Audio Source Separation via the Grouping of Partials in the Sum-of-Sinusoids Model

Nicholas Esterer

Department of Music Research
Schulich School of Music
McGill University
Montreal, Canada

November 2016

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Master of Arts.

Abstract

This thesis explores a strategy of audio source separation that relies on a classification of partials via their measured modulations. Perceptual studies have shown that signals with common amplitude- and frequency-modulation are heard as coming from the same source [35], [34]. To include these modulations in a sum-of-sinusoids model a nonlinear polynomial phase function is adopted whose parameters are estimated using the Distribution Derivative Method (DDM) [2]. For better estimation accuracy, a window is designed that has lower side-lobes than the canonical Hann window but that is also once-differentiable — a requirement of the DDM. These estimated parameters are used in a new partial tracking algorithm based on linear programming. The resulting partials are classified using the clustering technique of Gaussian mixture models [14] on frequency- and amplitude-modulation data. Principal components analysis is used to emphasize the parameter on which it would be best to perform classification. Once the partials have been classified into sources, the sources are synthesized from the measured sinusoidal parameters.

The additional information provided by the DDM (namely the frequency and log-amplitude slope) is incorporated into interpolating polynomials for the phase and amplitude of sinusoids. The quality of different model-orders for these polynomials is assessed on synthetic signals. The source separation system is evaluated on both simulated data and on a mixture of real recordings of percussive and plucked string instruments. In this latter case, it is shown that using amplitude-modulation is a good criterion for separation when there is little frequency-modulation.

Résumé

Dans ce mémoire nous explorons une stratégie de séparation de sources sonores s'appuyant sur une classification de partiels selon leurs modulations observées. Des études perceptives ont montré que des signaux dont les modulations d'amplitude et de fréquence sont communes sont perçus comme provenant d'une même source [35], [34]. Afin d'inclure ces modulations dans le modèle de synthèse additif, la phase est représentée par une fonction polynomiale non-linéaire dont les paramètres sont estimés par la méthode de distribution dérivée (Distribution Derivative Method - DDM) [2]. Afin d'améliorer la qualité d'estimation, nous avons conçu une fenêtre dont la résolution dynamique est meilleure que celle de la fenêtre canonique de Hann, tout en étant dérivable sur tout son domaine, propriété requise par la technique DDM. Les paramètres ainsi estimés sont utilisés par un nouvel algorithme de suivi de partiels fondé sur le principe d'optimisation par programmation linéaire. Les partiels trouvés sont alors classifiés en différentes sources par une technique de mélange gaussiens appliquées aux paramètres de modulation de fréquence et d'amplitude. Au préalable une analyse en composantes principales (PCA) est utilisée afin de faire ressortir les paramètres les mieux appropriés pour la classification. Une fois les partiels regroupés et classifiés en sources, celles-ci sont synthétisées en fonctions des paramètres associés aux trajets de partiels.

Plus précisément, les informations additionnelles fournies par la DDM (dérivée de la fréquence et de la log-amplitude) sont prises en compte selon plusieurs stratégies impliquant des polynômes de reconstruction de phase et d'amplitude d'ordres différents. La qualité des signaux re-synthétisés est alors évaluée pour chacune de ces stratégies.

Enfin, ce système de séparation de sources est testé sur un mélange de signaux synthétiques puis sur un mélange de signaux instrumentaux réels de percussion et de corde pincée. Dans ce dernier cas, nous montrons que la prise en compte de la modulation d'amplitude aide à la classification en l'absence de modulation en fréquence.

Acknowledgments

My time at McGill University has been a long one, so I have many people to thank.

First and foremost, I would like to thank Professor Philippe Depalle for his mentorship, support, and patience during the writing of this thesis, especially during the final months. He has a tremendous talent for understanding and organizing thought. I would not have been able to produce a document of this quality without his help.

I am grateful to Professor Gary Scavone both for the teaching assistance opportunities he granted me and for inspiring me in those first courses I took with him to further pursue Music Technology. I thank Professor Ichiro Fujinaga and Professor Stephen McAdams for the opportunities they provided to work on Music Technology projects, furthering my training in information technology and signal processing.

Much of my interest in music technology and signal processing was awoken by my colleague and friend, Scott Monk, to whom I express much gratitude.

Finally I would like to thank my parents, Denise and Rob Esterer, for their encouragement in all my endeavours.

Thank you all.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

This thesis was written in the information age where digital signals can be produced easily and are produced in great volumes. As stated in the introduction of [21], signals are the means by which information is transmitted. In the past, producing a digital signal was costly and required specialized equipment, motivating the user of the equipment to carefully plan the process that was to be documented by encoding the measurements into the signal. Now these tools are widely available and accessible to everyone, increasing both the variance in quality, but also the potency of information. For this reason, new signal processing techniques require ways of removing extraneous information, that is, data that have meaning and structure, but that are not pertinent to the information of interest. In this thesis, a producer of information is a *source*, and we have many sources transmitting information in a single signal.

Source separation is a difficult problem because it involves simultaneously estimating characteristics of the sources while separating them: for improved estimation, interference from other sources should be minimized; in order to remove interfering sources, their characteristics must be known. For the estimation problem, we must resort to using prior information: we assume we know the structure of the sources and can quantify in some way their characteristics, penalizing characterizations estimated by our system that do not match presumptions. For the separation problem, we often must resort to suboptimal solutions. These solutions may be adequate to aid a human in manual refinement of the sources, or serve as input to another technique.

1.2 Applications

Applications of source separation exist in a variety of disciplines and entire conferences are dedicated to the subject (see, for example, [64]). Here we will only consider applications pertinent to audio, acoustics and music. One of the most popular applications of audio source separation is for automated music transcription (see e.g., [1]). Having access to both a representation of the musical score and its constituent sounds would be very convenient for composers and sound engineers. Those interested in isolating individual voices in a recording of multiple speakers (for honourable or dishonourable purposes) would benefit from audio source separation — [42] discusses a strategy using characterizations of language to aid in the separation. There are no doubt many more applications of source separation in the field of audio signal processing.

1.3 Organization

The general strategy explored in this thesis is an iterative process with four distinct steps. (1) a model is chosen of the signals of interest. (2) realisations of this model are identified in the measurement signal. (3) once these have been identified, the parameters of the realisations are estimated. (4) the estimations are used to classify these realisations as one of a smaller set of higher-level objects. The structure of these objects is used to inform the selection of the new model, whose parameters are then estimated, etc.; the process can repeated to build up successively higher-level models.

can be

This thesis is structured according to these steps and is as follows. Chapter 2 discusses previous approaches to audio source separation and introduces the method adopted here. The following, Chapter 3, describes the signal model chosen to describe musical signals — the additive sinusoidal model with polynomial log-amplitude and phase — and describes a technique for estimating the parameters of models of arbitrary order. We introduce a new analysis window to have more control over the estimation process. Techniques to identify these models in signals are discussed in Chapter 4, where a classical technique of partial tracking is compared to a new linear programming formulation. Chapter 5 shows how the separated sources can be synthesized from the estimated model parameters using the additional information provided by the higher-order polynomial model. Finally, two experiments are carried out that demonstrate the use of amplitude- and frequency-

modulation to classify sources. The first, in Chapter 6 is on synthetic data and the second in Chapter 7 on a recording of percussion and plucked string instruments. It is in these latter chapters where classification is performed and its adaptation to the particular audio source separation problem is discussed. The Appendices A through C explain elements of these classification techniques.

