PCA to learn the set of spectral basis functions $\mathbf{\Psi}$ (PCA was used in this step because the results of chapter 2 show that this technique gives less redundant results than when ICA is used to learn $\mathbf{\Psi}$). Also, for each of the 3 rods mentioned above, $\mathbf{\Phi}$ and $\mathbf{\Psi}$ were learned from a set of 10 recordings from impacts on that rod (in this way we obtain different basis functions $\mathbf{\Phi}$ and $\mathbf{\Psi}$ for each rod). The ISAS method generated the sounds using these sets of basis functions and either the *original* coefficients in $\mathbf{V}$, or *interpolated* coefficients (where $\mathbf{V} = \{\mathbf{V}^1, \mathbf{V}^2, \ldots, \mathbf{V}^{10}\}$.) In summary, the techniques used to synthesize the sounds in each set differ in the way the coefficients $\mathbf{V}$ are used and the way the temporal basis functions in $\mathbf{\Phi}$ are learned. Below we give the specific details of each of these four sets of synthesized sounds (namely, $Y_{\mathbf{\Phi}-ICA,\mathbf{V}-original}$, $Y_{\mathbf{\Phi}-PCA,\mathbf{V}-original}$, $Y_{\mathbf{\Phi}-ICA,\mathbf{V}-interpolated}$, and $Y_{\mathbf{\Phi}-PCA,\mathbf{V}-interpolated}$).

Set $Y_{\mathbf{\Phi}-ICA,\mathbf{V}-original}$ contains 15 sounds obtained by the ISAS method with $\mathbf{\Phi}$ learned by spectral ICA, and using the *original* coefficients in $\mathbf{V}$: the method synthesized 5 sounds with the basis functions and coefficients obtained for each rod. Only one of the sets in $\mathbf{V}$ was used for each of the generated sounds. So, for example, to generate sound y_k, the method used the coefficients in $\mathbf{V}^k$. Set $Y_{\mathbf{\Phi}-PCA,\mathbf{V}-original}$ contains 15 sounds produced by the ISAS method in the same way, but with $\mathbf{\Phi}$ learned by spectral PCA.

As mentioned in section 3.6, in order to obtain new sounds, the coefficients in $\mathbf{V}$ can be manipulated, and one way of doing this is by interpolating them. Set $Y_{\mathbf{\Phi}-ICA,\mathbf{V}-interpolated}$ contains 45 sounds obtained by the ISAS method with $\mathbf{\Phi}$ learned by spectral ICA, and using *interpolated* coefficients: the method synthesized 15 sounds with the basis functions and interpolated coefficients obtained for each rod. These coefficients were the result of interpolating pairs of original coefficients in $\mathbf{V}$. For instance, sound y_{k_1,k_2} uses the interpolation of $\mathbf{V}^{k_1}$ and $\mathbf{V}^{k_2}$. Set $Y_{\mathbf{\Phi}-PCA,\mathbf{V}-interpolated}$ contains 45 sounds produced by the ISAS method in the same way, but with $\mathbf{\Phi}$ learned by spectral PCA.

Instead of having separate sets with the sounds produced with the original and interpolated coefficients, we could have merged the two sets that correspond to $\mathbf{\Phi}$ learned with spectral ICA, and analyzed only the results of set $(Y_{\mathbf{\Phi}-ICA,\mathbf{V}-original} \cup Y_{\mathbf{\Phi}-ICA,\mathbf{V}-interpolated})$. However, we

decided to keep the analysis of these sets separate to understand if using interpolated coefficients in the synthesis (when spectral ICA is used) causes different responses from those obtained when the original coefficients are used. The same applies to sets $Y_{\mathbf{\Phi}-PCA,\mathbf{V}-original}$ and $Y_{\mathbf{\Phi}-PCA,\mathbf{V}-interpolated}$. We did not merge these two sets to understand if using the original and interpolated coefficients in the synthesis (when spectral PCA is used) leads to different results in the user study.

Finally, to understand if the subjects were paying attention to the sounds, we wanted to use a set of sounds with some acoustic characteristics that were obviously not those of real impact sounds, but that still sounded like impact sounds. A simple method to synthesize such sounds is to randomly permute the spectral source signals in $\mathbf{C}$ (where $\mathbf{C} = \{\mathbf{C}^1, \mathbf{C}^2, \ldots, \mathbf{C}^{10}\}$). We built a set, which we call $Y_{\mathbf{C}-scrambled}$, with 60 such sounds. To synthesize these sounds we used the following procedure: for each of the 3 rods mentioned above, $\mathbf{\Phi}$ was learned by spectral ICA of a set of 10 recordings from impacts on that rod (we did not use $\mathbf{\Psi}$ here). Then, using these basis functions and randomly combining the spectral source signals in $\mathbf{C}$ we generated 20 sounds. For instance, sound y can be the result of combining the first spectral source signal from sound $Al8$, the second spectral source signal from sound $Al6$, the third from $Al10$, etc. The magnitude spectrogram used by the modified sinusoidal synthesis module (figure 3.3) to generate the sinusoidal part of y is $\mathbf{S} = \boldsymbol{\phi}_1(\mathbf{c}_1^{Al8})^T + \boldsymbol{\phi}_2(\mathbf{c}_2^{Al6})^T + \boldsymbol{\phi}_3(\mathbf{c}_3^{Al10})^T + \ldots$

Since these sounds are generated in a way that is inconsistent with the model, most sounds in set $Y_{\mathbf{C}-srambled}$ have pronounced *unreal* characteristics. While each spectral source signal is within the distribution of the corresponding temporal basis function, there are some combinations of spectral source signals that fall outside the original distribution because not all basis functions are orthogonal. As a consequence, the sounds may have some artifacts that are not found in the original sounds. However, with this procedure, it is still possible to generate plausible combinations of spectral source signals, that is, combinations that fall within the original distribution. Therefore, the set contains also some sounds that are acceptable and that ended up being classified more often as real than as synthesized.

The sets of basis functions mentioned above ($\mathbf{\Phi}$ and $\mathbf{\Psi}$) model the sinusoidal part of the signals, while the transients part is treated separately (figure 3.3). Here we did not perform any manipulations to the transients, and the attack a' of the synthesized sounds is equivalent to the transients sub-signal a. Following the results of section 4.1, in order to extract the sinusoidal and transients sub-signals (s and a, respectively) we used the sub-signal extraction approach ST (figure 3.4).

While good quality synthesized sinusoidal signals s' are obtained with both a 512-n point FFT and a 1024-n point FFT, the synthesized transients signals a' sound better when a 1024-n point FFT is used. (Like noted in section 4.1 for the sub-signals a, the synthesized signals a' obtained with a lower temporal resolution may show some artifacts that are less pronounced or even not present in the sub-signals obtained with a 1024-n point FFT.) For that reason, the sounds used in this study were obtained with different resolutions in the analysis of sub-signals s and a: the ISA method module used a 512-n point FFT, while the transients method module used a 1024-n point FFT. (The overlap was always half of the FFT sizes.) These values were used for all sounds, and even though some sounds in sets $Y_{\mathbf{\Phi}-,\mathbf{V}-}$ could have benefited from some individual tuning, we did not try to improve individual sounds by adjusting the parameters.

Results

There were 12 Carnegie Mellon University students, with ages ranging between 19 and 33, participating in this user study. There were 4 women and 8 men, but we noticed no gender effects. The subjects reported having normal hearing.

Figure 4.3 shows the average results for each set of sounds, that is, the vertical axis represents the average number of times sounds are classified as real (for the dark columns) or synthesized (for the white columns), normalized by the number of subjects. For example, the dark column of set X shows that, on average, the sounds in this set are classified as real by 72% of the subjects.

The sounds in sets $Y_{\mathbf{\Phi}-,\mathbf{V}-}$, are classified as real most of the times, that is, on average between 64% and 67% of the subjects classify these sounds as real. When we analyze each sound individually, the same conclusion holds: almost every sound in sets $Y_{\mathbf{\Phi}-,\mathbf{V}-}$ is classified more often as real than as synthesized. There are only 4 sounds (out of the 120 sounds in sets $Y_{\mathbf{\Phi}-,\mathbf{V}-}$) that have been classified more often as synthesized than as real. As mentioned before we did not try to improve individual sounds by adjusting the synthesis parameters, and perhaps these 4 sounds could have benefited from individual tuning. Yet, it is remarkable that without individual tuning, only 4 sounds had poorer results.

By itself, the fact that the sounds in sets $Y_{\mathbf{\Phi}-,\mathbf{V}-}$ on average are classified most of the time as real demonstrates that these sounds are *realistic*. This conclusion is even stronger when we compare the results obtained for these sets to the results obtained for set X, which is the set of real sounds. On average, the sounds in set X are classified as real by 72% of the subjects; even these sounds, which are actually real, are also classified as synthesized sometimes. The average values for set X are very

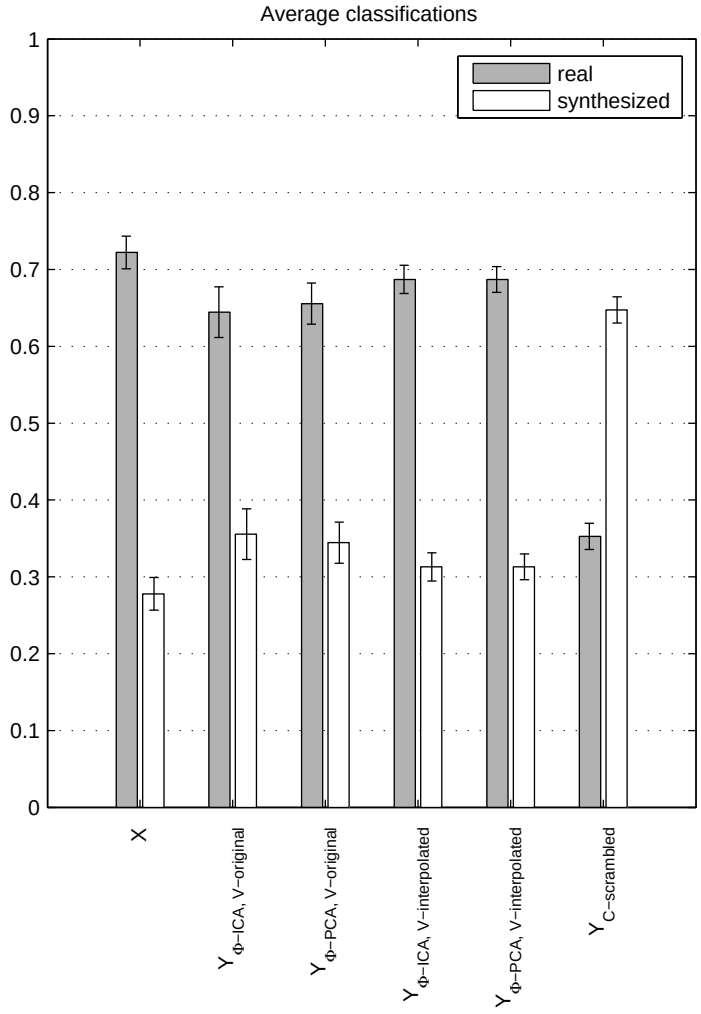


Figure 4.3: Average number of times the sounds are classified as real or synthesized, normalized by the number of subjects. The error bars show the standard error of the mean.

close to those of sets $Y_{\mathbf{\Phi}-,\mathbf{V}-}$, and as further analysis shows, there may be no statistically significant difference between the results of set X and sets $Y_{\mathbf{\Phi}-,\mathbf{V}-}$. Since the data are not continuous, here we used a χ^2 test (with a Yates correction)[2] to compare the results of set X to each of the sets $Y_{\mathbf{\Phi}-,\mathbf{V}-}$, $(Y_{\mathbf{\Phi}-ICA,\mathbf{V}-original} \cup Y_{\mathbf{\Phi}-ICA,\mathbf{V}-interpolated})$, and $(Y_{\mathbf{\Phi}-PCA,\mathbf{V}-original} \cup Y_{\mathbf{\Phi}-PCA,\mathbf{V}-interpolated})$ [Bech and Zacharov, 2006, DeGroot and Schervish, 2002]. The χ^2 values obtained do not exceed the critical value of 3.84 for $p = 0.05$, that is, $p > 0.05$ (a p-value of 0.05 or less denotes a statistically significant effect). Therefore, we cannot conclude that there is a statistically significant effect, i.e.,

[2]We used a Yates correction in every χ^2 test mentioned in this section because all the tests performed here have only one degree of freedom.

we cannot conclude that the real sounds (from set X) and synthesized sounds (from sets $Y_{\mathbf{\Phi}-,\mathbf{V}-}$) are classified in different ways. In other words, the results of the χ^2 test lead us to not reject the hypothesis that real and synthesized sounds are classified in the same way.

Another observation that can be made is that we do not need to use the original coefficients in $\mathbf{V}$. We can use other coefficients, provided they fall within the distribution of the original coefficients. One way of obtaining such coefficients is through interpolation (as in sets $Y_{\mathbf{\Phi}-ICA,\mathbf{V}-interpolated}$ and $Y_{\mathbf{\Phi}-PCA,\mathbf{V}-interpolated}$). A χ^2 test on sets $Y_{\mathbf{\Phi}-ICA,\mathbf{V}-original}$ and $Y_{\mathbf{\Phi}-ICA,\mathbf{V}-interpolated}$ gives $p > 0.30$ and a χ^2 test on sets $Y_{\mathbf{\Phi}-PCA,\mathbf{V}-original}$ and $Y_{\mathbf{\Phi}-PCA,\mathbf{V}-interpolated}$ gives $p > 0.40$, which do not indicate any statistically significant effect between the sounds obtained with interpolated and original coefficients.

So far we have not looked into the results obtained for set $Y_{\mathbf{C}-scrambled}$. These are very different from the results obtained for the other sets. As figure 4.3 shows, the sounds in set $Y_{\mathbf{C}-scrambled}$, on average are classified as synthesized by 65% of the participants, which shows that most of these sounds are not realistic or at least not as realistic as the sounds in the other sets. This is the only set for which the χ^2 test indicates a statistically significant effect ($p < 0.005$).

