Department of Communication Technology
Aalborg University

Institute of Electronic Systems

Title:
Perceptual Unitary ESPRIT Algorithm

Project Period:
February 1st, 2003 to
June 3rd, 2003

Project Group:
03gr1048

Authors:
Mikkel N. Schmidt
Jens Brøsted Seiersen

Supervisors:
Jesper Jensen
Søren Holdt Jensen

Number of Copies:
9

Number of Pages:
104

Abstract:
In this dissertation, we propose a new algorithm for estimating the most perceptually relevant parameters of a constant amplitude sinusoidal audio model. We propose to incorporate a psychoacoustic model in the Unitary ESPRIT algorithm which is a subspace based parameter estimation method providing accurate parameter estimates at a low computational cost. We thoroughly study the Unitary ESPRIT algorithm and the MPEG-1 Psychoacoustic Model 1. Then, we discuss and propose methods to include psychoacoustic information in the Unitary ESPRIT algorithm. Finally, we study the characteristics of the proposed algorithm for a wide range of different deterministic and real speech and audio signals. The proposed algorithm is found to be a robust, accurate, and efficient method for estimating perceptually relevant parameters for constant amplitude sinusoidal audio modeling.

Preface

This Master's thesis is submitted in partial fulfillment of the requirements for the degree of Master of Science in Electronic Engineering. The work described in this thesis was carried out between Feburary and June 2003 on the speech communication specialization at Aalborg University, and builds on the work by Jensen et al. [1] on incorporation psychoacoustic priciples in subspace based parameter estimation for sinusoidal audio coding.

We would like to thank our two supervisors, Jesper Jensen, Delft University of Technology, and Søren Holdt Jensen, Aalborg University, for their advice and support.

A CD-ROM containing most of the referenced papers, wave-file demonstrations, and MATLAB source code implementing the proposed algorithms is enclosed.

The project proposal written by Jesper Jensen and Søren Holdt Jensen is included in appendix E.

Aalborg University, January 3rd 2003.

Mikkel N. Schmidt

Jens Brøsted Seiersen

Contents

Chapter 1

Introduction and foundations

"The beginning of knowledge is the discovery of something we do not understand."

Frank Herbert (1920 – 1986)

In this chapter: *We start by giving a general introduction to the work described in this dissertation, briefly summarizing some of the previous work on which this builds. This leads us to defining a set of goals we seek to achieve in this project. Finally, we give an outline of the structure of remainder of this dissertation and introduce some common notation used throughout.*

1.1 Introduction

In sinusoidal modeling of audio signals, the signal of interest is divided into segments and each segment is modelled as a finite sum of sinusoids with different frequencies, amplitudes, and phases. Some models include other parameters such as a temporal envelope given by an exponential damping factor [1] [2] [3].

Sinusoidal models have found many applications in digital audio signal processing. Sinusoidal modeling has proven to be efficient at modeling speech signals [4] of which some signal regions are known to be periodic and quasi stationary. More recently it has been shown that sinusoidal models also can be used in low bit-rate audio coding [5]. Because sinusoidal models provide a very flexible signal representation they are well suited for performing speech transformations such as time-scale and pitch-scale modifications [4]. For this reason, sinusoidal models have been used in many types of music synthesis [6]. Also, sinusoidal models have been applied in the enhancement of speech degraded by additive noise [7].

For use in low bit-rate audio coding, sinusoidal modeling can be used as one of the central components in a hybrid coding scheme. An example of a hybrid audio model is the sines+transients+noise (STN) model [5], [8], which consists of three components. First, the tonal part of the signal is modelled by a sinusoidal model. The residual, i.e. the part of the signal not modelled by the sinusoidal model, is then passed to a transient coder, which explicitly models the

transients in the signal. Finally, the residual from the transient coder is modelled as a filtered noise process.

Because of the masking properties of the human auditory system, some frequency components will be less perceptually relevant because they are masked by other components in the signal. Thus, when sinusoidal models are used in audio coding, the main challenge is, based on a given signal segment, to estimate a limited set of parameters which best describe the signal segment in a perceptual sense.

A well known class of parameter estimation techniques are the so-called subspace based methods. For a thorough unified approach to subspace based signal analysis methods, confer e.g. [9]. These methods can provide robust and accurate parameter estimates, however the perceptual relevance of the estimates are not taken into account.