1.4 Notation

1.4.1 Vectors and matrices

While scalars are typeset normally — x is an example of a scalar — vectors and matrices are typeset in a boldface font, with matrices written with a capital letter, e.g., $\boldsymbol{x}$ is a vector and $\boldsymbol{X}$ a matrix. If a number is written instead of a symbol, we mean a vector whose entries are all that number, e.g., $\mathbf{1}$ is the vector of all 1s, $\mathbf{0}$ the vector of all 0s. The ith entry of a vector $\boldsymbol{x}$ is written x_i and the entry in the ith row and jth column of a matrix $\boldsymbol{X}$ is written $X_{i,j}$. Both are scalars and therefore typeset normally. Sometimes we might find it convenient to extract a column vector or row vector from the matrix $\boldsymbol{X}$. We write $\boldsymbol{x}_{i,:}$ to extract all columns from the ith row and $\boldsymbol{x}_{:,j}$ to extract all rows from the jth column. These are the ith row vector and jth column vector respectively. The orientation of a vector will be clear from the context, but in general $\boldsymbol{x}$ is a column vector while $\boldsymbol{y}_{i,:}$ and $\boldsymbol{x}^T$ are row vectors.

1.4.2 Operators

Inner product

We will be dealing with objects in vector spaces. The operator $\langle x, y \rangle$ takes two objects in a vector space V, $x, y \in V$ and maps them to an element $k \in \mathbb{K}$ of a field $\mathbb{K}$. For this thesis, the field will always be the field of real numbers $\mathbb{R}$ or complex numbers $\mathbb{C}$. The vector space can be the set of vectors of N elements in $\mathbb{R}^N$ or $\mathbb{C}^N$, in which case the inner product is defined, for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{K}^N$, $k \in \mathbb{K}$

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y} = k$$

The inner product is also defined on the vector space of functions Φ mapping from a set S to a field $\mathbb{K}$, $\Phi : \forall f$ s.t. $f(s) = k, s \in S, k \in \mathbb{K}$ in which case the inner product on $g, f \in \Phi$ is defined as

$$\langle g, f \rangle = \int_{-\infty}^{\infty} g(x)\overline{f(x)}dx$$

and $\overline{a}$ gives the complex conjugate of a.

General outer operators

The outer operator $\cdot \otimes_{\mathcal{O}} \cdot$ will only be defined for vectors in this thesis. It operates on the two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{K}^N$ and is defined as

$$\boldsymbol{x} \otimes_{\mathcal{O}} \boldsymbol{y} = \boldsymbol{W}$$

where the ith row and jth column of $\boldsymbol{W}$ are

$$w_{i,j} = \mathcal{O}(x_i, y_j)$$

Canonically, the operator $\mathcal{O}$ is multiplication in which case

$$\boldsymbol{x} \otimes_{\times} \boldsymbol{y} = \boldsymbol{W}$$

where the ith row and jth column of W are

$$W_{i,j} = x_i y_j$$

This outer product is also known as the *Kronecker product*, and we will omit the operator subscript when that is the case, i.e., we will simply write $\otimes$. As stated above, however, $\mathcal{O}$ can be defined arbitrarily as any function taking two inputs a returning a single output.

Point-wise operators

If an operator on matrices $\circ$ is written with a period preceding it, i.e., $.\circ$ it means perform that operation on each element individually. Some examples follow.

For matrix $\boldsymbol{X} \in \mathbb{K}^{M,N}$ and $p \in \mathbb{K}$

$$\boldsymbol{X}^{\cdot p} = \boldsymbol{W}$$

where

$$W_{i,j} = X_{i,j}^{p}$$

For matrices $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{K}^{M,N}$

$$\boldsymbol{X}.\boldsymbol{Y} = \boldsymbol{W}$$

where

$$W_{i,j} = X_{i,j} Y_{i,j}$$

(contrast these with canonical matrix multiplication).

1.4.3 Random variables

Many authors denote random variables with a normally typeset uppercase letter. We will use this convention only when convenient, but will always state explicitly that a certain variable is random. We distinguish between discrete and continuous random variables in our notation.

Discrete random variables

If a random variable X can only take on values in a discrete set, we say that this random variable is discrete-valued. Formally a discrete set Γ of size N is one for which there exists an isomorphism $\mathscr{I}$ that maps Γ on to the subset of the integers $[1, \ldots, N]$. The probability that X takes on the value x is written $\mathrm{p}(X = x)$ for discrete random variables.

Continuous random variables

If a random variable X can take on values in a set Γ isomorphic to $\mathbb{R}$ we say this random variable is continuous-valued. The probability that X takes on the value x is written $\mathrm{p}(x)$ for continuous random variables.

1.4.4 Complex numbers

A complex number $z \in \mathbb{C}$ can be described in Cartesian notation as

$$z = a + jb, a, b \in \mathbb{R}$$

or in polar notation as

$$z = \alpha \exp(j\omega), \alpha, \omega \in \mathbb{R}$$

where $j = \sqrt{-1}$. j is also often used to denote an index variable. It will be clear from the context when the imaginary number is meant and when the index is meant.

1.4.5 Logarithms

The logarithm base-e[1] of x is written $\log(x)$. The logarithm of any other base b will be denoted as such: $\log_b(x)$.

1.4.6 Ultimate values

This thesis presents a number of iterative algorithms whose steps will be counted using an index l and whose solutions will take on the values x_l. We use the $*$ notation to refer to these values at the final iterates: the index of the final iteration is written[2] l^* and its value $x_{l^*} = x^*$.

[1] e is Euler's constant.

[2] Arbitrary letters can be used.

Chapter 2

Methodology

Most audio source separation strategies use some combination of two general methodologies: at one end of this continuum are those that use physical or structural models of the sources and at the other, those that use purely statistical models. Here we give a brief overview of some previously proposed techniques.

2.1 Additive sinusoidal model

The additive sinusoidal model [52], [36], is a convenient model with wide-spread use in the computer music community. Various authors have applied this model to the source separation problem. For example, in [61] a prior estimation of the fundamental frequencies is used in tandem with temporal and spectral smoothness constraints to separate sources estimated via an additive model. Similar to this thesis, [6] uses an additive sinusoidal model to provide frequency-modulation cues to a latent component technique, and in [31], common amplitude-modulation and fundamental frequency estimation are use to separate sources, the additive sinusoidal model being used to reconstruct the phases of overlapping harmonics using the fundamental frequencies. The additive sinusoidal model is also the model adopted in this thesis and its use is justified in Chapter 3.