Music versus no music knowledge

There were 8 participants who reported having some knowledge of music and/or acoustics, while the remaining 4 participants reported not having that type of knowledge. We noticed some differences in the results of these two groups. Figure 4.4a shows the average results for the group with music knowledge and figure 4.4b shows the average results for the other group. The vertical axes show the average number of times sounds are classified as real (for the dark columns) or synthesized (for the white columns), normalized by the number of subjects in each group. For example, the dark column of set X in figure 4.4a, shows that, on average, the sounds in this set were classified as real by 65% of the subjects with music knowledge. An interesting difference between these two figures is that the white bars are much lower for the group with no music knowledge. In other words, people with music knowledge tend to classify sounds as synthesized more often, perhaps because they have learned to listen to sounds in a different way (for instance, while learning to play a musical instrument) and as a result may be attentive to more cues in the sounds. A χ^2 test on the answers of these two groups, confirms that there is statistically significant difference in the way the groups classify the sounds ($\chi^2 = 60.96$, $p < 0.005$).

Interestingly, this does not mean that people with music knowledge classify the sounds better.

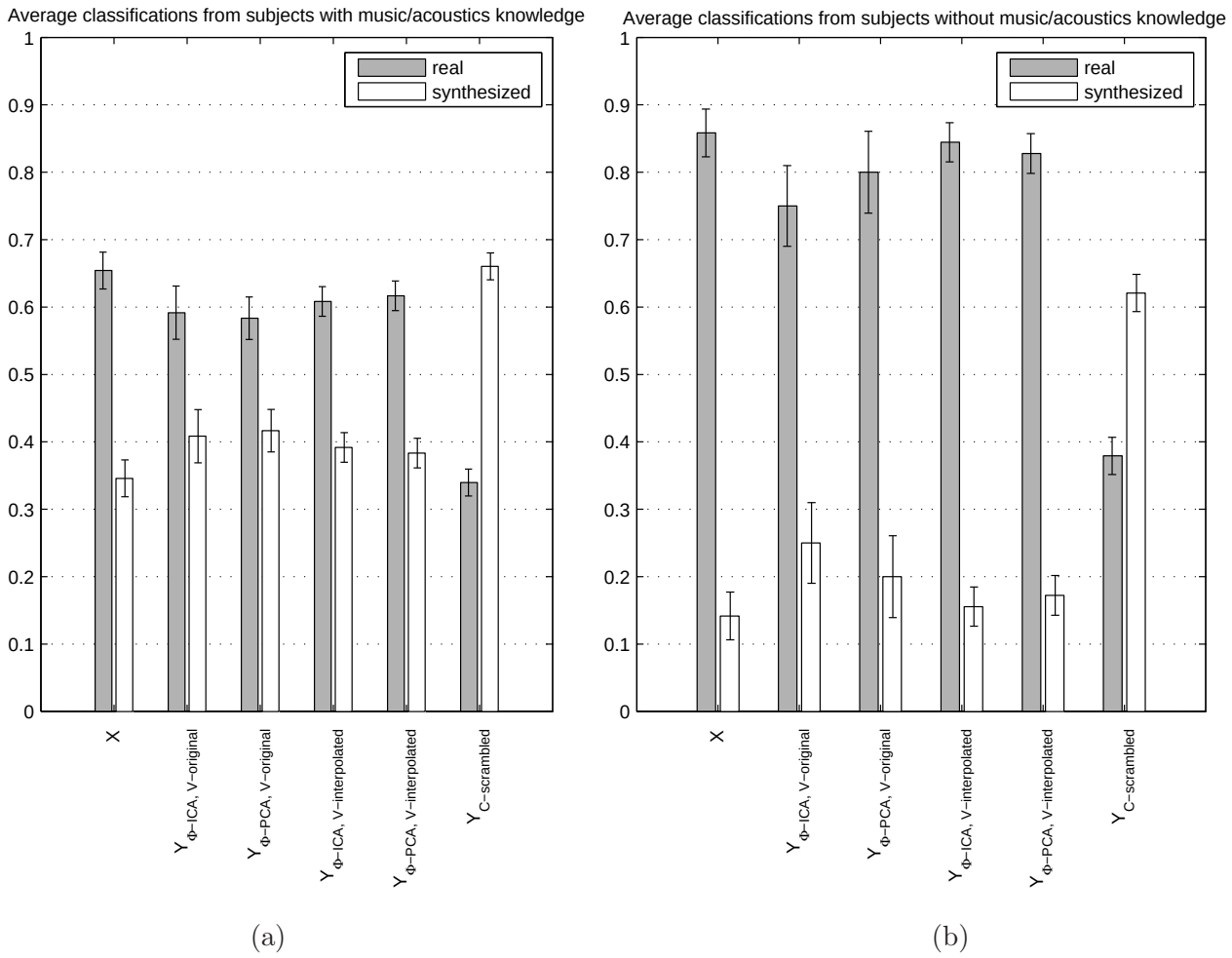


Figure 4.4: Average number of times the sounds are classified as real or synthesized, normalized by the number of subjects in each group. The error bars show the standard error of the mean. (a) Results from the group with music knowledge. (b) Results from the group with no music knowledge.

As it can be observed in figure 4.4a, this group tends to classify more synthesized sounds as synthesized, yet they also tend to classify more real sounds as synthesized. We analyzed the results of this group further to check if there was any difference between the classification of the real sounds (set X) and the other sounds. The results were similar to those obtained before (for all participants). The χ^2 values are smaller than the tabulated value 2.71 for $p = 0.10$. Since $p > 0.10$ we cannot conclude that the real sounds (from set X) and synthesized sounds (in sets $Y_{\mathbf{\Phi}-,\mathbf{V}-}$) are classified in different ways by people with music knowledge. In other words, we cannot reject the hypothesis that real and synthesized sounds are classified in the same way by this group.

Analysis by material

Since sounds from different rods have different acoustic properties, an interesting question is if people classify sounds from different rods differently. In other words, do people tend to classify sounds from a certain rod more often as real (or synthesized) than sounds from other rods? In fact it turned out that sounds from the steel rod (on average) were classified more often as real than the sounds from the other rods, while sounds from the zinc plated rod were classified more often as synthesized than the sounds from the other rods.

These differences may be due to the lack of tuning of the synthesis parameters. However, there may be other reasons. When participants were asked if they had used any cues to decide if the sounds were real or synthesized, they reported paying attention to echo, resonance, ringing, duration, power of attack and how power fades away. In particular, some participants mentioned that they assumed that shorter sounds and sounds with no vibration were synthesized. On average, impacts on the zinc plated rod are shorter and have less (or weaker) ringing effects, which may account for the difference in classification between these sounds and the impacts on the other rods.

Since people tend to classify the synthesized impacts on the zinc plated rod more often as synthesized than the impacts on the other rods, we can ask if the impacts on this rod are better classified than other impacts. In other words, is there any difference between the way the participants classified the real and synthesized impacts on this rod? It seems that this is not the case: since the χ^2 values obtained for these impacts do not exceed the tabulated value 2.71 for $p = 0.10$ (i.e., $p > 0.10$) we cannot conclude that the real sounds (from set $X \cap I_{Zn}$) and synthesized sounds (from sets $Y_{\mathbf{\Phi}-,\mathbf{V}-} \cap I_{Zn}$) are classified in different ways, where I_{Zn} is the set of all impacts on the zinc plated rod.

4.3 Number of temporal basis functions in Φ needed to synthesize sounds

As seen in the results of chapters 2 and 3, only a few temporal basis functions in $\mathbf{\Phi}$ are needed to account for most of the variance in the data. This number is especially low when spectral PCA is used to learn $\mathbf{\Phi}$. While in the previous user study we used all basis functions to synthesize the sounds, here we want to determine how many basis functions in $\mathbf{\Phi}$ are actually needed for that purpose (when $\mathbf{\Phi}$ is learned by spectral PCA). For that we did a user study that uses sounds

generated using different subsets of $\mathbf{\Phi}$. Briefly, in this user study, subjects were presented several pairs of synthesized impact sounds, and they were asked if the sounds in each pair were equal or different. Below we give more details on the stimuli, protocol and results of this user study.

Protocol

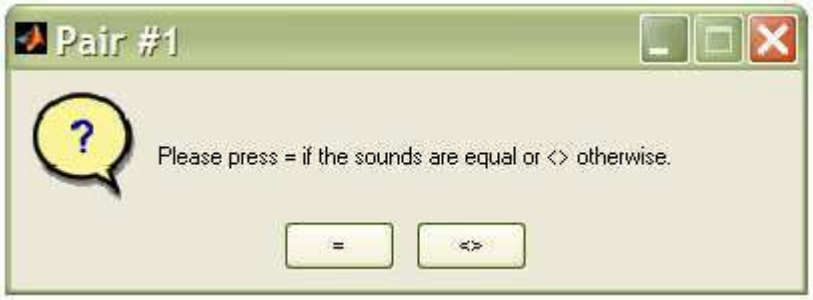
The same pairs of sounds were used for all subjects, and each pair was presented only once per participant. The order of presentation varied: to avoid having presentation order effects, every time the study was run a different random order was assigned to the pairs of sounds, and the order of the sounds within each pair was also varied randomly. In this way, each subject heard the pairs of sounds in a different order.

The test was performed on a computer and the sounds were heard through headphones. Subjects were explained that they would be presented pairs of sounds, and that they had to classify the sounds in each pair either as *equal* or *different*. In order to familiarize the subjects with the process and the type of sounds used, there were 5 training trials. Subjects could adjust the sound level to a comfortable setting at that time. No information was given as whether the pairs of sounds in the training trials were equal or different.

In order to avoid having sounds played when the subjects were distracted or not expecting them, before each pair was presented, a dialog box popped up asking the subject to press a button when he/she was ready to hear the next pair of sounds (figure 4.5a). The two sounds in the pair were played in sequence with a short silence interval (about 1 second) between them. Once the pair had been played, another dialog box popped up asking the subject to classify the sounds as equal or different (figure 4.5b). Since this is a forced choice study, subjects were instructed to make their best guess whenever in doubt about whether the sounds were or were not equal.

The subjects heard 54 pairs of sounds. Since each sound lasts less than 2 seconds, and there is a silence interval of about 1 second between the sounds, hearing all the pairs takes less than 5 minutes. However, since subjects take a few seconds to decide what button to press (= or <>), they would spend around 15 minutes doing the study (of course this time was subject dependent). Even though this was a very short test, in order to avoid subject tiredness effects, participants were able to take rest intervals in between any 2 pairs of sounds.

(a)



(b)

Figure 4.5: Main dialog boxes used in the user study. (a) The pairs of sounds were played only when subjects were ready to hear them. (b) Subjects classified the sounds in each pair either as equal or different. They did not have a neutral option.

Stimuli

To synthesize the sounds used in this user study we used the set of temporal basis functions $\mathbf{\Phi}$ learned by spectral PCA of 10 impacts on the aluminum rod, 10 impacts on the steel rod, and 10 impacts on the wooden rod (that is, we obtained three different sets $\mathbf{\Phi}$, one for each of the three rods used in this user study). In order to isolate the effects obtained by using different subsets of basis functions in $\mathbf{\Phi}$, we considered all spectral basis functions in $\mathbf{\Psi}$ (and the unchanged coefficients in $\mathbf{V}^k$). Since using all basis functions in $\mathbf{\Psi}$ and the unchanged coefficients in $\mathbf{V}^k$ is equivalent to using the whole set of spectral source signals $\mathbf{C}^k$ (see equation 2.10), we actually computed the spectrograms (used in the modified sinusoidal synthesis module - see figure 3.3) with $\mathbf{\Phi}$ and $\mathbf{C}^k$, and did not use the second part of the model (see equation 2.9).

For each of the rods, we used the spectral source signals $\mathbf{C}^k$ of only two impacts (say $\mathbf{C}^{k_1}$ and $\mathbf{C}^{k_2}$), and for each impact (k_1 and k_2) we synthesized 9 sounds: one using $\boldsymbol{\phi}_1$, one using $\boldsymbol{\phi}_1$ to $\boldsymbol{\phi}_2$, one using $\boldsymbol{\phi}_1$ to $\boldsymbol{\phi}_3$, ..., one using $\boldsymbol{\phi}_1$ to $\boldsymbol{\phi}_6$, one using $\boldsymbol{\phi}_1$ to $\boldsymbol{\phi}_8$, one using $\boldsymbol{\phi}_1$ to $\boldsymbol{\phi}_{10}$, and one

using $\mathbf{\Phi}$. The sounds in each pair were always synthesized with the spectral source signals in only one of the sets $\mathbf{C}^{k_1}$ or $\mathbf{C}^{k_2}$, and one of the sounds in the pair was always synthesized with all basis functions in $\mathbf{\Phi}$. So for instance, considering a pair that contains a sound computed with 3 basis functions and the spectral source signals in $\mathbf{C}^{k_1}$, $y_{k_1,\boldsymbol{\phi}_1:\boldsymbol{\phi}_3}$, the other sound in the pair, $y_{k_1,\mathbf{\Phi}}$, is obtained using all basis functions and the same set of spectral source signals. The spectrograms used in the modified sinusoidal synthesis module to generate the sinusoidal parts of $y_{k_1,\boldsymbol{\phi}_1:\boldsymbol{\phi}_3}$ and $y_{k_1,\mathbf{\Phi}}$ are, respectively:

$$(\mathbf{S}^{k_1,\boldsymbol{\phi}_1:\boldsymbol{\phi}_3})^T = \sum_{i=1}^{3} \boldsymbol{\phi}_i (\mathbf{c}_i^{k_1})^T \,,$$

and

$$(\mathbf{S}^{k_1,\mathbf{\Phi}})^T = \mathbf{\Phi}\mathbf{C}^{k_1} \,.$$

The basis functions mentioned above model the sinusoidal part of the signals. They were obtained by the ISA method module, which here used a 512-n point FFT to compute the spectrograms of the sinusoidal sub-signals. Like in the previous user study and following the results of section 4.1, in order to extract the sinusoidal sub-signals s we used approach ST, described in section 3.4.2 and figure 3.4. Here, the transients were not used because in this user study it was not important if the sounds sounded somewhat less realistic, and we wanted to isolate the effects obtained by varying the number of temporal basis functions used in the synthesis. Therefore, the synthesized sounds consisted only of the synthesized sinusoids ($y = s'$).