Only recent research has shown that subspace based parameter estimation techniques can be combined with a perceptual distortion measure and thus can be used to extract the most perceptually relevant signal parameters [1]. Jensen et al. [1] have showed how to combine a subspace method known as HTLS[1] first described by Van Huffel [10] with a recently developed psychoacoustic model [11], resulting in a model denoted the perceptual exponential sinusoidal model (P-ESM). Subjective comparison tests showed that signals modelled with the proposed algorithm "were of considerable higher perceptual quality" [1] than those modelled with the HTLS algorithm.

However, regarding the algorithm proposed by Jensen et al., two possible drawbacks are identified:

1. The signal model employed is based on exponentially damped sinusoids. Although the exponential sinusoidal model has been shown to outperform the constant amplitude sinusoidal model with regard to modeling transient segments of speech and audio [2], [3], [12], it requires as much as four parameters for each sinusoid: amplitude, frequency, phase, and damping factor. Thus, for stationary segments, where the damping factor is of little use, the constant amplitude sinusoidal model provides a more compact signal representation.

2. The HTLS algorithm used by Jensen et al. is computationally quite complex. Although the algorithm provides very accurate parameter estimates, subspace based parameter estimation methods which provide even better estimation accuracy at an equal or lower computational cost do exist [13].

This project aims at overcoming these two drawbacks by means of incorporating a psychoacoustic model in the so-called Unitary ESPRIT algorithm [13]. Unitary ESPRIT is a subspace based algorithm for estimating parameters of constant amplitude sinusoids. Since the basis functions in Unitary ESPRIT are constant amplitude sinusoids, exponential damping factors are not included in the model. Also, the Unitary ESPRIT algorithm provides a better estimation accuracy at a computational cost equal to that of the HTLS algorithm used in [1].

[1] The HTLS algorithm is based on a Hankel data matrix and employs total least squares techniques. Hence the name, HTLS.

1.2 Project goals

The goal of this work is to design an algorithm which incorporates a psychoacoustic model in the Unitary ESPRIT algorithm for use in a sinusoidal audio coder. To do this, we wish to

- Perform extensive studies of the Unitary ESPRIT algorithm.
- Study a well known psychoacoustic model, namely the MPEG-1 Psychoacoustic Model 1.
- Design an novel algorithm which incorporates a psychoacoustic model in the Unitary ESPRIT algorithm.
- Analyse the proposed new algorithm for a wide range of deterministic and natural signals.

1.3 Dissertation outline

The remainder of this dissertation is structured as follows:

Sinusoidal Modeling: We start by a brief discussion of different classes of audio coders to give an overview of how the sinusoidal model relates to other methods. Then, we present the sinusoidal signal model. We discuss a wide variety of parameter estimation techniques and different segmentation and reconstruction methods.

Unitary ESPRIT: We extensively describe the Unitary ESPRIT algorithm — a subspace based algorithm for estimating the frequencies in a sinusoidal model. We start by describing the ESPRIT algorithm which is the basis for the Unitary ESPRIT algorithm. Then we move on to describe the details of the Unitary ESPRIT algorithm which provides an increased estimation accuracy while having equal or lower computational complexity compared to the ESPRIT algorithm.

Psychoacoustic Model: We start by reviewing the psychoacoustic phenomena on which perceptual masking models are based. Then we thoroughly study a well known psychoacoustic model, namely the MPEG-1 Psychoacoustic Model 1.

Perceptual Unitary ESPRIT: We present a novel approach to estimating the most perceptually relevant parameters in a sinusoidal audio model: We propose a method which incorporates the perceptual model from the MPEG-1 standard in the Unitary ESPRIT algorithm. We start by discussing how the perceptual distortion of a signal can be measured, based on information from a psychoacoustic model. Then, we introduce methods for incorporating perceptual knowledge in the estimation of signal frequencies by means of Unitary ESPRIT as well as in the estimation of amplitudes and phases.

Tests: We perform a series of experiments with the proposed Perceptual Unitary ESPRIT algorithm for a wide range of deterministic and real speech and audio signals. To examine the effects of the psychoacoustic model, we compare the Perceptual Unitary ESPRIT algorithm with the Unitary ESPRIT algorithm. Then, we relate the proposed algorithm to the P-ESM algorithm introduced by Jensen et al. [1].

Conclusions and Discussion: We summarize the results obtained in this work and discuss strengths and weaknesses of the proposed algorithm.