2.2 Multiple fundamental frequency estimation

This technique assumes the signal considered can be described in a format akin to the musical score — a collection of notes each with times indexing their beginning and end

and a frequency, the fundamental, describing the perceived pitch of the note. Multiple fundamental frequency estimation for the purposes of music transcription dates back to the 1970s [20, ch. 20] [41]. This is related to audio source separation because the resulting high-level representation — the estimated score — can be used to synthesize signals corresponding to subsets of notes in the score, e.g., if a particular instrument is desired, its notes are extracted and then a signal is synthesized using stored recordings of the instrument or instrument-modeling synthesis techniques. The technique has become quite developed, see [20, ch. 20] for a review of modern techniques.

There are some drawbacks to the technique. Many musical signals of interest such as unpitched percussion, do not always have a perceivable fundamental frequency. A musical score describes notes as having distinct boundaries in time and frequency, which is not always true when one considers musical gestures such as portamento or *dal niente*[1]. Nevertheless, the production of even a crude score from a musical signals is useful in applications such as automated music transcription (e.g., [51]) or music catalogue query [37].

2.3 Principal latent component analysis (PLCA) and non-negative matrix factorizations (NMF)

2.3.1 Motivation

The power spectrum of a signal and consequently its spectrogram are always non-negative valued. If a signal is considered as consisting of a sum of original source signals, these source signals will have non-negative spectrograms as well. The following two techniques attempt to determine these spectrograms solely from a spectrogram of their mixture using techniques for determining latent variables. We will refer to techniques of this sort as *latent variable models*. The technique is discussed in a bit more detail in the following to demonstrate how it differs from the additive technique explored in this thesis.

2.3.2 Approaches

The PLCA and NMF approaches to the audio source separation problem are very popular. An early and highly cited work that applies NMF to polyphonic music transcription is

[1] "Out of nothing": usually accompanying a crescendo and indicating that the player start from silence and gradually increase their playing dynamic.

[54]. Since then many variations on this approach have been proposed. A technique using smoothness based on spectral difference and sparseness as regularization terms is presented in [62]. In [60] vectors in the resulting matrices are forced to be a linear combination of predefined "basis spectra", chosen for their harmonic and perceptual properties. [1] explores the uses of different divergences for the purposes of up-mixing and noise removal.

2.3.3 PLCA

In this formulation, the spectrogram (defined in Section 3.2), being non-negative like a probability distribution, is considered as such

$$c|X(t,f)| = \mathrm{p}(t,f)$$

where c is a constant so that the distribution integrate to 1. Explicitly, we consider the probability that energy in the spectrogram lie in the vicinity of time t and frequency f. With the hope of retrieving the spectrograms of the P underlying sources, it is proposed that the distribution is actually the distribution of K random variables each being chosen with probability $\mathrm{p}(Z=k)$. The pair of random variables from component k, T_k and F_k are assumed independent, i.e., $\mathrm{p}(t,f|Z=k) = \mathrm{p}(t|Z=k)\mathrm{p}(f|Z=k)$. Each random variable, it is hoped, describes a source ($K=P$) or a part of a source ($K>P$). In addition, each of these underlying distributions has marginal distributions $\mathrm{p}(t|Z=k)$ and $\mathrm{p}(f|Z=k)$. The marginal distributions and $\mathrm{p}(Z=k)$ can be estimated using the expectation maximization algorithm with the following update rules for the lth iteration of the algorithm [55]

$$\mathrm{p}_{l+1}(Z=k|t,f) = \frac{\mathrm{p}_l(Z=k)\mathrm{p}_l(t|Z=k)\mathrm{p}_l(f|Z=k)}{\sum_{j=0}^{K-1}\mathrm{p}_l(Z=j)\mathrm{p}_l(t|Z=j)\mathrm{p}_l(f|Z=j)}$$

$$\mathrm{p}_{l+1}(t|Z=k) = \frac{\sum_f \mathrm{p}(t,f)\mathrm{p}_{l+1}(Z=k|t,f)}{\sum_s\sum_f \mathrm{p}(s,f)\mathrm{p}_{l+1}(Z=k|s,f)}$$

$$\mathrm{p}_{l+1}(f|Z=k) = \frac{\sum_t \mathrm{p}(t,f)\mathrm{p}_{l+1}(Z=k|t,f)}{\sum_t\sum_g \mathrm{p}(t,g)\mathrm{p}_{l+1}(Z=k|t,g)}$$

$$\mathrm{p}_{l+1}(Z=k)\frac{\sum_t\sum_f \mathrm{p}(t,f)\mathrm{p}_{l+1}(Z=k|t,f)}{\sum_{j=0}^{K-1}\sum_t\sum_f \mathrm{p}(t,f)\mathrm{p}_{l+1}(Z=j|t,f)}$$

After convergence, the marginal distribution $\mathrm{p}_{l^*}(t|Z=k)$ gives the distribution of energy of the kth component over time. Similarly, the marginal distribution $\mathrm{p}_{l^*}(f|Z=k)$ gives the distribution of energy of the kth component over frequency. Once the set of components $\{\tilde{k}\}$ belonging to the pth source has been determined, we can synthesize the spectrogram of this source as

$$|X_p(t,f)| = \frac{1}{c}\sum_{j\in\{\tilde{k}\}} \mathrm{p}_{l^*}(t,f|Z=j)\mathrm{p}_{l^*}(Z=j)$$

PLCA can be extended by the use of "kernel distributions" that allow the specification of marginal distributions with both time and frequency extent, and "entropic priors" that encourage sparsity in the resulting marginal distributions [53].

2.3.4 NMF

Instead of considering $|X(t,f)|$ as a probability distribution, we consider it at discrete frequencies mc_f and times nc_t with $m,n \in \mathbb{N}$, the entry at the mth row and nth column of matrix $V_{m,n} = |X(nc_t, mc_f)|$ with non-negative entries. We seek an approximate factorization of $\boldsymbol{V} \in \mathbb{R}_+^{M\times N}$ into matrices $\boldsymbol{W} \in \mathbb{R}_+^{M\times K}$ and $\boldsymbol{H} \in \mathbb{R}_+^{K\times N}$ such that

$$\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H} \tag{2.1}$$

This can be done by solving the program

$$\min \mathcal{D}(\boldsymbol{V}, \boldsymbol{W}\boldsymbol{H})$$

subject to

$$\boldsymbol{V} \geq \boldsymbol{0}$$

$$\boldsymbol{W} \geq \boldsymbol{0}$$

$$\boldsymbol{H} \geq \boldsymbol{0}$$

The particular choice of function (D) that measures divergence leads to different update equations in the iterative procedure for finding $\boldsymbol{W}$ and $\boldsymbol{H}$.