Results

There were 12 Carnegie Mellon University students, with ages ranging between 19 and 33, participating in this user study. All subjects reported having normal hearing.

Figure 4.6 shows the average number of times subjects classified the sounds in each pair as *equal*. The number of temporal basis functions used to synthesize one of the sounds in the pairs increases from left to right in the abscissa (the other sound in the pairs was always synthesized using all basis functions in $\mathbf{\Phi}$). The last point in the plots (marked with $\mathbf{\Phi}$ in the abscissa) corresponds to the pairs with two equal sounds: ($y_{k,\mathbf{\Phi}}$, $y_{k,\mathbf{\Phi}}$). If there were no random guessing, these values would be 1 (that is, 100%). The average number of times people misclassified these pairs, and which corresponds to random guessing [Levitt, 1971], was: 12.5% for aluminum and steel, and 4.2% for wood.

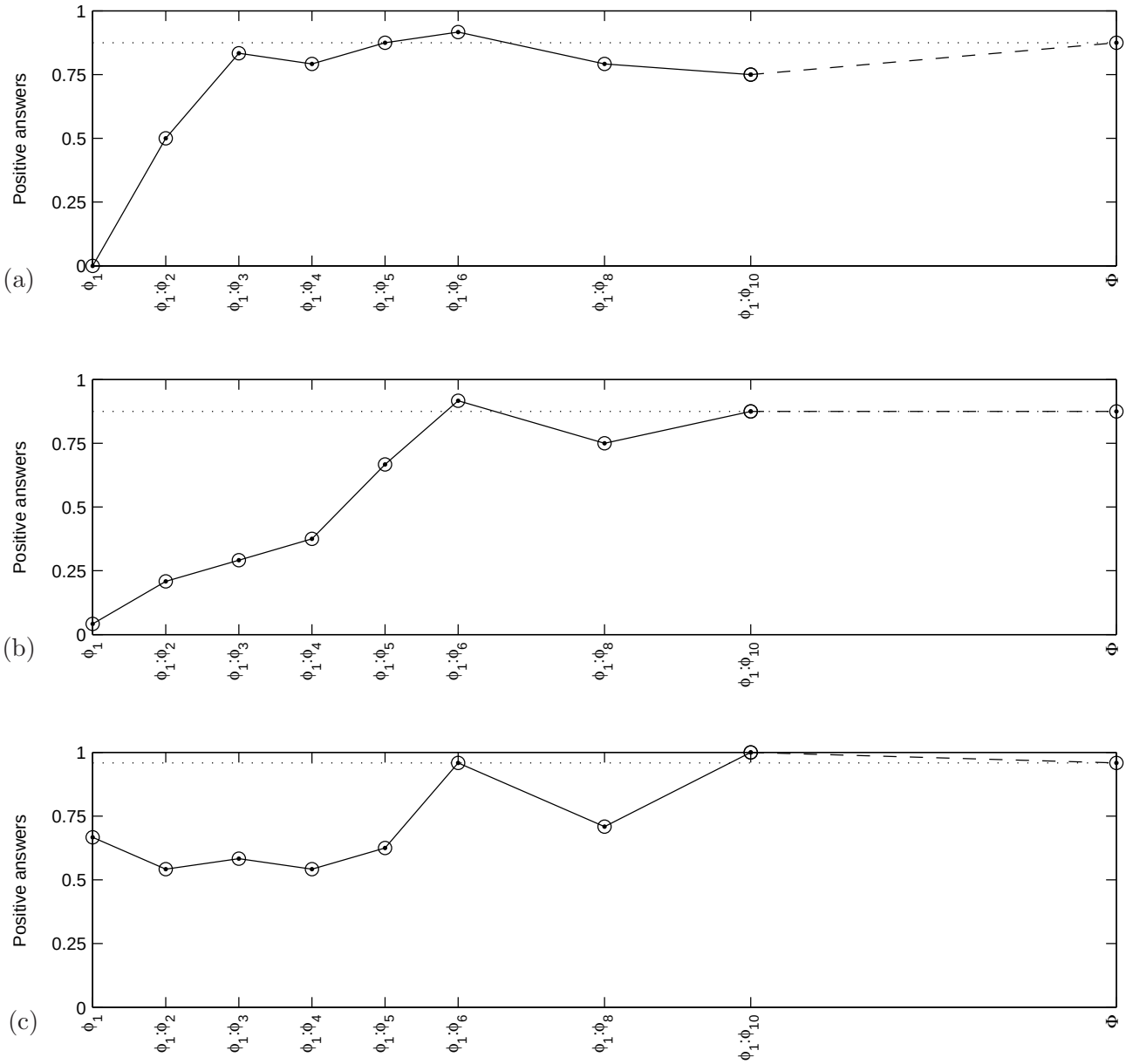


Figure 4.6: Average number of positive answers. The graphs show the average number of times people classified the sounds in each pair as *equal*, normalized by the number of subjects. Each plot corresponds to a different rod: (a) the aluminum rod, (b) the steel rood, and (c) the wooden rod. The abscissa corresponds to the synthesized pairs of sounds, it indicates which temporal basis functions in $\mathbf{\Phi}$ were used in the synthesis of one of the sounds in the pair (the other sound in the pair used all basis functions). Each point in the graphs corresponds to the answers for 2 pairs of sounds, for instance the points with abscissa $\boldsymbol{\phi}_1 : \boldsymbol{\phi}_3$ show the answers for the pairs ($y_{k_1,\boldsymbol{\phi}_1:\boldsymbol{\phi}_3}$, $y_{k_1,\mathbf{\Phi}}$) and ($y_{k_2,\boldsymbol{\phi}_1:\boldsymbol{\phi}_3}$, $y_{k_2,\mathbf{\Phi}}$).

In the first two plots, as the number of temporal basis functions used increases, people tend to classify the sounds in the pairs more often as equal. In other words they perceive less differences between the sound obtained with all basis functions in $\mathbf{\Phi}$, and the other sound in the pair. In the third plot, the average number of times people classify the sounds as equal remains approximately the same until the first 6 temporal basis functions are used. In all three plots, when the first 6 temporal basis functions in $\mathbf{\Phi}$ are used, the average number of times people say the sounds in the pairs ($y_{k,\boldsymbol{\phi}_1:\boldsymbol{\phi}_6}$, $y_{k,\mathbf{\Phi}}$) are equal is the same or higher than for the pairs with the two equal sounds ($y_{k,\mathbf{\Phi}}$, $y_{k,\mathbf{\Phi}}$). That value is almost reached with only 3 temporal basis functions for the aluminum sounds.

In chapters 2 and 3 it was seen that when spectral PCA is used to learn the temporal basis functions $\mathbf{\Phi}$, the first 6 basis functions explain most of the variance in the data (96% of the variance in the original signals, and 98% of the variance in the sinusoidal sub-signals). The results of this user study show that there is no need for the synthesized sounds to account for all the variance in the data, that is, there is no need to use the complete set of temporal basis functions $\mathbf{\Phi}$. More specifically, the results suggest that only the first 6 temporal basis functions are needed to synthesize the sounds, because people tend to not distinguish the sounds made with all temporal basis functions from those made with just 6 of them.

These results show that by considering a low dimensional representation of the data, that is, by considering only a few basis functions, we can obtain perceptually-satisfactory synthesized sounds. While this user study only tests the set of temporal basis functions $\mathbf{\Phi}$ learned by spectral PCA, we predict that not very different results will be obtained if we test the number of basis functions needed when $\mathbf{\Phi}$ is learned by spectral ICA. In that case, we will probably need a few more temporal basis functions $\boldsymbol{\phi}_i$ to describe and synthesize the sounds because we need more basis functions to explain the same percentage of variance of the data (figures 2.13 and 3.9), but we predict that the number of temporal basis functions in $\mathbf{\Phi}$ needed will still be low (maybe around 10). Even though this user study only tests the number of temporal basis functions in $\mathbf{\Phi}$ needed, similar results can be obtained when we test the number of spectral basis functions in $\mathbf{\Psi}$ needed to describe and synthesize the sinusoidal sub-signals, and the number of spectral basis functions in $\mathbf{\Upsilon}$ and temporal basis functions in $\mathbf{\Gamma}$ needed to describe and synthesize the transients sub-signals (because in all these cases, a small number of basis functions can explain most of the variance in the data). Therefore, we can represent the data with only a few parameters, as we only need a few basis

functions in $\mathbf{\Phi}, \mathbf{\Psi}, \mathbf{\Upsilon}$ and $\mathbf{\Gamma}$ to describe and synthesize the sounds.

4.4 Summary

This chapter described some tests that evaluate the synthesized sounds. The first test showed that it is possible to decompose the original signals into sinusoidal and transients sub-signals without losing the properties that make them sound real (section 4.1). Better results were obtained when a higher temporal resolution was used in the spectrograms from which the transients sub-signal is extracted, and with the first extraction approach described in section 3.4.2, that is, approach *ST*.

The second test is a more thorough user study that shows that the sounds synthesized by the ISAS method are realistic (section 4.2). In order to evaluate the complete ISAS method, both its analysis and synthesis parts were used to produce the sounds for this study. Here, we used real impacts on rods as well as impacts synthesized using the original coefficients and interpolated coefficients. The results of this study show that the sounds synthesized by the ISAS method (even when modified - by interpolating the coefficients) sound real, and it shows that we cannot conclude that the real and synthesized sounds are classified in different ways. This study also shows that there is a statistically significant difference between the classifications given by people with music knowledge and those given by people without music knowledge, but that this does not imply that people with music knowledge classify the sounds better.

Finally, the last test was a user study designed to determine the number of temporal basis functions in $\mathbf{\Phi}$ needed to synthesize the sounds (section 4.3). The results show that using a small number of temporal basis functions is enough (like the first 6 basis functions in $\mathbf{\Phi}$) as people tend to not distinguish the sounds made with all temporal basis functions from those made with just 6 of them. These results suggest that we only need a low dimensional representation of the data, i.e. we only need to consider a few basis functions, to obtain perceptually-satisfactory synthesized sounds.

Chapter 5

Conclusions

Natural sounds of the same type have a rich variability in their acoustic structure, but in spite of this variability, when these sounds are heard they are often perceptually very similar. For example, even though the waveforms of different impacts on the same object can be very different, they have some common intrinsic structures that listeners can identify. Our goal here was not to study sound perception, but rather to construct a method that characterizes the structures that are common to sounds of the same type as well as their variability.

We proposed a statistical method for learning a representation of the intrinsic structures of impact sounds, which we referred to as the Intrinsic Structure Analysis (ISA) method (chapter 2). The fact that the method does not require any prior knowledge of the physics, acoustics or dynamics of the objects, can offer advantages over previous knowledge-based models. In addition, the method is not restricted to learning an explicit set of properties of the sounds and it learns the properties that best characterize the statistics of the data. We showed that, by using redundancy reduction techniques, namely PCA and ICA, the method learns basis functions that represent the underlying temporal and spectral structures of the sounds.

The ISA method can be used to characterize the structures of a single sound or those common to a set of impact sounds, in which case it also captures the natural variability in the structures. Obviously, if the ISA method receives different inputs, it produces different outputs, but if the sounds are of the same type, the structures that the method learns are comparable. For instance, the results of analyzing one sound versus several sounds of the same type are very similar.

We have discussed two possible ways of analyzing the spectrograms: spectral analysis and temporal analysis. Also, we have seen that the results obtained by spectral ICA are smoother and

more easily interpretable than the results of temporal ICA, which are noisier and not as easily interpretable. This difference is due to spectral analysis matching the statistics of the data better than temporal analysis. This property applies to all sounds that have the same type of bin and frame joint distribution as the sounds used here. In addition, we have seen that the ISA method can use one of two models: a model M_b that decomposes the spectrograms into spectral source signals and whose temporal basis functions $\mathbf{\Phi}$ are learned by spectral analysis, or a model M_r that decomposes the spectrograms into temporal source signals and whose spectral basis functions $\mathbf{\Theta}$ are learned by temporal analysis (both models have also a second set of basis functions, $\mathbf{\Psi}$ and $\mathbf{\Lambda}$, respectively, learned by ICA or PCA of matrices of source signals). Nonetheless, since spectral analysis is better suited to explain the structures of the spectrograms than temporal analysis, in this dissertation the ISA method uses model M_b, that is it uses spectral analysis of the spectrograms.

Since the method is not restricted to learn explicit features (or structures) of the sounds, the representations obtained include new information that was not represented by previous physical models. For instance, features that are more abstract than simple decay rate or average spectra, like features that characterize ringing, or decay shapes that are not exponential, can now be modeled and easily extracted from the sounds. In particular, spectral ICA is able to decompose spectrograms into a small number of underlying features that characterize acoustic properties such as ringing, resonance, sustain, decay, and onsets.

The ISA method has many applications, such as the study of the intrinsic structures of impact sounds, sound recognition, sound clustering, and sound perception, and although here we have only considered impact sounds, namely impacts on rods, we predict that the method can be used to represent other types of transient acoustic events. The method considers only the spectral (and temporal) content of the signals. Nonetheless, there is also complex structure in the phase of the signals, which is important for synthesizing sound waveforms from the model. Since phase information is not perceptually relevant in the steady state (or periodic) regions of the sounds [Roads, 1996, Ladefoged, 1996], the structures learned by the ISA method can be used to resynthesize the periodic portions of the sounds. On the other hand, phase information is vital for the perception of the attack transients, which are perceptually important portions of the sounds. Without correct phase alignment of the components in a signal, the resynthesis of attack transients is not successful, as the synthesized attacks will sound noisier and less sharp. So, while the model is able to represent the high structure variability in the transients portions of the sounds, it cannot be successfully used

to resynthesize them. In response to this problem, we proposed an extension to the ISA method, which we referred to as the Intrinsic Structure Analysis and Synthesis (ISAS) method, that can be used when the goal of the application includes the synthesis of the sounds (chapter 3).