1.4 Notation

Throughout this dissertation we will use the following notation:

$a_{i,j} = \{\boldsymbol{A}\}_{i,j}$	The i,jth element of the matrix $\boldsymbol{A}$,
$\boldsymbol{a}_{i,:}$	The ith row of $\boldsymbol{A}$,
$\boldsymbol{a}_{:,j}$	The jth column of $\boldsymbol{A}$,
$\boldsymbol{A}^H$	Hermitian (complex conjugate) transpose of $\boldsymbol{A}$,
$\boldsymbol{A}^*$	Complex conjugate of $\boldsymbol{A}$,
$\boldsymbol{A}^\dagger$	The pseudoinverse of $\boldsymbol{A}$,
$tr(\boldsymbol{A})$	Trace of $\boldsymbol{A}$, i.e. the sum of the diagonal elements of $\boldsymbol{A}$,
$diag(a_1, \ldots, a_m)$	An $m \times m$ matrix with $a_1, \ldots, a_m$ on the main diagonal and zeros elsewhere,
$\mathcal{R}(\boldsymbol{A})$	The range (column space) of $\boldsymbol{A}$,
$\mathcal{N}(\boldsymbol{A})$	The null space of $\boldsymbol{A}$,
$\lambda_i(\boldsymbol{A})$	The ith eigenvalue of $\boldsymbol{A}$,
$rank(\boldsymbol{A})$	Rank of $\boldsymbol{A}$, i.e. the number of independent columns (or rows) of $\boldsymbol{A}$,
$Re(\boldsymbol{A})$	Real part of $\boldsymbol{A}$,
$Im(\boldsymbol{A})$	Imaginary part of $\boldsymbol{A}$,
$x(k) * y(k)$	The convolution of $x(k)$ and $y(k)$,
$\|\| \cdot \|\|_2$	The ℓ_2 norm,
$\|\| \cdot \|\|_F$	The Frobenius norm,
k	Time index,
l	Frequency bin index.

Chapter 2

Sinusoidal modeling

“Get your facts first, and then you can distort them as much as you please.”

Mark Twain (1835 – 1910)

In this chapter: *We start by a brief discussion of different classes of audio coders to give an overview of how the sinusoidal model relates to other methods. Then, we present the sinusoidal signal model. We discuss a wide variety of parameter estimation techniques and different segmentation and reconstruction methods.*

2.1 Audio coding taxonomy

To give an understanding of how the sinusoidal audio model fits in the field of audio coding, we here give a brief overview of the taxonomy of audio coders.

When transmitting digital audio over communication channels or storing audio on a digital storage medium, it is often necessary or most cost efficient to compress the signal in order to optimally utilize the capacity of the communication channel or the storage medium. For this purpose, a number of audio compression schemes have been devised [14], [15].

Audio compression algorithms can generally be divided in two groups: lossless and lossy algorithms.

Lossless audio coders exploit redundancies in the signal to give a bit-exact but more compact description of the signal. However, the compression obtainable by lossless coding is limited. Because of the stochastic nature of audio signals, they do no lend themselves well to lossless compression.

Lossy audio coders achieve better compression by allowing some noise or pertubations of the signal. As long as the noise and signal pertubations are kept below the audible limit, they will not be detectable by the human auditory system, and thus the audio compression can be virtually unnoticable.

Lossy audio coders can be divided in two primary classes: waveform coders and parametric coders.

Waveform coders attempts to accurately describe the waveform of the signal. An example is the time/frequency transform coder, in which a signal is divided into a number of frequency bands which are quantized individually such that the quantization noise can be shaped in a perceptually optimal manner. Waveform coders have shown good results for medium to high bit-rate audio coding.

Parametric coders assume an underlying model of the signal which can be exploited. A typical example is voice coders (vocoders) which are based on a parametric model of the human speech production system. Parametric coders have shown good results for medium to low bit-rate audio coding.

Audio coding using the sinusoidal model is an example of a lossy parametric method.

2.2 The sinusoidal signal model

In the sinusoidal signal model, a signal is represented by a series of consecutive possibly overlapping segments, each modeled by a finite sum of sinusoids of different frequencies, amplitudes, and phases.

2.2.1 Assumptions

There are two primary underlying assumptions in the sinusoidal model:

1. Each signal segment can be adequatly represented by a finite sum of sinusoids.
2. The parameters of the signal change slowly, such that the model parameters for one signal segment can be considered constant.