The Kullback-Leibler divergence [30]

The *Kullback-Leibler* divergence function for measuring the divergence between two matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ is

$$\mathcal{D}_{\mathrm{KL}}(\boldsymbol{X},\boldsymbol{Y}) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} X_{m,n}\log(Y_{m,n}) - X_{m,n} + Y_{m,n}$$

and can be minimized using the following update equations for the lth iteration[2]

$$H_{a,b}^{l+1} = H_{a,b}^{l}\frac{\sum_{j=0}^{M-1} W_{j,a}^{l}V_{j,b}^{l}/(W^{l}H^{l})_{j,m}}{\sum_{j=0}^{M-1} W_{j,a}^{l}}$$

$$W_{a,b}^{l+1} = W_{a,b}^{l}\frac{\sum_{j=0}^{N-1} H_{b,j}^{l+1}V_{a,j}^{l}/(W^{l}H^{l+1})_{a,j}}{\sum_{j=0}^{N-1} H_{b,j}^{l+1}}$$

The Itakura-Saito divergence [11]

Another divergence popular for factorizing spectrograms is the *Itakura-Saito* divergence

$$\mathcal{D}_{\mathrm{IS}}(\boldsymbol{X},\boldsymbol{Y}) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\frac{X_{m,n}}{Y_{m,n}} - \log\left(\frac{X_{m,n}}{Y_{m,n}}\right) - 1$$

This divergence is scale-invariant, meaning that $\mathcal{D}_{\mathrm{IS}}(c\boldsymbol{X},c\boldsymbol{Y}) = \mathcal{D}_{\mathrm{IS}}(\boldsymbol{X},\boldsymbol{Y})$, which makes it well suited for audio signals which have a large dynamic range. Put another way, divergences involving large values in $\boldsymbol{V}$ and $\boldsymbol{WH}$ will be weighted similarly to divergences involving small values, which is not the case with the Kulback-Leibler divergence. The Itakura-Saito divergence is minimized through the following update equations

$$\boldsymbol{H}^{l+1} = \boldsymbol{H}^{l}.\frac{\boldsymbol{W}^{lT}((\boldsymbol{W}^{l}\boldsymbol{H}^{l})^{\cdot -2}.\boldsymbol{V}^{l})}{\boldsymbol{W}^{lT}(\boldsymbol{W}^{l}\boldsymbol{H}^{l})^{\cdot -1}}$$

$$\boldsymbol{W}^{l+1} = \boldsymbol{W}^{l}.\frac{((\boldsymbol{W}^{l}\boldsymbol{H}^{l+1})^{\cdot -2}.\boldsymbol{V}^{l})\boldsymbol{H}^{l+1T}}{(\boldsymbol{W}^{l}\boldsymbol{H}^{l+1})^{\cdot -1}\boldsymbol{H}^{l+1T}}$$

Once convergence has been obtained the kth column of matrix $\boldsymbol{W}$ will contain values representing the level of activation of the kth component at the frequency corresponding

[2]It can be shown that these update equations are equivalent to those for PLCA [53].

to the row index and the kth row of $\boldsymbol{H}$ the level of activation of the kth component at the time corresponding to the column index. If the set of components $\{\tilde{k}\}$ belonging to the pth source has been determined, we can synthesize the spectrogram of this source as

$$|X_p(nc_t, mc_f)| = \sum_{j \in \{\tilde{k}\}} W_{:,j} H_{j,:}$$

2.3.5 Synthesis of factorized spectrograms

Synthesizing the original signal is less straightforward as the phase information contained in the STFT was discarded to obtain a non-negative spectrogram. We can simply use the original phases of the STFT used to compute the spectrum with the new magnitude information from $|X_p(t, f)|$ but the resulting signal may have artifacts due to the phase information of unwanted sources that remains in the STFT. A technique to lessen these artifacts using constrained Wiener filtering has been proposed in [29]. One may also choose to invert the spectrogram without any phase information by using an algorithm that iteratively reconstructs the phase part of the STFT while minimizing the error between the spectrogram of the reconstructed signal and its original power spectrum, starting from an initial guess [17]. Each iteration requires transforming the signal to its spectrogram and then back to a time-domain signal, requiring considerable computational effort.

2.3.6 Extensions and shortcomings

As with PLCA, certain extensions can be integrated into NMF to encourage particular solutions. For example, to promote sparseness in the resulting matrices, i.e., to encourage that fewer entries be non-zero, one can add the regularization term

$$\alpha \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} W_{m,k}^2 + \beta \sum_{k=0}^{K-1} \sum_{n=0}^{N-1} H_{k,n}^2$$

to the divergence to penalize matrices with large entries. α and β are terms that control the influence of this regularization. A variety of regularization terms are possible [5, ch. 3].

Equation 2.1 can be seen as the sum

$$V \approx \sum_{k=-0}^{K-1} \boldsymbol{w}_{:,k} \otimes \boldsymbol{h}_{k,:}$$

i.e., the sum of all the outer products between columns of $\boldsymbol{W}$ and rows of $\boldsymbol{H}$. From this perspective, we can see the columns of $\boldsymbol{W}$ as a collection of K spectral "templates" and the rows of $\boldsymbol{H}$ as their time-varying gains. What this means is, sounds exhibiting frequency-modulation require a large number of columns in $\boldsymbol{W}$. This large number of columns makes the organizing of templates into sources a difficult task.

2.4 An approach using amplitude- and frequency-modulation

Perceptual studies have shown that sounds modulated synchronously in amplitude or frequency are heard as one sound, whereas asynchronously modulated sounds are heard as distinct [35] [34]. Here we define the modulation of parameters θ_i and θ_j as being synchronous if they are given as functions of time, $\theta_i = f_i(t)$ and $\theta_j = f_j(t)$ and there is an affine transform $\mathscr{A}$ such that $\mathscr{A}\{f_i\}(t) \approx Af_j(t) + B$ where A and B are constants that do not vary with time (at least for the time that we observe the signal). If we can accurately measure these parameters and they are typical of the sounds we are trying to separate, then we can design techniques to reliably separate these sounds from acoustic mixtures. This involves picking those parameters classified as belonging to the same sound, discarding the rest, and resynthesizing from these parameters. The task of audio source separation therefore comprises the following tasks:

- Decide on a signal model for the sound of interest, with parameters that can be estimated and that are similar for similar sounds. We have chosen the additive sinusoidal model with a higher-order polynomial description of amplitude and phase.

- Locate regions in the signal where these signals are thought to be present using a peak-picking technique. Estimate the signal parameters at these locations. Here we use the Distribution Derivative Method [2] to estimate these parameters.

- Use these measurements as input to a partial tracking algorithm. We compare the effectiveness of the original peak matching procedure of McAulay and Quatieri [36] and a new linear programming formulation of the problem.

- Classify the partials and group them as sets of parameters coming from the same source. One of these sources will be the sound of interest. The classification is carried out on the frequency- and amplitude-modulation parameters. Principal components analysis is used to preprocess the data before classification.

- Choose a group of parameters (partials) representing the sound of interest and synthesize the separated signal from them. We compare the quality of synthesis for three models. The first assumes constant amplitude and linear phase at the analysis step resulting in linear amplitude and cubic phase at the synthesis step [36]. The second assumes quadratic amplitude and phase at the analysis step but constrains the amplitude to be cubic at the synthesis step. The final model assumes quadratic amplitude and phase at the analysis step resulting in a quartic model for phase and amplitude at the synthesis step.