Because it is an extension of the ISA method, the ISAS method has the same main characteristics as the former method. Furthermore, it can deal with the synthesis of both the periodic and transients portions of the sounds. We have shown that the sounds synthesized by this method are realistic (as they are classified more often as real than as synthesized), and that we cannot conclude that the real sounds used in this dissertation and the sounds synthesized by the ISAS method are classified in different ways (chapter 4). These results are also valid when the sounds are modified (by interpolations) provided that the final synthesized sounds still fall within the distribution of the original sounds.

While the modified sounds reported in this dissertation were obtained by interpolating the coefficients associated to impacts on the same rod and approximately the same position, we are now exploring the synthesis of new sounds obtained with the interpolation of the coefficients associated to impacts on different positions of the rod. Also, while the work presented here used rods of different materials but with the same length and diameter, we are now working with a larger set of sounds that includes impacts on rods of different sizes. For instance, we are interested in exploring how much of the structure of impacts on a certain material varies when the diameter of the rods change.

As mentioned earlier, the methods presented aim for low dimensional representations of the sounds. While the statistical techniques that they use, decompose the sounds into a substantial number of underlying features, it is possible to recover a very good approximation of the initial representation of the sounds with just a few of those features (meaning that we can use these dominant features to represent the sounds in a lower dimensional space). Such low dimensional characterizations of the data can present advantages over previous physical models. For example, since impact sounds can have hundreds of partials [van den Doel et al., 2002], modeling them with equation 2.1 would mean using a very big N. When the objective is to model only the perceptually relevant portions of the sound, much less partials can be used (that is, N can be substantially smaller), yet determining which partials should be used is also a difficult question [van den Doel et al., 2002]. On the other hand, our methods only require a small number of temporal and spectral basis functions to represent a very good approximation of the initial representation of the sounds

and it is easy to determine which are the dominant basis functions. Also, it has been seen that using a small number of temporal basis functions is enough to synthesize the sounds, as people tend to not distinguish the sounds made with all temporal basis functions from those made with just 6 of them, and that similar results are expected for the remaining types of basis functions (namely, the spectral basis functions that represent the sinusoidal sub-signals and the spectral and temporal basis functions that represent the transients sub-signals). Therefore, we only need a low dimensional representation of the data, i.e. we only need to consider a few basis functions, to obtain perceptually-satisfactory synthesized sounds.

The structures learned by the proposed methods are constrained by the initial representations of the data. Using other initial representations may lead to learning structures different from the ones seen in this dissertation. While the Fourier method describes signals as linear combinations of sinusoidal waves, other methods that use amplitude modulated sinusoids or other decomposition functions and that may be better suited to represent transients may be worth considering. An example of such a representation is Prony's method (which was briefly described in section 2.1). Since this method describes the signals as linear combinations of exponentially decaying sinusoidal waves with varying frequency, amplitude, phase and damping factor, it may be suitable to describe sounds with sharp decays, such as impacts. However, such approaches may fail to account for the rich structure variability of real sounds, especially those of the attack transients and those of noise due to the roughness of the surfaces. Another method that bears many similarities to Prony's method and that may also be considered is MP with a dictionary of damped sinusoidal waves [Goodwin, 1997]. Since the sinusoidal atoms can start at any point in time, this method may be better suited to represent transients that can have any starting point in the waveform (section 3.4.1). Additionally, the dictionary of damped sinusoidal waves can be merged with a dictionary of Gabor atoms to combine the advantages of both dictionaries [Goodwin and Vetterli, 1999]. Thus, effects that are not as easily explained by the damped sinusoidal atoms (as sounds that have slow increases in amplitude) are fitted by Gabor atoms. Still, these techniques may fail to fully describe the structure of attacks and noisy portions of the sounds, and they may need to be combined with other (possibly learned) more appropriate atoms. Another alternative is the spike code [Lewicki, 2002b, Smith and Lewicki, 2005, Smith, 2006]. Since this method (with a dictionary of gammatone or learned atoms) gives an efficient representation of both transients and sustain portions of the sound, it may be a better candidate for the initial representation of the data.

To conclude, the methods proposed in this dissertation are statistical (and consequently data-driven) approaches for modeling distributions of related impact sounds. The methods do not require any prior knowledge of the physics, acoustics or dynamics of the objects and are able to learn features that are closely related to the acoustic properties of the sounds. They provide low dimensional representations of the data and at the same time characterize the variability of sounds of the same type. While other data-driven and statistical approaches have been proposed in the literature, there are some fundamental differences between those approaches and the methods proposed here. As seen in section 2.6, there have been more studies that use redundancy reduction techniques and spectrograms or other time-frequency representations, but their essential goals are different. They focus on problems like source separation, event detection, etc., while we focus on the representation of the structures and variability of sounds of the same type. Also, to the best of our knowledge, our approach is the first to partition individual sounds according to the temporal behavior of the partials. Several statistical approaches to model specific parameters of the sounds have been proposed, especially to model parameters that describe the noise portions of the sounds [e.g. Serra and Smith, 1990, Desainte-Catherine and Hanna, 2000]. Also, there are some physical models that use real sounds to fit the parameters of their (knowledge based) equations, such as the amplitudes, frequencies and decay rates of the vibration modes [e.g. van den Doel et al., 2001, Pai et al., 2001]. The advantage of our methods is that they are not tight to specific parameters and they can learn the features that best characterize the statistics of the data. There are also some non-physical methods (many of which were described in section 3.2) that extract their equations' parameters from real sounds. However, these methods were usually developed to analyze, modify and synthesize only one sound at a time. On the contrary, our methods can deal with ensembles of sounds, and while they can be used to represent the structures of a single sound, they were developed to represent the structures common to sounds of the same type.

Appendix A

Models M_r and M_b

This appendix gives a more detailed description of the two models (M_r and M_b) presented in section 2.2. Here we include illustrations of the matrices used in the models and details on matrix rearrangements.

The spectrograms can be defined as an ordered set of T bins or as a sequence of F frames. Section A.1 describes model M_r, which represents the spectrograms as a sequence of frames, while section A.2 describes model M_b, which represents the spectrograms as an ordered set of bins. Model M_r is presented first because this model may be more intuitive to the reader, but if the reader is only interested in model M_b, section A.1 can be safely skipped.

A.1 Model M_r

Model M_r considers the spectrogram of sound k, that is $\mathbf{S}^k$, as a sequence of frames. In order to express the spectrogram in this way, let us assume we have a set of I basis functions $\mathbf{\Theta}$, represented as a matrix of size $(F \times I)$, where each column vector $\boldsymbol{\theta}_i$ is a spectral basis function. Also, let us assume that for each sound k and basis function $\boldsymbol{\theta}_i$ we have a *temporal source signal* $\mathbf{p}_i^k$ that scales basis function $\boldsymbol{\theta}_i$ over time. The set of these signals, $\mathbf{P}^k$, can be represented as a matrix of size $(I \times T)$, where the ith row contains vector $(\mathbf{p}_i^k)^T$.

The linear combination of basis function $\boldsymbol{\theta}_i$ with the corresponding temporal source signal $\mathbf{p}_i^k$ gives a $(F \times T)$ matrix $\mathbf{R}^{k,i}$ that represents part of the structure in $\mathbf{S}^k$. The whole structure in $\mathbf{S}^k$

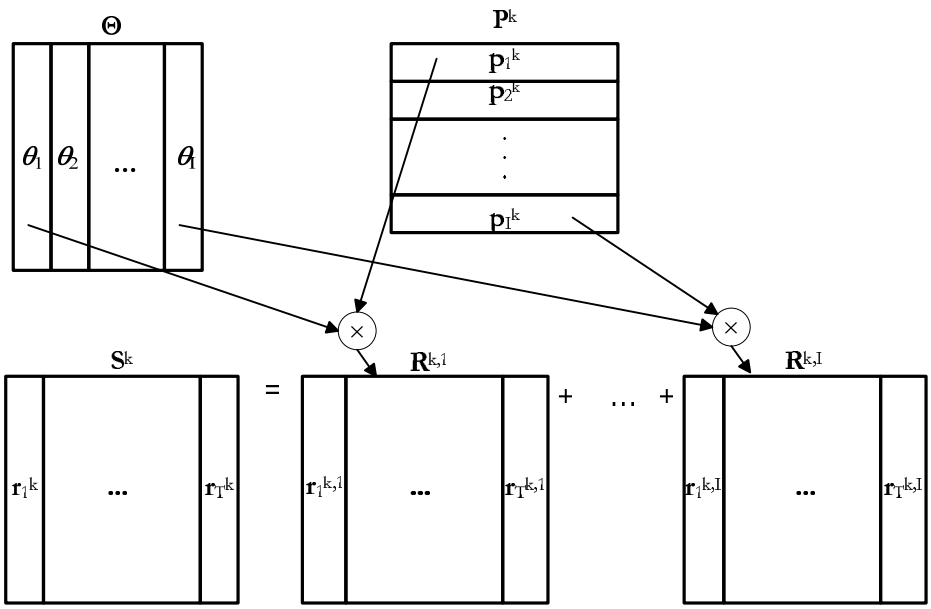


Figure A.1: Illustration of equation A.1. (For simplicity, here we did not mark row vectors with the transpose symbol (T).)

is represented by the whole set of such matrices (see figure A.1):

$$\mathbf{S}^k = \sum_{i=1}^{I} \mathbf{R}^{k,i} = \sum_{i=1}^{I} \boldsymbol{\theta}_i \, (\mathbf{p}_i^k)^T = \boldsymbol{\Theta} \, \mathbf{P}^k \, . \tag{A.1}$$

If we consider only the tth column (frame) in $\mathbf{S}^k$, equation A.1 can be rewritten as:

$$\mathbf{r}_t^k = \sum_{i=1}^{I} \mathbf{r}_t^{k,i} = \sum_{i=1}^{I} \boldsymbol{\theta}_i \, p_{i,t}^k \, . \tag{A.2}$$

where $\mathbf{r}_t^k$ is the tth frame of $\mathbf{S}^k$, $\mathbf{r}_t^{k,i}$ is the tth column of $\mathbf{R}^{k,i}$, and the scalar $p_{i,t}^k$ is the tth value in $\mathbf{p}_i^k$ (that is, its value at time frame t). The basis $\boldsymbol{\Theta}$ can be used to describe a single sound, as in equation A.1, or the spectral regularities of a set of related sounds:

$$(\mathbf{S}^1, \mathbf{S}^2, \ldots, \mathbf{S}^K) = \boldsymbol{\Theta} \, (\mathbf{P}^1, \mathbf{P}^2, \ldots, \mathbf{P}^K) \, . \tag{A.3}$$

So far, the model only describes the spectral structure. Yet, it can be extended to describe the temporal structure inherent in the temporal source signals $\mathbf{p}_i^k$ for an ensemble of related sounds.

We can first construct a new matrix $\mathbf{Q}^i$ of size $(T \times K)$ where each column is the ith temporal source signal for a different sound, that is, $\mathbf{Q}^i = (\mathbf{p}_i^1, \ldots, \mathbf{p}_i^K)$, for K sounds (see figure A.2). In

total there will be I $\mathbf{Q}^i$ matrices, that is, as many $\mathbf{Q}^i$ matrices as spectral basis functions $\boldsymbol{\theta}_i$. Let us call the kth column of matrix $\mathbf{Q}^i$ as $\mathbf{q}^i_k$ (note that $\mathbf{q}^i_k = \mathbf{p}^k_i$).

We will assume we have a set of J basis functions $\boldsymbol{\Lambda}^i$, represented as a matrix of size $(T \times J)$ where each column vector $\boldsymbol{\lambda}^i_j$ is a temporal basis function. Also, let us assume we have a vector of coefficients $\mathbf{g}^i_j$ for each basis function, such that the jth vector scales basis function $\boldsymbol{\lambda}^i_j$ across sounds. The set of these vectors is represented as matrix $\mathbf{G}^i$ of size $(J \times K)$, whose jth row is the vector $(\mathbf{g}^i_j)^T$. The linear combination of a basis function $\boldsymbol{\lambda}^i_j$ with the corresponding vector of coefficients produces a $(T \times K)$ matrix $\mathbf{Q}^{i,j}$ that represents part of the structure in $\mathbf{Q}^i$. The whole structure in $\mathbf{Q}^i$ is obtained by adding all matrices $\mathbf{Q}^{i,j}$ (see figure A.2):

$$\mathbf{Q}^i = \sum_{j=1}^{J} \mathbf{Q}^{i,j} = \sum_{j=1}^{J} \boldsymbol{\lambda}^i_j \, (\mathbf{g}^i_j)^T = \boldsymbol{\Lambda}^i \, \mathbf{G}^i \, . \tag{A.4}$$

If we consider only one column in $\mathbf{Q}^i$, the previous equation can be rewritten as:

$$\mathbf{q}^i_k = \sum_{j=1}^{J} \mathbf{q}^{i,j}_k = \sum_{j=1}^{J} \boldsymbol{\lambda}^i_j \, g^i_{j,k} \, . \tag{A.5}$$

where $\mathbf{q}^i_k$ is the kth column of $\mathbf{Q}^i$, $\mathbf{q}^{i,j}_k$ is the kth column of $\mathbf{Q}^{i,j}$, and the scalar $g^i_{j,k}$ is the value of $\mathbf{g}^i_j$ for sound k. We can now consider the previous equation at a given time frame t and express $q^i_{k,t}$ as follows:

$$q^i_{k,t} = \sum_{j=1}^{J} q^{i,j}_{k,t} = \sum_{j=1}^{J} \lambda^i_{j,t} \, g^i_{j,k} \, , \tag{A.6}$$

where the scalars $q^i_{k,t}$, $q^{i,j}_{k,t}$, and $\lambda^i_{j,t}$ are the values of $\mathbf{q}^i_k$, $\mathbf{q}^{i,j}_k$,and $\boldsymbol{\lambda}^i_j$ at time frame t, respectively.