The first assumption is valid for such signals as voiced regions of speech and musical instrument sounds such as trumpets and violins. The frequency spectrum of these types of signals consist of distinct harmonically related frequency components. The second assumption can be considered valid, when the signal segments are chosen such that the segment rate is high compared to the dynamics of the analyzed signal.

2.2.2 Signal segment model

The sinusoidal model for each signal segment consists of a deterministic part, a sum of sinusoids, plus a stochastic part, a noise term, which is included to account for any part of a real life signal which is not adequatly modeled by the sum of sinusoids. Often, the noise term is assumed stationary with zero mean.

Consider a signal, $x(k)$, consisting of D sinusoids with constant amplitudes and phases in additive zero-mean stationary real Gaussian noise, $n(k)$, with known covariance. If the amplitude,

normalized frequency, and phase of the ith sinusoid are given by S_i, Ω_i, and ϕ_i respectively, the signal can be written as

$$x(k) = \sum_{i=1}^{D} S_i \cos(\Omega_i k + \phi_i) + n(k). \tag{2.1}$$

Equivalently, using the Euler relation

$$S\cos(\Omega k + \phi) = \frac{S}{2}\left(e^{j\Omega k + j\phi} + e^{-j\Omega k - j\phi}\right). \tag{2.2}$$

we may express the signal as a sum of $d = 2D$ complex sinusoids (cisoids). Let s_i denote the complex amplitude of the ith cisoid and ω_i its normalized frequency. Then, the signal can be written as

$$x(k) = \sum_{i=1}^{d} s_i z_i^k + n(k), \tag{2.3}$$

where the signal poles $z_i = e^{j\omega_i}$ lie on the unit circle. Since $x(k)$ is a real signal, the signal poles, z_i, occur in complex conjugate pairs or on the real axis. If the signal poles do not lie on the unit circle, this signal model corresponds to a sum of exponentially damped cisoids.

For use in the derivation of the Unitary ESPRIT algorithm, we need to express the signal model in matrix–vector notation. Equivalent to the signal model above, we may write

$$\boldsymbol{x}(k) = \boldsymbol{A}\boldsymbol{s}(k) + \boldsymbol{n}(k), \tag{2.4}$$

where $\boldsymbol{x}(k), \boldsymbol{n}(k) \in \mathbb{R}^m$ are given by

$$\begin{aligned} \boldsymbol{x}(k) &= [x(k), \ldots, x(k+m-1)]^T, \\ \boldsymbol{n}(k) &= [n(k), \ldots, n(k+m-1)]^T, \end{aligned}$$

and the complex amplitudes of the d cisoids are given by $\boldsymbol{s}(k) \in \mathbb{C}^d$

$$\boldsymbol{s}(k) = [s_1 e^{jk\omega_1}, \ldots, s_d e^{jk\omega_d}]^T = \boldsymbol{\Phi}^k \boldsymbol{s}_0,$$

where

$$\begin{aligned} \boldsymbol{\Phi} &= diag(e^{j\omega_1}, \ldots, e^{j\omega_d}), \\ \boldsymbol{s}_0 &= [s_1, \ldots, s_d]^T. \end{aligned}$$

$\boldsymbol{A}$ is a matrix where each column, $\boldsymbol{a}_{:,i}$, corresponds to each of the d complex sinusoids

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ e^{j\omega_1} & e^{j\omega_2} & \ldots & e^{j\omega_d} \\ e^{j2\omega_1} & e^{j2\omega_2} & \ldots & e^{j2\omega_d} \\ \vdots & \vdots & & \vdots \\ e^{j(m-1)\omega_1} & e^{j(m-1)\omega_2} & \ldots & e^{j(m-1)\omega_d} \end{bmatrix} \in \mathbb{C}^{m\times d}. \tag{2.5}$$

Definition 2.1. *If the matrix $\boldsymbol{V} \in \mathbb{C}^{p\times q}$ can be written in the following form it is said to be Vandermonde [16, p. 183]*

$$\boldsymbol{V} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ v_0 & v_1 & \cdots & v_{q-1} \\ \vdots & \vdots & & \vdots \\ v_0^{p-1} & v_1^{p-1} & \cdots & v_{q-1}^{p-1} \end{bmatrix}.$$