It should be noted that the strategy for source separation relies on the same perceptual principle as in [6]. The work here differs in many respects. We have chosen to use solely the additive sinusoidal model as a model of the signals considered. In addition to frequency-modulation, we incorporate amplitude-modulation to aid in the source separation. The source separation itself differs technically from their approach as well. Chapter 7 uses only the amplitude-modulation to classify partials into sources. This is similar to the strategy explored in [31], but does not use a prior estimation of the fundamental frequencies of the sources.

Chapter 3

Signal Modeling

3.1 Introduction

To build tools to analyse and synthesize signals some structure must be imposed on the signals. The structures chosen can reflect something about the behaviour of these signals as observed in the field, as we will see with sinusoidal models. Other structures are chosen phenomenologically — we do not really know the underlying mechanism behind the production of these signals, but a particular structure is chosen for its mathematical or conceptual convenience, such as is the case when we consider higher-order models for sinusoidal phase and amplitude.

We begin the chapter with what could be seen as a mathematical analog of the musical score: time-frequency representations. Through this we will justify the adoption of a sinusoidal model for musical signals. Finding this inadequate to describe the signals of interest with sufficient quality, the sinusoidal model is generalized to incorporate modulations in frequency and amplitude. A technique is described to estimate the parameters of these more complex models which requires windows that are everywhere differentiable — we design a new window having desirable properties close to those of well-known optimal windows, but that is everywhere differentiable.

3.2 Time-frequency representations

As most musical instruments are resonating media, and excited resonating media are well described as linear time-invariant (LTI) auto-regressive (AR) structures, many popular

models of musical audio are some variation of this description [12]. Strictly speaking, incorporating moving-average (MA) structures into a model of musical signals could improve its quality, but such a model would preclude the sum-of-sinusoids model adopted later in this thesis.

An LTI auto-regressive structure is a signal that can be described using the following *difference equation*:

$$x(n) = \sum_{k=1}^{K} a_k x(n-k) + b_0 v(n)$$

Here x is the output of the system (what is heard or measured) and v is the input. K is the order of the model. Both are general functions of time which, in the case of properly sampled digital audio, can be considered at discrete times $n \in \mathbb{Z}$ without any loss of information [7, ch. 2]. $a_k, b_0 \in \mathbb{C}$ and are constants. Casually you can think of the output of the system at time n as being a linear combination of past outputs, plus some of the scaled input.

AR structures are excited in various ways: some are bowed, others struck, etc. To characterize the above structure we excite it with a simple signal, the *Kronecker delta*

$$\delta(n) = \begin{cases} 1 & n = 0 \\ 0 & \text{otherwise} \end{cases}$$

This Kronecker delta input will yield its *impulse response* from which we can derive many properties of the AR structure.

As an example take the case where $K = 1$ and $a_1 = r\exp(j\omega)$, $r, \omega \in \mathbb{R}$, $|r| < 1$. Then the difference equation is

$$x(n) = r\exp(j\omega)x(n-1) + v(n)$$

Exciting this with the Kronecker delta we get

$$\begin{aligned} x(0) &= 1 \\ x(1) &= r\exp(j\omega) \\ x(n) &= r^n \exp(j\omega n) \end{aligned}$$

which is a complex exponential starting at $n = 0$ and periodic in $n_T = \frac{2\pi}{\omega}$ multiplied by the real-valued exponential r^n. In other words, the output is a damped sinusoid. From this it is not hard to see that if we can estimate the coefficients a_k, we can then know the frequencies, amplitudes and damping factors of the sinusoids that are output when this structure is excited by an impulse (the Kronecker delta). This principle is presented as a motivation for the following techniques and is not pursued here. The interested reader is referred to [33] for more information.

An alternative method for determining the frequencies and amplitudes of sinusoids in mixture is to take the inner product of the signal with a complex exponential of known frequency

$$X(\omega) = \sum_{n=-\infty}^{\infty} x(n) \exp(-j\omega n)$$

The function $X(\omega)$ will be large if $x(n)$ contains a complex exponential of frequency ω and small if it does not, effectively indicating which sinusoidal functions are present in the signal. This transformation of a signal as a function of time n into one as a function of frequency ω is known as the *Discrete-time Fourier Transform* (DTFT).

To create a variety of pitches and timbres, typically the media of musical instruments are not static, but vary in time. That means the sets of sinusoids describing the state of the media and its excitation also change in time. To account for this we consider many small intervals of signal where we assume its characteristics are roughly static. We can then piece these time-intervals together afterwards to get a description of the signal in both time and frequency. To do this, we multiply the signal by a window w which makes the signal 0 outside of the interval of interest. We then test what sinusoids with frequencies ω are present at different times τ, giving a function of two variables

$$X(\tau, \omega) = \sum_{n=-\infty}^{\infty} x(n) w(n - \tau) \exp(-j\omega n) \tag{3.1}$$

This transformation of a signal of time n into one of time τ and frequency ω is known as the *Discrete-time Short-time Fourier Transform* (DTSTFT).

One further point about the window should be discussed. The Fourier transform of the product of two functions, like we have in Equation 3.1, is equal to the convolution of the Fourier transform of each function separately. If we denote the Fourier transform operator

as $\mathcal{F}$, for functions g and f we have

$$\mathcal{F}(gf) = \mathcal{F}(g) * \mathcal{F}(f)$$

where $*$ is the convolution operator. The value $X(\tau, \omega)$, which will be a complex number, can be seen as describing the amplitude and phase of a sinusoid at that frequency and time. If the Fourier transform of the window function is not purely real its imaginary part will offset the phase of this sinusoid. It is usually simpler to avoid this complication. The Fourier transform of a real even function

$$f(n) = f(-n), f(n) \in \mathbb{R}$$

is real, so we choose windows with this property. See [19, p. 52] for a concrete illustration of this.

3.3 Polynomial phase models

The DTFT and DTSTFT are very useful because they are invertible [45][1] and fast algorithms exist for their computation by digital computer [59]. If the presence of a sinusoid is determined, e.g., by finding τ^* and ω^* such that X is maximized, that signal can be removed or altered easily.

One drawback of these transforms is they only project onto sinusoidal functions of linear phase, i.e., functions of constant frequency. In general, musical signals are not linear combinations of sinusoids of constant frequency (consider, for example, vibrato). We could decide to project onto a different family of functions and considerable effort has been devoted to finding alternatives (see [28] for a review). In the case of musical signals, however, we have some prior information about the mechanics of the their production and can make certain assumptions about the underlying functions.

3.3.1 Sinusoidal Representations

Many musical acoustic signals are quasi-harmonic [12], meaning that they consist of a sum of sinusoids whose frequencies are roughly integer multiples of a fundamental frequency.

[1] Provided proper sampling in time and frequency.