If instead of considering the rows of $\mathbf{G}^i$, we consider its columns, where $\mathbf{h}^k_i$ is the kth column of $\mathbf{G}^i$, we can define a new matrix $\mathbf{H}^k$ of size $(J \times I)$ with all the vectors that refer to sound k from all $\mathbf{G}^i$ matrices: each column in $\mathbf{H}^k$ is the kth column of a different $\mathbf{G}^i$, that is $\mathbf{H}^k = (\mathbf{h}^k_1, \ldots, \mathbf{h}^k_I)$ (see figure A.3). Considering $h^k_{i,j}$ as the jth value of $\mathbf{h}^k_i$, we have that $h^k_{i,j} = g^i_{j,k}$, and since $p^k_{i,t} = q^i_{k,t}$, we can rewrite equation A.6 as follows:

$$p^k_{i,t} = \sum_{j=1}^{J} \lambda^i_{j,t} \, h^k_{i,j} \, . \tag{A.7}$$

Combining equations A.2 and A.7 it follows that the frames of $\mathbf{S}^k$ can be expressed as:

$$\mathbf{r}^k_t = \sum_{i=1}^{I} \sum_{j=1}^{J} \boldsymbol{\theta}_i \, \lambda^i_{j,t} \, h^k_{i,j} \, . \tag{A.8}$$

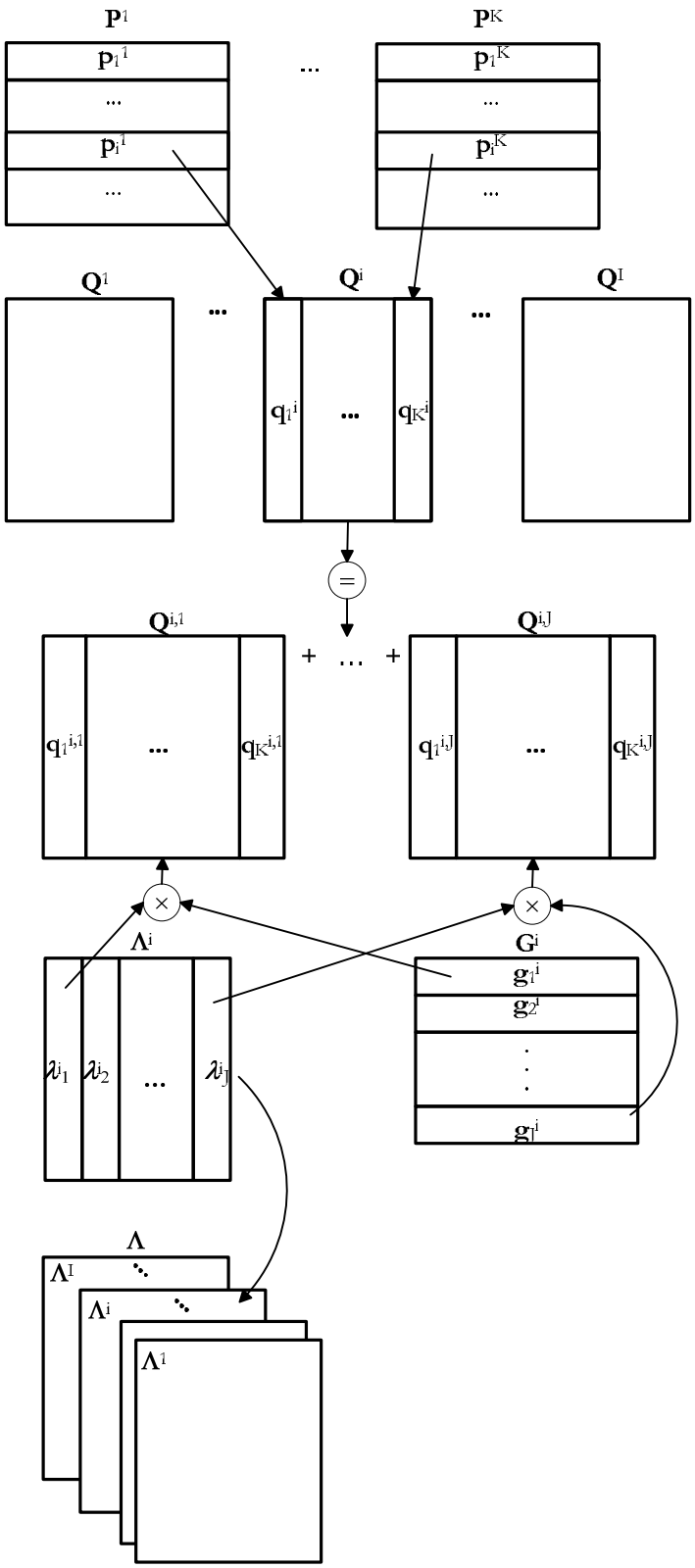


Figure A.2: Illustration of equation A.4. (For simplicity, here we did not mark row vectors with the transpose symbol (T).)

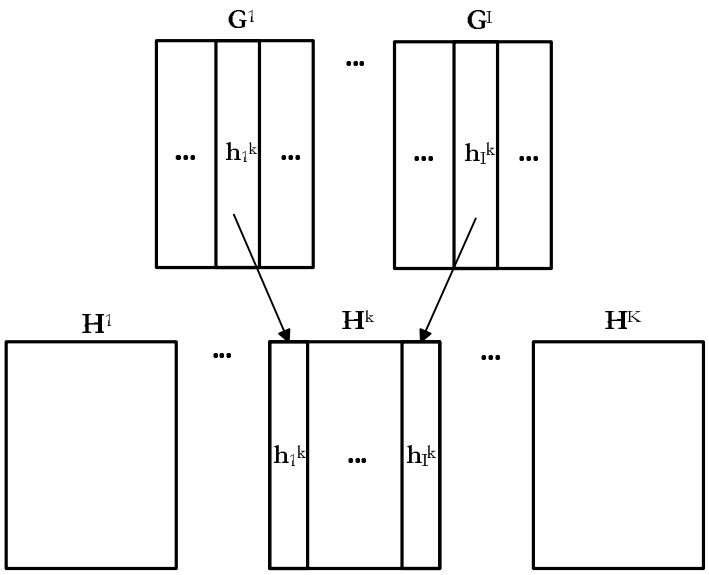


Figure A.3: Illustration of matrixes $\mathbf{H}$.

Thus, $\mathbf{S}^k$ can be modeled by a set of spectral basis functions $\mathbf{\Theta}$, a set of temporal basis functions $\mathbf{\Lambda}$ (where $\mathbf{\Lambda}$ is a three-dimensional matrix of size $(T \times J \times I)$ whose slices are the matrixes $\mathbf{\Lambda}^i$), and a set of coefficients $\mathbf{H}^k$, that is, $\mathbf{S}^k = M_r(\mathbf{\Theta}, \mathbf{\Lambda}, \mathbf{H}^k)$.

A.2 Model M_b

Model M_b considers the spectrogram of sound k, that is $\mathbf{S}^k$, as an ordered set of bins. In order to express the spectrogram in this way, we will assume we have a set of I basis functions $\mathbf{\Phi}$, which is represented as a matrix of size $(T \times I)$ whose column vectors are the temporal basis functions $\boldsymbol{\phi}_i$. We will also assume that for each sound k we have a set of *spectral source signals* represented as a matrix $\mathbf{C}^k$ of size $(I \times F)$, where the ith row contains vector $(\mathbf{c}_i^k)^T$. This spectral signal scales basis function $\boldsymbol{\phi}_i$ across frequencies. (Section 2.5.1 shows how to find $\mathbf{\Phi}$ and $\mathbf{C}^k$.)

By linearly combining a basis function with the corresponding spectral source signals, we obtain a $(T \times F)$ matrix $\mathbf{B}^{k,i}$ that represents part of the structure in $\mathbf{S}^k$. The whole structure in $\mathbf{S}^k$ is obtained by linearly combining matrices $\mathbf{B}^{k,i}$ (see figure A.4):

$$(\mathbf{S}^k)^T = \sum_{i=1}^{I} \mathbf{B}^{k,i} = \sum_{i=1}^{I} \boldsymbol{\phi}_i \, (\mathbf{c}_i^k)^T = \mathbf{\Phi}\, \mathbf{C}^k \,. \tag{A.9}$$

If we consider only the fth column (or bin) in $\mathbf{B}^{k,i}$ (i.e., $\mathbf{b}_f^{k,i}$), equation A.9 can be rewritten

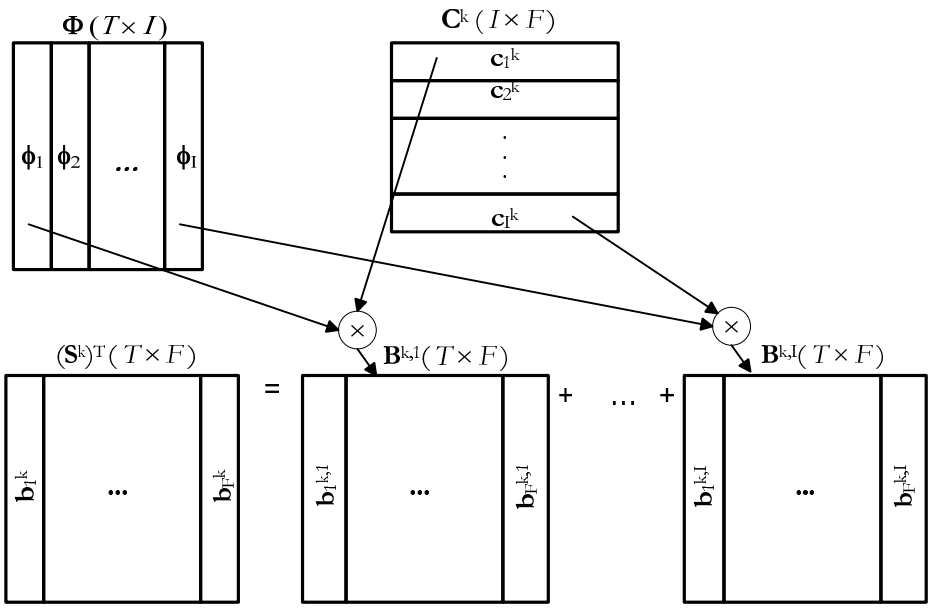


Figure A.4: Illustration of equation A.9. (For simplicity, here we did not mark row vectors with the transpose symbol T.)

as follows:

$$\mathbf{b}_f^k = \sum_{i=1}^{I} \mathbf{b}_f^{k,i} = \sum_{i=1}^{I} \boldsymbol{\phi}_i \, c_{i,f}^k \,. \tag{A.10}$$

where $\mathbf{b}_f^k$ is the transpose of the fth bin of $\mathbf{S}^k$, and the scalar $c_{i,f}^k$ is the value of $\mathbf{c}_i^k$ at frequency bin f. The basis $\boldsymbol{\Phi}$ can be used to describe a single sound, as in equation A.9, or the temporal regularities of a set of related sounds:

$$((\mathbf{S}^1)^T, (\mathbf{S}^2)^T, \ldots, (\mathbf{S}^K)^T) = \boldsymbol{\Phi}\,(\mathbf{C}^1, \mathbf{C}^2, \ldots, \mathbf{C}^K)\,. \tag{A.11}$$

As explained in section 2.5 and appendix B, the temporal basis functions $\boldsymbol{\Phi}$ and the spectral source signals $\mathbf{C}^k$ can be obtained by spectral ICA or PCA. (Note that equation A.9 does not correspond exactly to the implementation of the method, because the implementation uses the extended matrices, as mentioned in section 2.3. See appendix B for exact details on the input and output to the function calls of spectral ICA and PCA.)

This model thus far describes the temporal structure, but not the spectral structure inherent in the spectral source signals $\mathbf{c}_i^k$. We can extend the model to describe the spectral structure in these signals.

We can first construct a new matrix $\mathbf{D}^i$ of size $(F \times K)$ where each column is the ith spectral

source signal for a different sound, that is, $\mathbf{D}^i = (\mathbf{c}_i^1, \ldots, \mathbf{c}_i^K)$, for K sounds. In total there will be I $\mathbf{D}^i$ matrices, one for each temporal basis function $\boldsymbol{\phi}_i$. Let us rename the kth column of $\mathbf{D}^i$ to $\mathbf{d}_k^i$ (note that $\mathbf{d}_k^i = \mathbf{c}_i^k$).

We will assume we have a set of J basis functions $\boldsymbol{\Psi}^i$, represented as a matrix of size $(F \times J)$ where each column vector $\boldsymbol{\psi}_j^i$ is a spectral basis function. Also, we will assume we have a vector of coefficients $\mathbf{u}_j^i$ for each basis function, such that the jth vector scales basis function $\boldsymbol{\psi}_j^i$ across sounds. The set of these vectors, $\mathbf{U}^i$, is represented by a matrix of size $(J \times K)$, where the jth row is the vector $(\mathbf{u}_j^i)^T$. The linear combination of a basis function $\boldsymbol{\psi}_j^i$ with the corresponding vector of coefficients produces an $(F \times K)$ matrix $\mathbf{D}^{i,j}$ that represents part of the structure in $\mathbf{D}^i$. The whole structure in $\mathbf{D}^i$ is a linear combination of matrices $\mathbf{D}^{i,j}$ (see figure A.5):

$$\mathbf{D}^i = \sum_{j=1}^{J} \mathbf{D}^{i,j} = \sum_{j=1}^{J} \boldsymbol{\psi}_j^i \, (\mathbf{u}_j^i)^T = \boldsymbol{\Psi}^i \, \mathbf{U}^i \,. \tag{A.12}$$

As explained in section 2.5 and appendix B, the spectral basis functions $\boldsymbol{\Psi}$ and matrices of coefficients $\mathbf{U}^i$ can be obtained by ICA or PCA.

When we consider only one column in $\mathbf{D}^i$, we can rewrite the previous equation as follow:

$$\mathbf{d}_k^i = \sum_{j=1}^{J} \mathbf{d}_k^{i,j} = \sum_{j=1}^{J} \boldsymbol{\psi}_j^i \, u_{j,k}^i \,, \tag{A.13}$$

where $\mathbf{d}_k^{i,j}$ is the kth column of $\mathbf{D}^{i,j}$, and the scalar $u_{j,k}^i$ is the value of $\mathbf{u}_j^i$ for sound k. We can now consider the equation A.13 at a given frequency bin f and express $d_{k,f}^i$ as follows:

$$d_{k,f}^i = \sum_{j=1}^{J} d_{k,f}^{i,j} = \sum_{j=1}^{J} \psi_{j,f}^i \, u_{j,k}^i \,, \tag{A.14}$$

where the scalars $d_{k,f}^i$, $d_{k,f}^{i,j}$, and $\psi_{j,f}^i$ are, respectively, the values of $\mathbf{d}_k^i$, $\mathbf{d}_k^{i,j}$,and $\boldsymbol{\psi}_j^i$ at frequency bin f.