For these kinds of signals, most of the energy can be attributed to sinusoids and so the signal can be described by a small number P of sinusoids with slowly varying amplitude and phase, plus some noise. The model is

$$x(n) = \sum_{p=1}^{P} A_p(n) \exp(j\phi_p(n)) + \varepsilon \tag{3.2}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, σ^2 quantifies the power of the noise, and $A_p(n), \phi_p(n) \in \mathbb{R}$ are the functions of amplitude and phase respectively for the pth sinusoid. In the following, we consider equivalent sinusoidal mixtures of complex-valued polynomial phase and log-amplitude exponentials

$$x(n) = \sum_{p=1}^{P} \exp(\mathcal{P}_p(n)) + \varepsilon$$

where $\mathcal{P}_p$ is the complex valued polynomial of order Q describing the argument of the pth sinusoid, i.e.,

$$\mathcal{P}_p(n) = \sum_{q=0}^{Q} c_q n^q$$

and $c_q \in \mathbb{C}$. Note the form in Equation 3.2 can be retrieved as

$$\mathcal{P}_p(n) = \log(A_p(n)) + j\phi_p(n) = \Re\{\mathcal{P}_p(n)\} + j\Im\{\mathcal{P}_p(n)\}$$

In Chapter 5 we will see how estimations of these parameters at different times can be interpolated to create higher-order phase and log-amplitude functions with improved synthesis quality.

3.4 Polynomial phase parameter estimation

Assuming this sinusoidal model, how can we then estimate the parameters describing a signal of interest? Recently a set of techniques have been developed that use some combination of derivatives of the analysis window w or the signal x to estimate the polynomial coefficients directly [18]. For this thesis we will only consider a technique that does not estimate derivatives of the signal and only requires a once-differentiable analysis window as it is relatively easy to implement and suits our purposes.

The following is adapted from [2]. Consider the inner product of the signal $x(n) = \exp(\mathcal{P}_p(n))$ and a known analysis *atom* $\psi(n)$

$$\langle x, \psi \rangle = \int_{-\infty}^{\infty} x(n)\overline{\psi}(n)dn$$

Differentiating with respect to n, we obtain by the product rule

$$\frac{dx}{dn}(n)\overline{\psi}(n) + x(n)\frac{d\overline{\psi}}{dn}(n) = \left(\sum_{q=1}^{Q} qc_q n^{q-1}\right) x(n)\overline{\psi}(n) + x(n)\frac{d\overline{\psi}}{dn}(n)$$

If $\psi(t)$ is 0 outside of some interval $n \in [-T, T]$ then

$$\sum_{q=1}^{Q} qc_q \int_{-T}^{T} n^{q-1} x(n)\overline{\psi}(n)dn + \left\langle x, \frac{d\overline{\psi}}{dn} \right\rangle = 0$$

If we define the operator $\mathcal{T}^\alpha : (\mathcal{T}^\alpha x)(n) = n^\alpha x(n)$ we can write

$$\sum_{q=1}^{Q} qc_q \left\langle \mathcal{T}^{q-1} x, \overline{\psi} \right\rangle = -\left\langle x, \frac{d\overline{\psi}}{dn} \right\rangle$$

From this we can see that to estimate the coefficients c_q, $1 \leq q \leq Q$ we simply need R atoms with $R \geq Q$ to solve the linear system of equations

$$\sum_{q=1}^{Q} qc_q \left\langle \mathcal{T}^{q-1} x, \overline{\psi}_r \right\rangle = -\left\langle x, \frac{d\overline{\psi}_r}{dn} \right\rangle \tag{3.3}$$

for $1 \leq r \leq R$. To estimate c_0 we write the signal we are analysing as

$$s(n) = \exp(c_0) \exp\left(\sum_{q=1}^{Q} c_q n^q\right) + \eta(n)$$

$\eta(n)$ is the error signal, or the part of the signal that is not explained by our model. We also define a function $\gamma(n)$, the part of the signal whose coefficients have already been estimated

$$\gamma(n) = \exp\left(\sum_{q=1}^{Q} c_q n^q\right)$$

Computing the inner product $\langle s, \gamma \rangle$, we have

$$\langle s, \gamma \rangle = \langle \exp(c_0)\gamma, \gamma \rangle + \langle \eta, \gamma \rangle$$

The inner-product between η and γ is 0, by the orthogonality principle [26, ch. 12]. Furthermore, because $\exp(c_0)$ does not depend on n, we have

$$\langle s, \gamma \rangle = \exp(c_0) \langle \gamma, \gamma \rangle$$

so we can estimate c_0 as

$$c_0 = \log\left(\langle s, \gamma \rangle\right) - \log\left(\langle \gamma, \gamma \rangle\right) \tag{3.4}$$

The estimation of the coefficients of a phase and log-amplitude polynomial using this method is known as the *Distribution Derivative Method (DDM)*.

3.5 Choosing atom ψ

As we are dealing with mixtures of sinusoids of small bandwidth, in addition to the finite-time support constraint, we desire atoms whose inner-product is only significant within a finite bandwidth of interest. To construct these atoms, we multiply the Fourier atom by the window w

$$\psi_{\tau,\omega}^{\mathcal{F}_w}(n) = w(n-\tau)\exp(-j\omega(n-\tau))$$

A good overview of different windows and their properties is given in [19]. We require that the window be at least once-differentiable and zero outside of a certain interval, therefore, somewhat informally, we require

$$\lim_{n \to T} \psi(n) = \psi(T) = 0$$

3.5.1 The Hann window

The *Hann* window possesses this property

$$w_h(n) = \begin{cases} 0.5 + 0.5\cos\left(\frac{n}{T}\pi\right) & -T \leq n \leq T \\ 0 & \text{otherwise} \end{cases}$$

The Hann window is a member of a class of windows constructed by summing scaled harmonically related cosine functions, subject to the constraint that the scaling coefficients sum to 1 so that the window have a value of 1 at $n = 0$. Letting $T = N/2$, where N is the length of the window

$$w(n) = \begin{cases} \sum_{m=0}^{M-1} a_m \cos\left(\frac{2\pi}{N} mn\right) & -\frac{N}{2} \leq n \leq \frac{N}{2} \\ 0 & \text{otherwise} \end{cases}$$

With $M = 2$ and $a_0 = a_1 = 0.5$, we have the Hann window.

The simple expression for its calculation and good trade-off between main-lobe width and side-lobe height make the Hann window a popular choice in many signal processing applications. The expression for its Fourier transform is such that fast digital implementations of windowing a signal by a Hann window involve no multiplies [19, p. 183]. A recursive implementation of the DTSTFT is possible when windowing with the Hann window, which is important for applications where little storage is available [58, p. 102]. In spite of all its merits, other windows have been proposed that have superior qualities, such as lower side-lobes.

3.5.2 Continuous Blackman-Harris windows

A family of windows with certain properties superior to the Hann window is the *Blackman-Harris* family of windows. These are also sum-of-cosine windows and so are easily tabulated. To design these windows, optimization techniques were used to search for coefficients giving minimum height of the highest side-lobe (maximum out-of-band rejection) [47]. The 4-term window whose coefficients a are listed in Table 3.1 has a maximum side-lobe level of -92 dB, just shy of the quantization noise of a 16-bit linear pulse code modulated signal (-96 dB). As can be seen in Figure 3.2, this window has a very large main-lobe which means

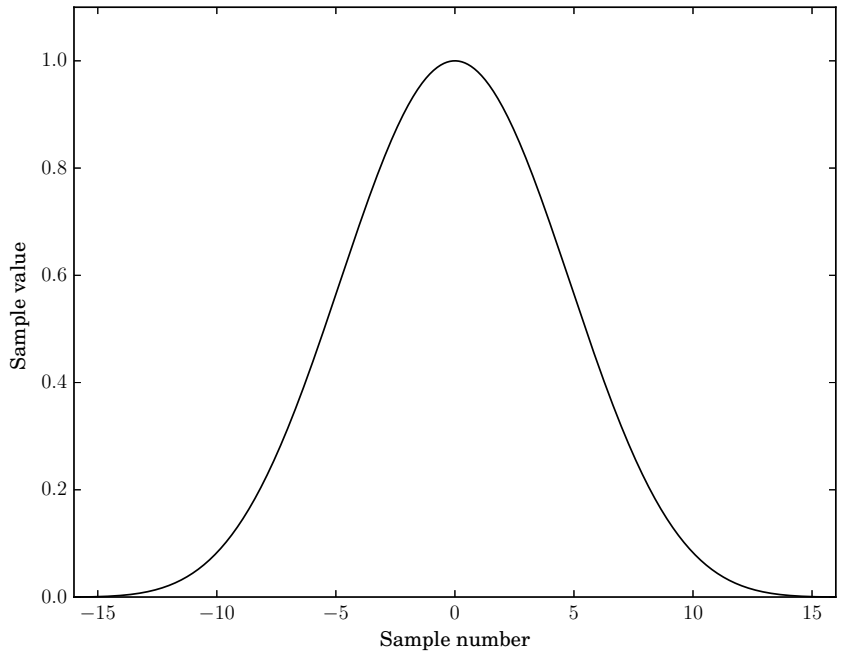

Fig. 3.1: **Minimum 4-Term Blackman-Harris: Time-domain**

two sinusoids of similar frequency will be difficult to resolve. Furthermore, as shown in Figure 3.3, the window has a discontinuity at its boundaries, e.g., $w\left(\frac{N}{2}\right) \neq 0$, and is not once-differentiable. In any case the window is valuable in that it effectively nulls any influence of signals outside of a bandwidth of interest. The shape of the Blackman-Harris window in the time- and frequency-domains can be seen in Figures 3.1 and 3.2 respectively.

It should be clarified that when we compare the widths of the main-lobes of two windows, we compare two windows of the same length. Of course, the bandwidth of a window can also be decreased by increasing its length, at the expense of time-resolution. When searching for windows superior to the Hann window, we are motivated by our ability to describe the

Table 3.1

Window	a_0	a_1	a_2	a_3
Minimum	0.35857	0.48829	0.14128	0.01168
$\mathcal{C}^1$	0.35874	0.48831	0.14127	0.01170

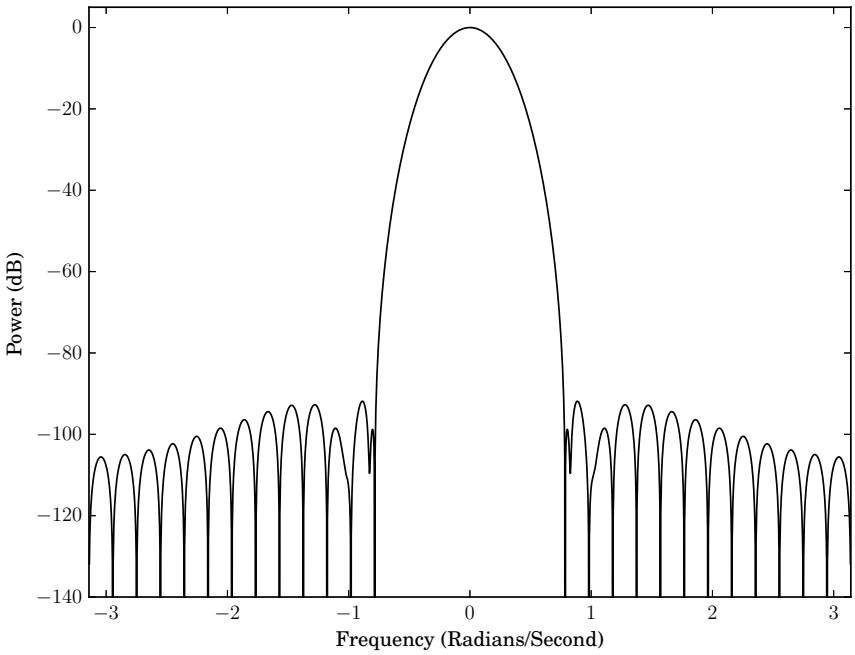

Fig. 3.2: **Minimum 4-Term Blackman-Harris: Frequency-domain**

signal with more detail between two analysis frames than would be possible with a simple linear-phase sinusoid model. Using a longer window to decrease the main-lobe width is not problematic in this case, but we would still like a high level of signal rejection outside of the bandwidth of interest for improved estimation accuracy of the signal parameters. For this reason, we search for windows that have very low side-lobe height and are also once-differentiable, without caring so much about the width of the main-lobe. To find a window with properties similar to the 4-term Blackman-Harris window but without a discontinuity, we solve the optimization problem

$$\min ||a - \tilde{a}||_2$$

subject to

$$w_{\tilde{a}}\left(\frac{N}{2}\right) = w_{\tilde{a}}\left(\frac{-N}{2}\right) = 0$$

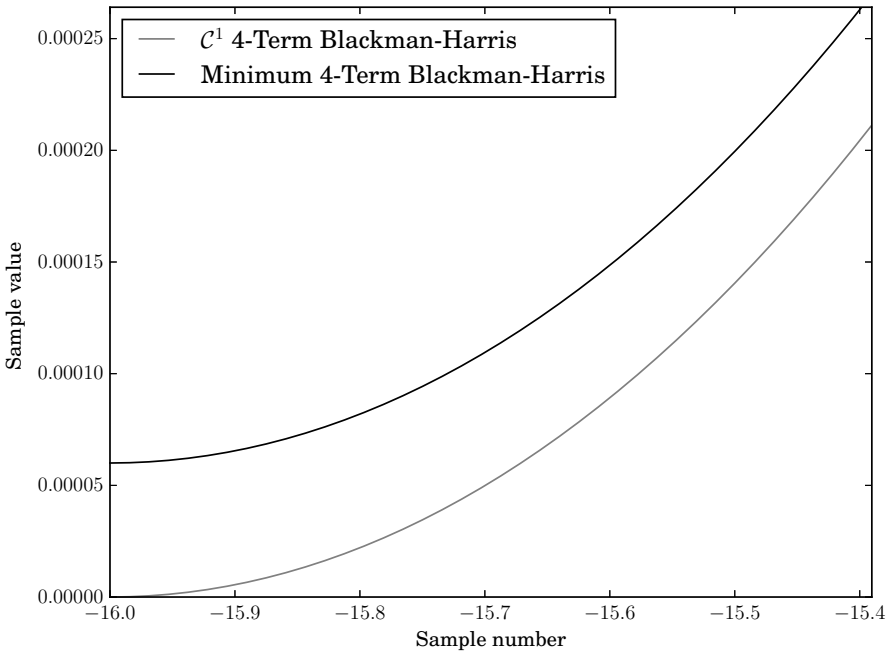

Fig. 3.3: **Comparison of endpoints of window in time-domain.** Here we observe that the $\mathcal{C}^1$ 4-term Blackman-Harris window approaches 0 at the boundary of its time support, whereas the Minimum 4-Term Blackman-Harris window does not.