If instead of considering the rows of $\mathbf{U}^i$, we consider its columns, where $\mathbf{v}_i^k$ is the kth column of $\mathbf{U}^i$, we can define a new matrix $\mathbf{V}^k$ of size $(J \times I)$ with all vectors that refer to sound k from all $\mathbf{U}^i$ matrices: each column in $\mathbf{V}^k$ is the kth column of a different $\mathbf{U}^i$, that is $\mathbf{V}^k = (\mathbf{v}_1^k, \ldots, \mathbf{v}_I^k)$ (see figure A.6). Considering $v_{i,j}^k$ as the jth value of $\mathbf{v}_i^k$, we have that $v_{i,j}^k = u_{j,k}^i$, and since $c_{i,f}^k = d_{k,f}^i$, we can rewrite equation A.14 as follows:

$$c_{i,f}^k = \sum_{j=1}^{J} \psi_{j,f}^i \, v_{i,j}^k \,. \tag{A.15}$$

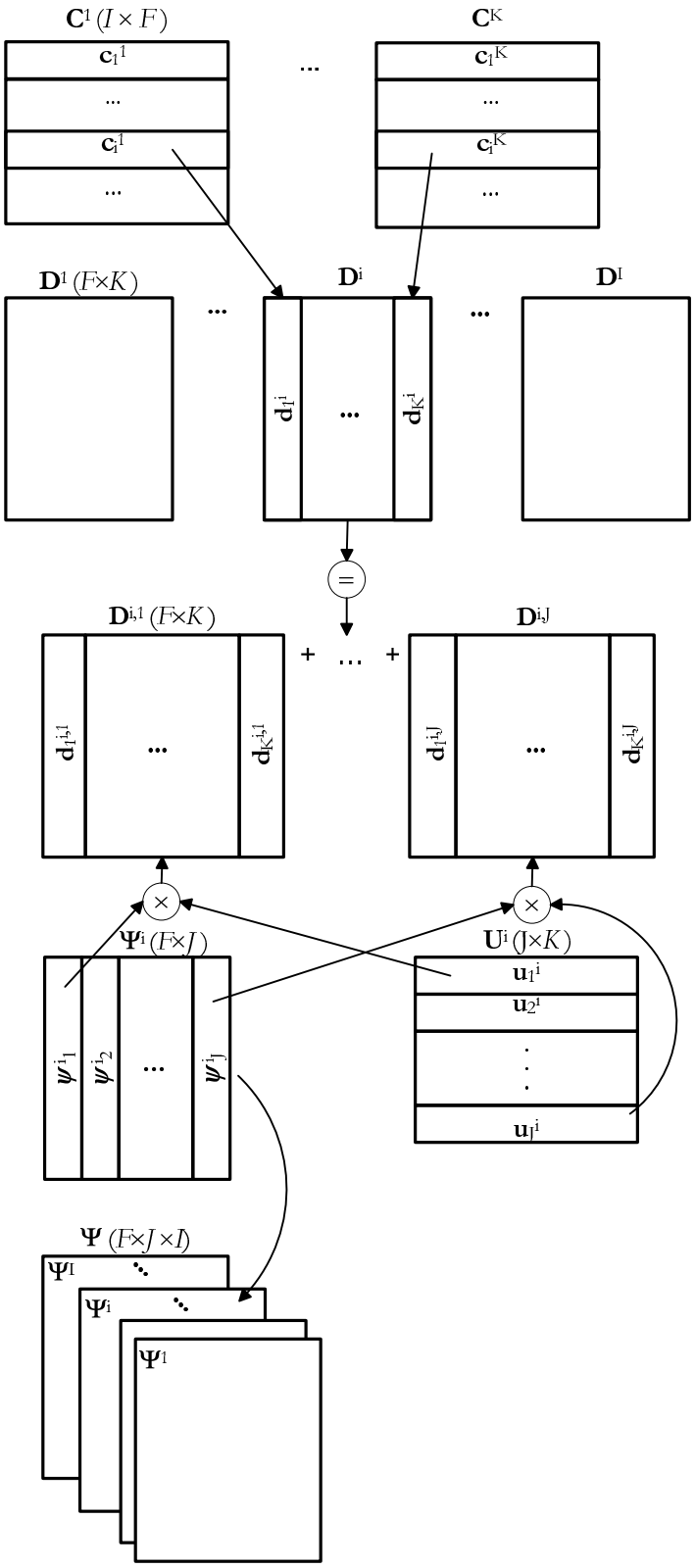


Figure A.5: Illustration of equation A.12. (For simplicity, here we did not mark row vectors with the transpose symbol T.)

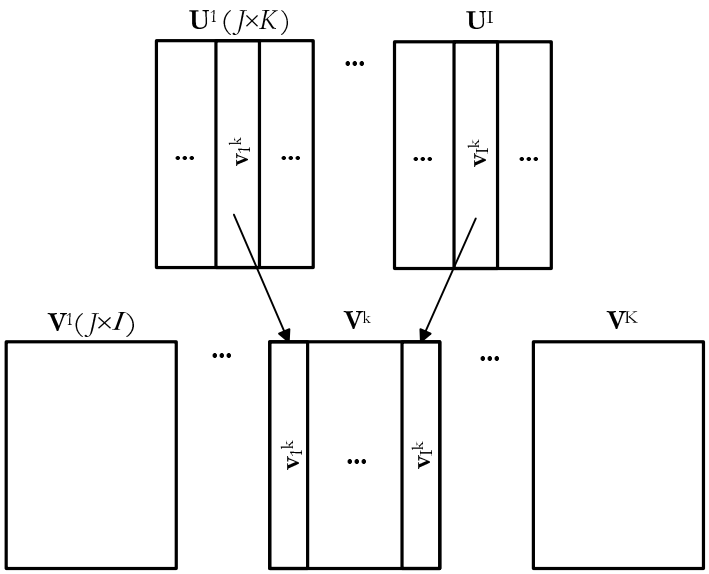


Figure A.6: Illustration of matrices $\mathbf{V}$.

Finally, combining equations A.10 and A.15 it follows that the bins of $\mathbf{S}^k$ can be expressed as

$$\mathbf{b}_f^k = \sum_{i=1}^{I} \sum_{j=1}^{J} \boldsymbol{\phi}_i \, \psi_{j,f}^i \, v_{i,j}^k \,, \tag{A.16}$$

which shows how $\mathbf{S}^k$ can be modeled by a set of temporal basis functions $\boldsymbol{\Phi}$, a set of spectral basis functions $\boldsymbol{\Psi}$ (where $\boldsymbol{\Psi}$ is a three-dimensional matrix of size $(F \times J \times I)$ whose slices are the matrices $\boldsymbol{\Psi}^i$), and a set of coefficients $\mathbf{V}^k$, that is, $\mathbf{S}^k = M_b(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \mathbf{V}^k)$.

Appendix B

PCA and ICA function calls

This appendix gives details of the implementation of the ISA method (with model M_b) introduced in chapter 2. As explained in section 2.3, we use PCA and ICA to learn the sets of basis functions $\mathbf{\Phi}$ and $\mathbf{\Psi}$. In order to learn $\mathbf{\Phi}$, the method uses spectral analysis, namely spectral PCA and spectral ICA of spectrograms (see section 2.4 for the definition of spectral analysis). Afterwards, it learns $\mathbf{\Psi}$ by applying PCA and ICA to the spectral source signals $\mathbf{C}^k$ associated with $\mathbf{\Phi}$.

The implementation of the method with spectral PCA applies MATLAB's built in `princomp` function to $(-\mathbf{X}, \mathbf{X})^T$, where $\mathbf{X}$ is the horizontal concatenation of transposed spectrograms,

$$\mathbf{X} = ((\mathbf{S}^1)^T, (\mathbf{S}^2)^T, \ldots, (\mathbf{S}^K)^T),$$

and $(-\mathbf{X}, \mathbf{X})^T$ is the vertical concatenation of $-\mathbf{X}^T$ and $\mathbf{X}^T$. This produces matrices $\mathbf{\Phi}$ and $\mathbf{C}^T$, where $\mathbf{\Phi}(:, i) = \boldsymbol{\phi}_i$, and $\mathbf{C}$ is a matrix that contains $\mathbf{C}^1, \mathbf{C}^2, \ldots \mathbf{C}^K$ (where $\mathbf{C}^k$ contains vectors $\mathbf{c}_1^k, \ldots, \mathbf{c}_I^k$ as well as the results relative to $-\mathbf{X}$), or more specifically, $\mathbf{C}(i, :)$ is a vector that contains $\mathbf{c}_i^k$ (and $-\mathbf{c}_i^k$) for $k \in 1, ..., K$.

The implementation of the model with spectral ICA applies `fastica` [Hyvärinen et al., 2001] to matrix $(-\mathbf{X}, \mathbf{X})$. The results shown in chapter 2 were obtained with option `g`, which specifies the nonlinearity used in the fixed-point algorithm, set to `tanh`, and option `lastEig`, which specifies the number of eigenvectors used in the computation, set to `50` (this option was set to `30` for the ISAS method in section 3.5.1). This produces matrices $\mathbf{\Phi}$ and $\mathbf{C}$ as defined earlier. Note that here a signal mixture is the horizontal concatenation of one transposed frame from each of the K spectrograms, and there are T signal mixtures. Therefore $I \leqslant T$ in equations 2.8−2.12. (See section 2.3 for the definition of signal mixture and for details on the size of the matrices.)

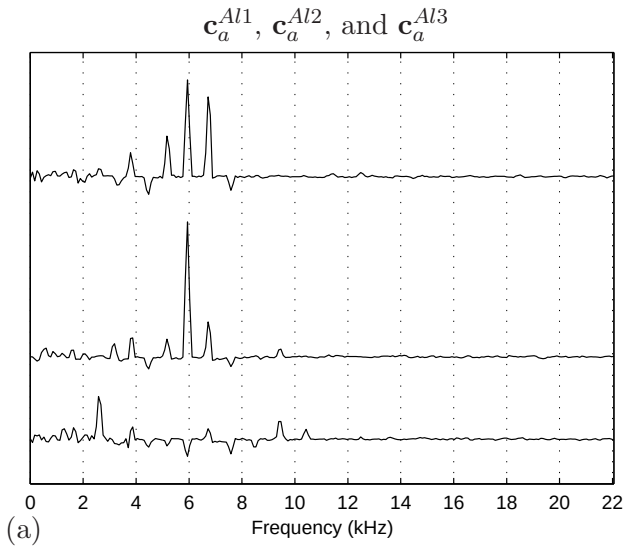
Once it finds the matrices of spectral signals $\mathbf{C}^1, \mathbf{C}^2, \ldots \mathbf{C}^K$, the implementation of the method rearranges them into matrices $\mathbf{D}^i$, where $\mathbf{D}^i = (\mathbf{c}_i^1, ..., \mathbf{c}_i^K)$ (see the top of figure A.5). For each matrix $\mathbf{D}^i$, it applies function `princomp` to $(\mathbf{D}^i)^T$ or `fastica` (with the same options as for spectral ICA) to $\mathbf{D}^i$. These calls produce matrices $\mathbf{\Psi}^i$, and $\mathbf{U}^i$. Finally, it obtains matrices $\mathbf{V}^k$ by rearranging matrices $\mathbf{U}^i$ (see figure A.6). Here a signal mixture is a row of $\mathbf{D}^i$. Since $\mathbf{D}^i$ is a matrix of size $(F \times K)$ there are F signal mixtures. Therefore, $J \leqslant F$ in equations 2.10−2.12.

Appendix C

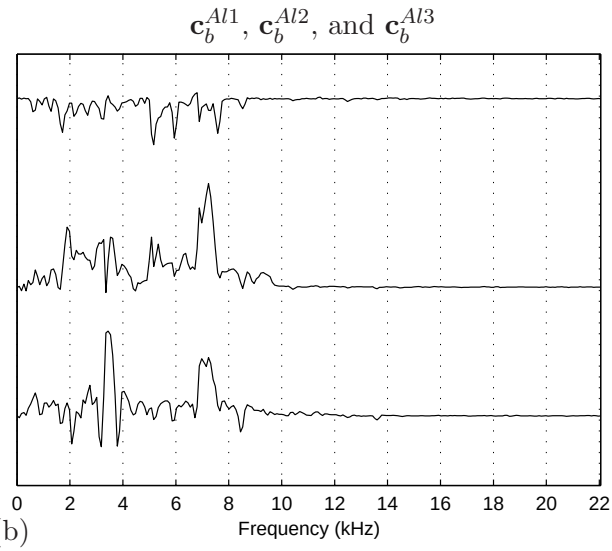
Spectral source signals $\mathbf{C}^k$

This appendix includes some figures of spectral source signals that complement chapter 2.

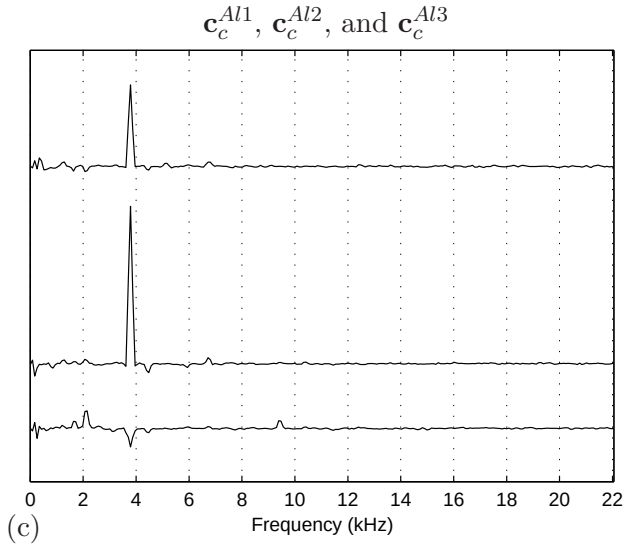
Figure C.1: (In the next page.) Spectral source signals ($\mathbf{C}^k$) obtained by spectral ICA of ten impacts on an aluminum rod. The spectral source signals corresponding to the basis functions shown in figure 2.10a and sounds Al1, Al2 and Al3, are shown from (a) to (f), always from top to bottom. For example, the spectral source signals in (a), $\mathbf{c}_a^{Al1}$, $\mathbf{c}_a^{Al2}$, and $\mathbf{c}_a^{Al1}$, from top to bottom, are associated to basis function $\boldsymbol{\phi}_a$ in figure 2.10a.