$$\sum_{m}^{M-1} a_m = 1$$

where

$$w_{\tilde{a}}(n) = \begin{cases} \sum_{m=0}^{M-1} \tilde{a}_m \cos\left(\frac{2\pi}{N} mn\right) & -\frac{N}{2} \leq n \leq \frac{N}{2} \\ 0 & \text{otherwise} \end{cases}$$

The solution $\tilde{a}^*$ is given in Table 3.1 and time- and frequency-domain plots are given in Figures 3.4 and 3.5 respectively. This window will be referred to as the $\mathcal{C}^1$ 4-Term Blackman-Harris window. Some figures of merit for the two windows are compared in Table 3.2 in a similar fashion to [19]. We see from comparing Figures 3.1 and 3.2 with 3.4 and 3.5 that the $\mathcal{C}^1$ 4-Term Blackman-Harris window is not too different from the Minimum 4-term Blackman-Harris window, but has the additional desirable property of differentiability everywhere in its domain. A comparison of the windows's endpoints is presented in Figure 3.3.

Re-emphasized the fact that 4-term window clearly goes down to zero.

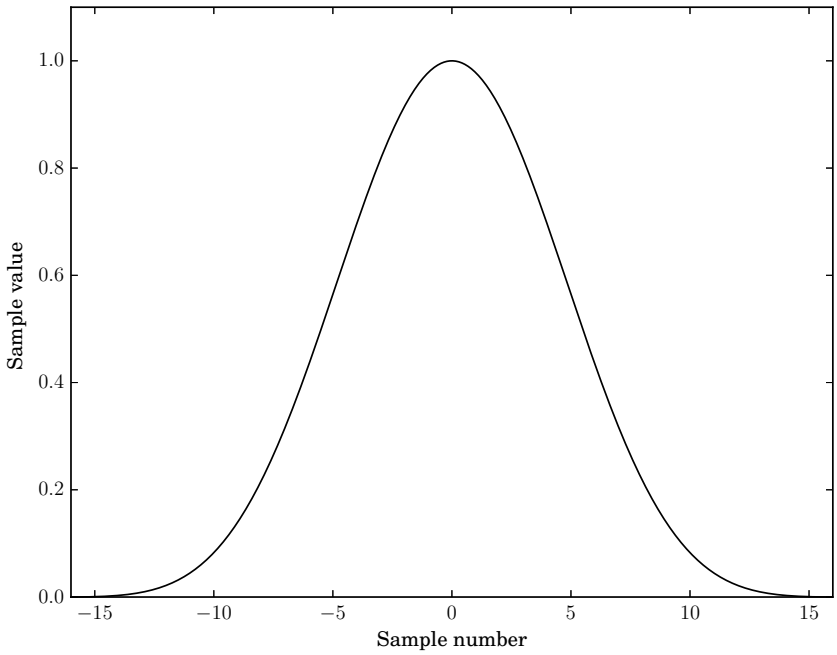

Fig. 3.4: $\mathcal{C}^1$ **4-Term Blackman-Harris: Time-domain**

3.6 Conclusion

In this chapter we have developed the rationale for adopting the sinusoidal model when analysing musical signals. In turn we have presented some techniques for estimating the parameters of these signals. Obviously these techniques only work as well as their assumptions are true — for the best results we should use these techniques only on signals that indeed contain sinusoids. For the signals considered in this thesis, we assume this to be true.

Table 3.2

Window	Highest side-lobe level (dB)	6-dB bandwidth in bins	Side-lobe fall-off (dB/octave)
Minimum	-92	2.72	6
$\mathcal{C}^1$	-90	2.66	12

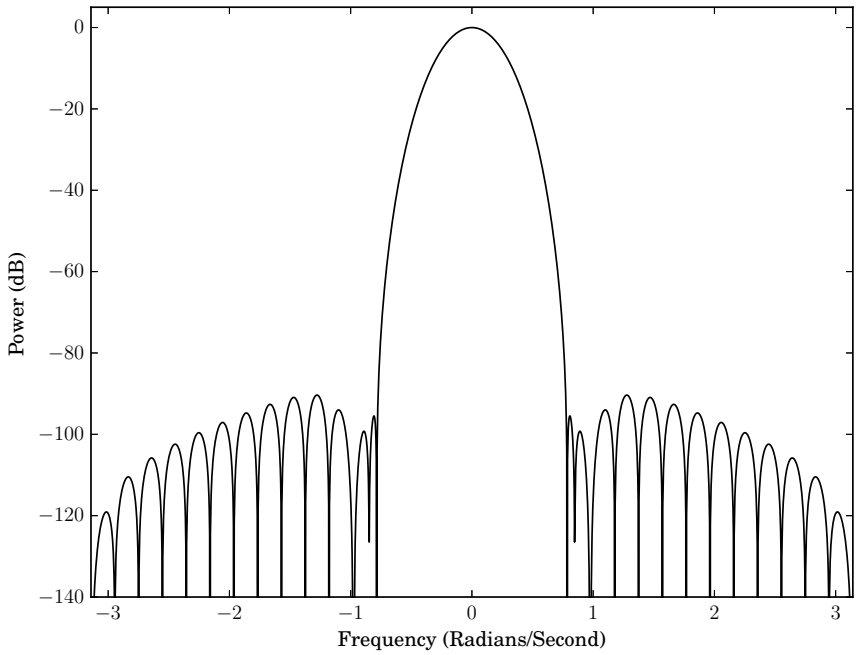

Fig. 3.5: $\mathcal{C}^1$ **4-Term Blackman-Harris: Frequency-domain**

The techniques presented in this chapter are typically used to estimate the parameters of short signals as we see in the use of window functions to limit the time-frequency extent of our analysis. For the source separation problem we are interested in larger sinusoidal objects — partials — whose global properties are more readily classified. We could use a very large analysis window and a high-order polynomial for phase when solving for the coefficients, but the size of the linear system to be solved in Equation 3.3 will increase quadratically in the order of the model. Furthermore, it is difficult to account for situations where the signal is corrupted or briefly absent. In these situations we may prefer to use interpolation to reconstruct the signal in the corrupted region. For these reasons, we prefer to make multiple estimations of the parameters of low-order models and connect those estimations thought of as belonging to a single partial. We will see in Chapter 5 that this will allow postulating a higher-order phase model. Before that is possible, however, we must determine how to connect multiple estimations to form partials.