$\mathbf{c}_a^{Al1}$, $\mathbf{c}_a^{Al2}$, and $\mathbf{c}_a^{Al3}$
0 2 4 6 8 10 12 14 16 18 20 22
Frequency (kHz)

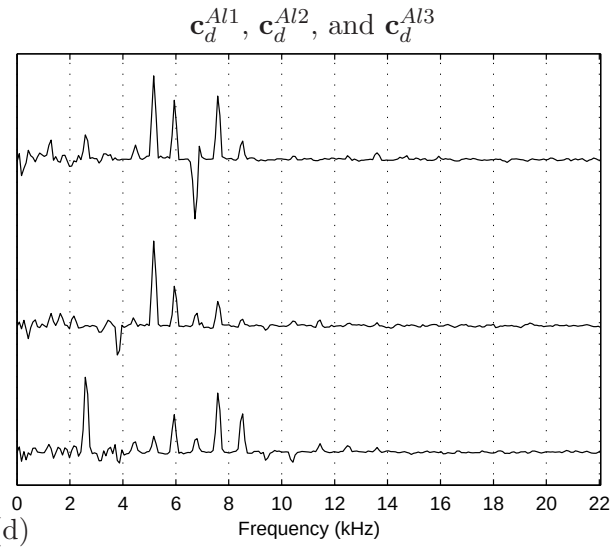
(a)


$\mathbf{c}_b^{Al1}$, $\mathbf{c}_b^{Al2}$, and $\mathbf{c}_b^{Al3}$
0 2 4 6 8 10 12 14 16 18 20 22
Frequency (kHz)

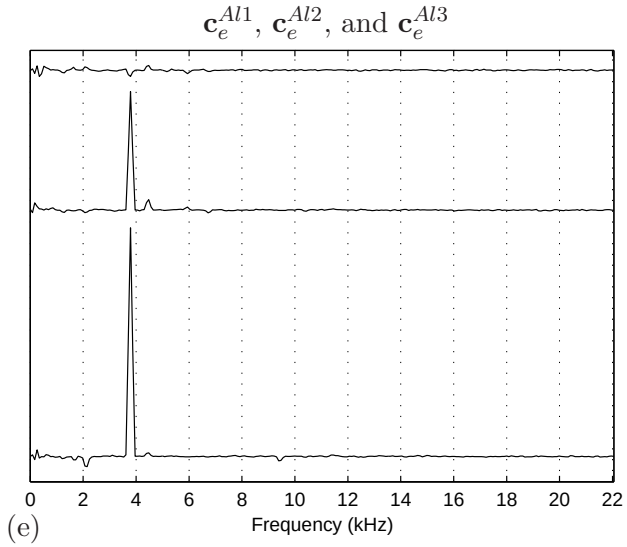
(b)


$\mathbf{c}_c^{Al1}$, $\mathbf{c}_c^{Al2}$, and $\mathbf{c}_c^{Al3}$
0 2 4 6 8 10 12 14 16 18 20 22
Frequency (kHz)

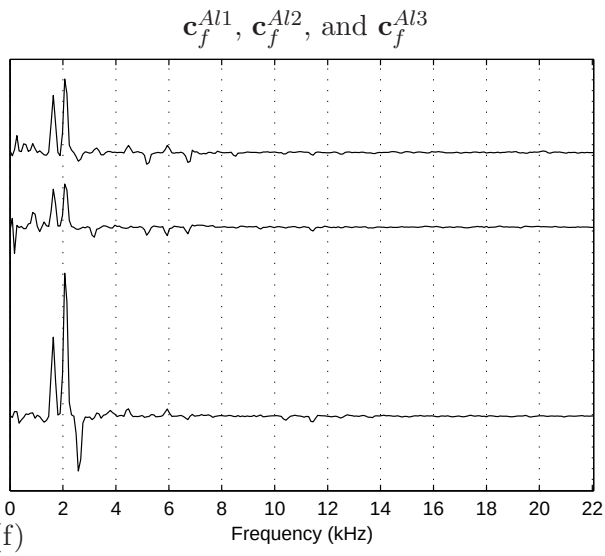
(c)


$\mathbf{c}_d^{Al1}$, $\mathbf{c}_d^{Al2}$, and $\mathbf{c}_d^{Al3}$
0 2 4 6 8 10 12 14 16 18 20 22
Frequency (kHz)

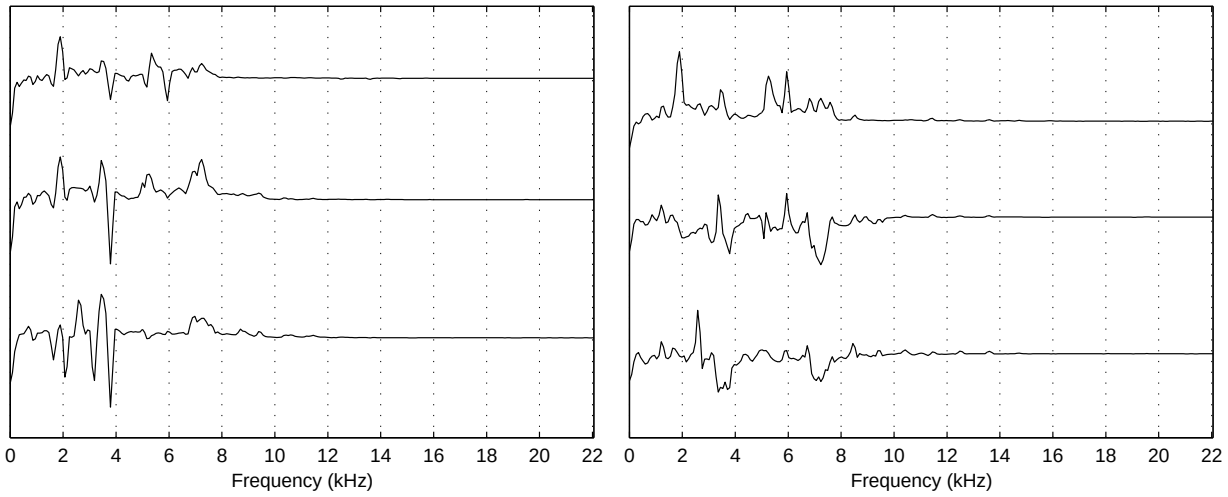
(d)


$\mathbf{c}_e^{Al1}$, $\mathbf{c}_e^{Al2}$, and $\mathbf{c}_e^{Al3}$
0 2 4 6 8 10 12 14 16 18 20 22
Frequency (kHz)

(e)


$\mathbf{c}_f^{Al1}$, $\mathbf{c}_f^{Al2}$, and $\mathbf{c}_f^{Al3}$
0 2 4 6 8 10 12 14 16 18 20 22
Frequency (kHz)

(f)



(a) (b)

Figure C.2: Spectral source signals ($\mathbf{C}^k$) obtained by spectral PCA of ten impacts on the aluminum rod. The spectral source signals associated with basis functions $\boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_3$ from figure 2.11, and sounds Al1, Al2 and Al3 are shown from top to bottom: (a) shows $\mathbf{c}_2^{Al1}$ to $\mathbf{c}_2^{Al3}$, and (b) shows $\mathbf{c}_3^{Al1}$ to $\mathbf{c}_3^{Al3}$.

Appendix D

Discrete cosine transform

The discrete cosine transform (DCT) of an impulse at the beginning (or left side) of the time window is a low frequency sinusoid. This is illustrated in figure D.1. In figure D.1a the signal, x_1 consists of an impulse at the beginning of the window (where time runs from left to right). Figure D.1b shows the DCT of that signal. Here, the impulse is represented by a low frequency sinusoid. As the impulse appears later in time (that is, towards the right side of the window), the frequency of the sinusoid increases, that is, low frequency sinusoids represent signals towards the left side of the window, while high frequency sinusoids represent signals towards the right side of the window. Figure D.2a shows an impulse, x_2, in the time domain. Since the impulse is located towards the right side of the window, its DCT is a high frequency sinusoid, which is plotted in figure D.2b.

The DCT represents the transients as sinusoids in the frequency domain. This transform maps the signal into a frequency by amplitude space and retains phase information. There are other transforms with similar properties, such as the discrete sine transform. On the other hand, the discrete Fourier transform (DFT) maps the signal into a frequency by amplitude and phase space. Using the real or imaginary DFT implies loss of information, and the magnitude DFT does not have the same properties, nor the correspondence between the frequency of the basis function and the time at which the transient occurs. For example, the magnitude DFT gives the same representation for x_1 and x_2; both signals are represented as constant energy spread over the whole spectrum (figures D.1c and D.2c).

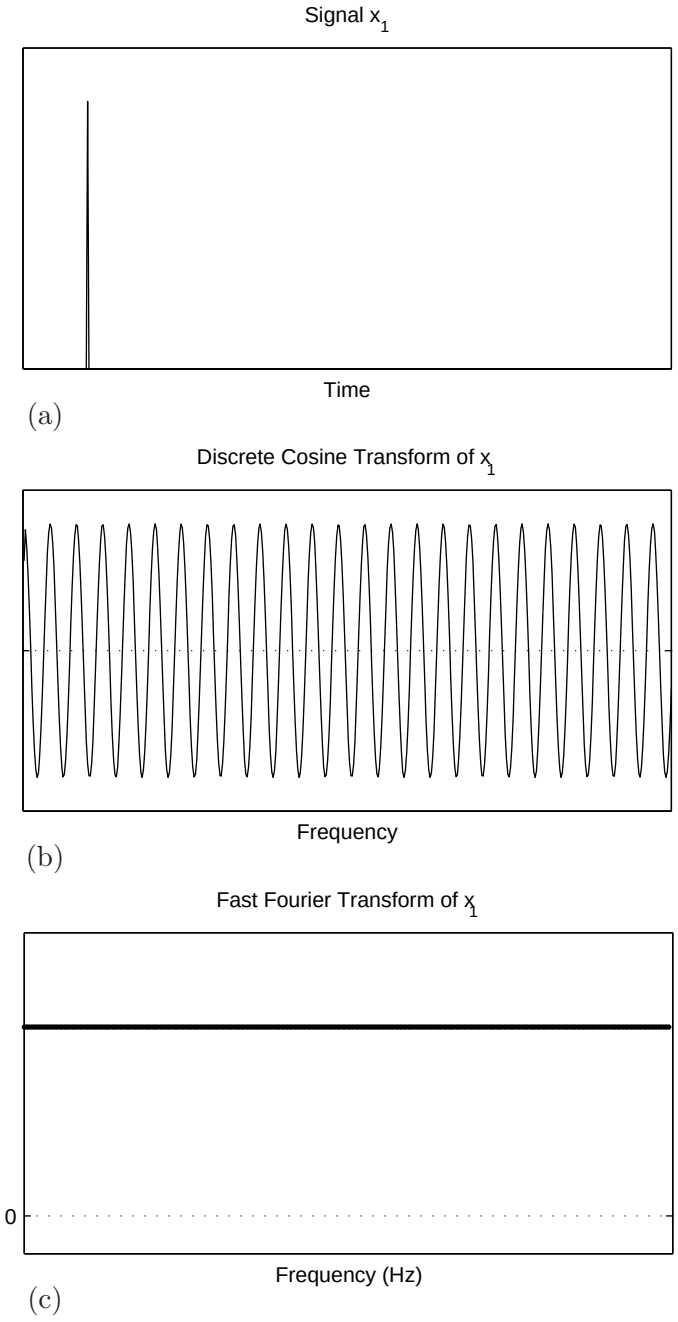


Figure D.1: DCT of an impulse x_1 located towards the beginning of the window. (a) The waveform x_1. (b) The DCT II of x_1 is a low frequency sinusoid. (c) The magnitude DFT of x_1 is constant.

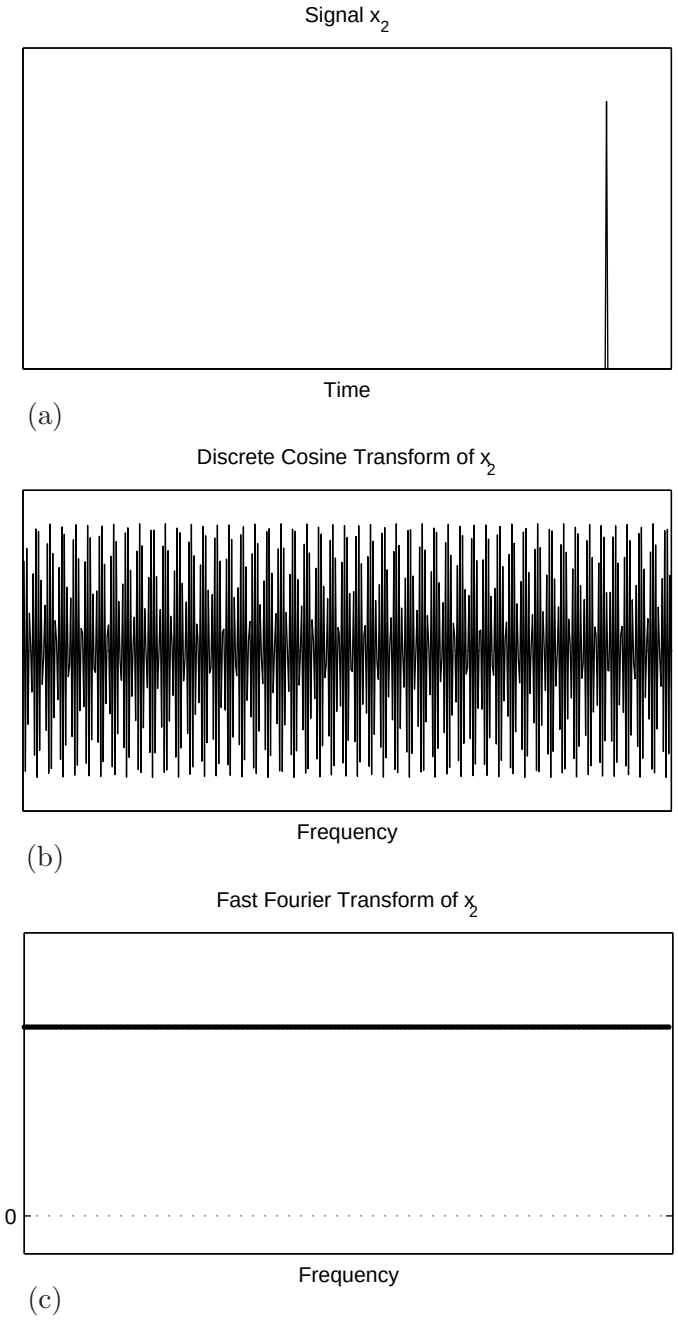


Figure D.2: DCT of an impulse x_2 located towards the end of the window. (a) The waveform x_2. (b) The DCT II of x_2 is a high frequency sinusoid. (c) The magnitude DFT of x_2 is constant.

Bibliography

M. Ali. *Adaptive Signal Representation with Applications in Audio Coding*. PhD thesis, University of Minnesota, 1996.

L.B. Almeida and F.M. Silva. Variable-frequency synthesis: an improved harmonic coding scheme. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 27.5.1–27.5.4, San Diego, CA, 1984.

F. Avanzini and D. Rocchesso. Modeling collision sounds: Non-linear contact force. In *Proceedings of the COST G-6 Conference on Digital Audio Effects, 2001*, December 2001a.

F. Avanzini and D. Rocchesso. Controlling material properties in physical models of sounding objects. In *Proceedings of the International Computer Music Conference 2001*, pages 91–94, September 2001b.

A.K. Barros, T. Rutkowski, F. Itakura, and N. Ohnishi. Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavelets. *IEEE transactions on neural networks*, 13(4):888–893, 2002.

S. Bech and N. Zacharov. *Perceptual Audio Evaluation-Theory, Method and Application*. John Wiley & sons, Ltd, 2006.

G. Borin, G. De Poli, and A. Sarti. Musical signal synthesis. In C. Roads, S.T. Pope, A. Picialli, and G. De Poli, editors, *Musical Signal Processing*. Swets & Zeitlinger Publishers, 1997.

J. Brown and P. Smaragdis. Independent component analysis for automatic note extraction from musical trills. *Journal of Acoustical Society of America*, 115(5):2295–2306, May 2004.

M. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. In *Proceedings of the International Computer Music Conference*, Berlin, Germany, 2000.

S. Cavaco and M.S. Lewicki. Statistical modeling of intrinsic structures in impact sounds. *Journal of the Acoustical Society of America*, 121(6):3558–3568, June 2007.

M. F. Christensen and S. van de Par. Efficient parametric coding of transients. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1340–1351, 2006.

G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Journal of Constructive Approximation*, 13:57–98, 1997.

Ph. 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.

Ph. Depalle, G. Garcia, and X. Rodet. Analysis of sound for additive synthesis: Tracking of partials using hidden markov model. In *Proceedings of International Computer Music Conference (ICMC'93)*, Tokyo, Japan, September 1993.

M. Desainte-Catherine and P. Hanna. Statistical approach for sounds modeling. In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx 00)*, Verona, Italy, December 2000.

M.H. DeGroot and M.J. Schervish. *Probability and Statistics*. Addison Wesley, third edition, 2002.

Y. Ding and X. Qian. Sinusoidal and residual decomposition and residual modeling of musical tones using the QUASAR signal model. In *Proceedings of the International Computer Music Conference*, pages 35–42, September 1997.

D. Ellis. Sinewave and sinusoid+noise analysis/synthesis. LabRosa home page, 25th of March 2003. http://labrosa.ee.columbia.edu/matlab/sinemodel/.

D. Ellis and B. Vercoe. A perceptual representation of sound for auditory signal separation. presented at the 123rd meeting of the Acoustical Society of America, Salt Lake City, May 1992.

K. Fitz, L. Haken, and P. Christensen. Transient preservation under transformation in an additive sound model. In *Proceedings of the International Computer Music Conference*, 2000.

J.L. Flanagan and R.M. Golden. Phase vocoder. *Bell System Technical Journal*, 45:1493–1509, November 1966.

D.J Freed. Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events. *Journal of Acoustical Society of America*, 87(1):311–322, January 1990.

W. Gaver. *Auditory display: sonification, audification and auditory interfaces*, chapter Using and Creating Auditory Icons, pages 417–446. Addison-Wesley, 1994.

W. Gaver. *Everyday Listening and Auditory Icons*. PhD thesis, University of California at San Diego, San Diego, CA, 1988.

E.B. George and M.J.T. Smith. Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones. *Journal of the Audio Engineering Society*, 40(6): 497–515, June 1992.

M. Goodwin. Matching pursuit with damped sinusoids. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 2037–2040, April 1997.

M. Goodwin. Residual modeling in music analysis/synthesis. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1005–1008, May 1996.

M. Goodwin and M. Vetterli. Matching pursuit and atomic signal models basd on recursive filter banks. *IEEE Transactions on Signal Processing*, 47(7):1890–1902, July 1999.

R. Gribonval and E. Bacry. Harmonic decomposition of audio signals with matching pursuit. *IEEE Transactions on Signal Processing*, 51(1):101–111, January 2003.

R. Gribonval, Ph. Depalle, X. Rodet, E. Bacry, and S. Mallat. Sound signals decomposition using a high resolution matching pursuit. In *Proceedings of International Computer Music Conference (ICMC'96)*, pages 293–296, 1996.

D. Griffin and J.S. Lim. Multiband excitation vocoder. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 36, issue 8, pages 1223–1235, 1988.

S. Hainsworth and M. Macleod. On sinusoidal parameter estimation. In *Proceedings of the International Conference on Digital Audio Effects (DAFx 03)*, London, UK, September 2003.

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

D.L. James, J. Barbič, and D.K. Pai. Precomputed acoustic transfer: Output-sensitive, accurate sound generation for geometrically complex vibration sources. *ACM Transactions on Graphics (SIGGRAPH)*, 25(3):987–995, 2006.

S.M. Kay and S.L. Marple Jr. Spectrum analysis - a modern perspective. *Proceedings of the IEEE*, 69(11):1380–1419, 1981.

H.-G. Kim, E. Berdahl, and T. Sikora. Study of MPEG-7 sound classification and retrieval. In *5th International ITG Conference on Source and Channel Coding*, Erlangen, Germany, 2004.

R.L. Klatzky, D.K. Pai, and E.P. Krotkov. Perception of material from contact sounds. *Presence: Teleoperators and Virtual Environments*, 9(4):399–410, August 2000.

P. Ladefoged. *Elements of acousitc phonetics*. The University of Chicago Press, second edition edition, 1996.

M. Lagrange, S. Marchand, M. Raspaud, and J.-B. Rault. Enhanced partial tracking using linear prediction. In *Proceedings of the International Conference on Digital Audio Effects (DAFx 03)*, London, UK, September 2003.

S. Lakatos, S. McAdams, and René Caussé. The representation of auditory source characteristics: simple geometric form. *Perception and Psychophysics*, 59(8):1180–1190, 1997.

C. Lambourg, A. Chaigne, and D. Matignon. Time-domain simulation of damped impacted plates. ii. numerical model and results. *Journal of Acoustical Society of America*, 109(4):1433–1447, April 2001.

J. Laroche and M. Dolson. Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing*, 7(3):323–332, May 1999a.

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*, pages 91–94, New Paltz, NY, October 1999b.

H. Levitt. Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, 49:467–477, 1971.

M. S. Lewicki and T. J. Sejnowski. Coding time-varying signals using sparse, shift-invariant representations. In *Advances in Neural Information Processing Systems*, volume 11, pages 730–736. MIT Press, 1999.

M.S. Lewicki. Efficient coding of time-varying signals using a spiking population code. In R.P.N. Rao, B.A. Olshausen, and M.S. Lewicki, editors, *Probabilistic Models of the Brain: Perception and Neural Function*, chapter 12, pages 223–234. MIT Press, Cambridge, MA, 2002a.

M.S. Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356–363, 2002b.

R.A. Lutfi. Auditory detection of hollowness. *Journal of Acoustical Society of America*, 110(2): 1010–1019, August 2001.

R.A. Lutfi and E.L. Oh. Auditory discrimination of material changes in struck-clamped bar. *Journal of Acoustical Society of America*, 102(4):3647–3656, December 1997.

S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, December 1993.

S. Marchand. Improving spectral analysis precision with an enhanced phase vocoder using signal derivatives. In *Proceedings of the Digital Audio Effects Workshop (DAFx 98)*, pages 114–118, Barcelona, Spain, November 1998.

J.S. Marques and L.B. Almeida. New basis functions for sinusoidal decomposition. In *Proceedings of EUROCON*, Stockholm, Sweden, 1988.

P. Masri. *Computer Modeling of Sound for Transformation and Synthesis of Musical Signal*. PhD thesis, University of Bristol, 1996.

P. Masri and A. Bateman. Improved modeling of attack transients in music analysis-resynthesis. In *Proceedings of the International Computer Music Conference*, pages 100–103, Hong Kong, August 1996.

S. McAdams and E. Bigand, editors. *Thinking in Sound, the cognitive psychology of human audition*, chapter Glossary. Oxford University Press, 1993.

R.J McAulay and T.F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34(4):744–754, August 1986.

R.J McAulay and T.F. Quatieri. Sinusoidal coding. In W.B. Kleijn and K.K. Paliwal, editors, *Speech Coding and Synthesis*, chapter 4, pages 121–173. Elsevier Science B.V., 1995.

J.A. Moorer. The use of the phase vocoder in computer music application. *Journal of the Audio Engineering Society*, 26(1/2):42–45, January/February 1978.

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

J.F. O'Brien, P.R. Cook, and G. Ess. Synthesizing sounds from physically based motion. In *SIGGRAPH*, 2001.

J.F. O'Brien, S. Shen, and C.M. Gatchalian. Synthesizing sounds from rigid-body simulations. In *Proceedings of ACM SIGGRAPH Symposium on Computer Animation*, pages 175–181, San Antonio, Texas, 2002.

D.K. Pai, K. van den Doel, D.L. James, J. Lang, J.E. Lloyd, J.L. Richmond, and S.H Yau. Scanning physical interaction behavior of 3d objects. In *SIGGRAPH*, 2001.

J.M. Picket. *The acoustics of speech communication, fundamentals, speech percption theory, and technology*. Allyn and Bacon, 1999.

M. Portnoff. Implementation of the digital phase vocoder using the fast Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(3):243–248, June 1976.

M. Portnoff. Short-time Fourier analysis of sample speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):374–390, 1981.

G.R.B. de Prony. Essai expérimentale et analytique, etc. *Paris Journal de l'Ecole Polytechnique*, 1(2):24–76, 1795.

M.S. Puckette. Phase-locked vocoder. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 1995.

C. Roads. *The Computer Music Tutorial*. The MIT Press, 1996.

A. Röbel. Transient detection and preservation in the phase vocoder. In *Proceedings of the International Computer Music Conference (ICMC'03)*, pages 247–250, Singapore, 2003.

A. Röbel. Onset detection in polyphonic signals by means of transient peak classification. MIREX Online Proceedings (ISMIR 2005), September 2005.

X. Rodet and F. Jaillet. Detection and modeling of fast attack transients. In *Proceedings of the International Computer Music Conference (ICMC'01)*, pages 30–33, 2001.

X. Serra. Musical sound modeling with sinusoids plus noise. In C. Roads, S.T. Pope, A. Picialli, and G. De Poli, editors, *Musical Signal Processing*. Swets & Zeitlinger Publishers, 1997.

X. Serra. *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition*. PhD thesis, Stanford University, 1989.

X. Serra and J. Smith. Spectral modeling synthesis: a sound analysis/synthesis based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24, 1990.

P. Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In C.G. Puntonet and A. Prieto, editors, *Independent Component Analysis and Blind Signal Separation, 5th International Conference, ICA*, volume 3195 of *Lecture Notes in Computer Science*, pages 494–499, Granada, Spain, 2004.

P. Smaragdis and J. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, October 2003.

E. Smith. *Efficient Auditory Coding*. PhD thesis, Carnegie Mellon University, 2006.

E. Smith and M.S. Lewicki. Efficient coding of time-relative structure using spikes. *Neural Computation*, 17(1):19–45, 2005.

E.C. Smith and M.S. Lewicki. Efficient auditory coding. *Nature*, 439(7079):800–805, 2006.

J.O. 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*, pages 290–297, 1987.

J.V. Stone. *Independent Component Analysis, A Tutorial Introduction*. MIT Press, 2004.

H. Thornburg and F. Gouyon. A flexible analysis-synthesis method for transients. In *Proceedings of International Computer Music Conference (ICMC'2000)*, pages 400–403, 2000.

K. van den Doel, P.G. Kry, and D.K. Pai. Foleyautomatic: physically-based sound effects for interactive simulation and animation. In *SIGGRAPH*, August 2001.

K. van den Doel, D.K. Pai, T. Adam, L. Kortchmar, and K. Pichora-Fuller. Measurements of perceptual quality of contact sound models. In *Proceedings of the International Conference on Auditory Display*, pages 345–349, Kyoto, Japan, 2002.

T.S. Verma. *A Perceptually Based Audio Signal Model with Application to Scalable Audio*. PhD thesis, Standford University, 1999.

T.S. Verma and T.H.Y. Meng. Extending spectral modeling synthesis with transient modeling synthesis. *Computer Music Journal*, 24(2):47–59, 2000.

T.S. Verma, S.N. Levine, and T.H.Y. Meng. Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals. In *Proceedings of the International Computer Music Conference*, pages 164–167, September 1997.

T. Virtanen. Separation of sound sources by convolutive sparse coding. In *Proceedings of the ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, Jeju, Korea, 2004.

Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. Huang. Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

W.A. Yost. *Fundamentals of hearing, an introduction*. Academic Press, fourth edition, 2000.