□

We note that $\boldsymbol{A}$ is Vandermonde and the columns of $\boldsymbol{A}$ are cisoids with unit amplitude and zero phase. The columns of $\boldsymbol{A}$ can be seen as the basic components of which the signal vector, $\boldsymbol{x}(k)$, is a linear combination. In the sequel we will assume that the d frequencies are distinct

$$\omega_i \neq \omega_j \quad \text{for all} \quad i \neq j,$$

and within the Nyquist bound

$$-\pi < \omega_i < \pi,$$

and that $m > d$ so that $\boldsymbol{A}$ is a tall matrix. Since all the frequencies are different, all the columns of $\boldsymbol{A}$ are independent and thus $\boldsymbol{A}$ is full rank (cf. e.g. [17, p. 409]) i.e. $rank(\boldsymbol{A}) = d$.

2.2.3 Signal matrix

For use in subspace based signal analysis, we need to express the signal model in terms of a signal matrix. When analyzing a signal block of length $N = m+M-1$ we may construct a signal matrix of size $m \times M$

$$\boldsymbol{X}(k) = [\boldsymbol{x}(k), \ldots, \boldsymbol{x}(k+M-1)] \in \mathbb{R}^{m \times M}.$$

This signal matrix is Hankel structured, i.e. it has constant anti-diagonals

$$\boldsymbol{X}(k) = \begin{bmatrix} x(k) & x(k+1) & \cdots & x(k+M-1) \\ x(k+1) & x(k+2) & \cdots & x(k+M) \\ \vdots & \vdots & & \vdots \\ x(k+m-1) & x(k+m) & \cdots & x(k+N-1) \end{bmatrix}. \tag{2.6}$$

Expressing the signal model from (2.4) in terms of the signal matrix yields

$$\boldsymbol{X}(k) = \boldsymbol{A}\boldsymbol{S}(k) + \boldsymbol{N}(k), \tag{2.7}$$

where the complex amplitude matrix and the noise matrix are given by

$$\boldsymbol{S}(k) = [\boldsymbol{s}(k), \ldots, \boldsymbol{s}(k+M-1)] \in \mathbb{C}^{d \times M},$$

and

$$\boldsymbol{N}(k) = [\boldsymbol{n}(k), \ldots, \boldsymbol{n}(k+M-1)] \in \mathbb{R}^{m \times M},$$

respectively. These matrices are also Hankel structured.

2.3 Segmentation

For a signal with slowly varying parameters, the assumption of stationarity only holds for short time segments. Therefore, before the signal is analyzed, it is segmented into frames

$$x_i(k) = \sqcap_N(k - ip)x(k), \tag{2.8}$$

where i denotes the frame index, p is the frame stride i.e. the number of samples between consecutive frames, and $\sqcap_N(k)$ is the rectangular window function defined as

$$\sqcap_N(k) = \begin{cases} 1 & k = 0, 1, \ldots, N-1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.9}$$

The length of the signal segments must be chosen such that the parameters of the signal can be considered constant in the segment. In order to better model signals with varying parameter–dynamics, a dynamic time segmentation scheme can be utilized to adapt the length of the analysis window to the signal. This can provide a more compact signal representation; however, this comes at the expense of an increased computational cost, since a large number of different length windows must be analyzed [18].

2.4 Parameter estimation

Estimation of parameters for the sinusoidal model can be done using a variety of different methods. In the following, we provide an abridged overview of the most important methods.

The parameter estimation in the sinusoidal model can be based on either parametric or non-parametric methods. In non-parametric parameter estimation methods, the signal is analyzed without regard to the underlying signal model using e.g. the discrete Fourier transform. In parametric parameter estimation methods, such as the class of subspace based methods, the assumption that the signal is generated by an underlying model is exploited.

2.4.1 Fourier transform based methods

The mathematical tool used by most non-parametric parameter estimators is the discrete Fourier transform (DFT) which transforms a windowed signal segment into the frequency domain using [19]

$$X(\omega) = \sum_{k=0}^{N-1} w(k)x(k)e^{-j\omega k}. \tag{2.10}$$

Because the signal segment has a limited length, the frequency spectrum will have a limited resolution corresponding to the reciprocal of the segment length. Often, the signal segment is zero-padded prior to taking the DFT which increases the frequency resolution but does not add new information. It corresponds to interpolating the frequency spectrum. Frequencies which are harmonically related to the sampling frequency will correspond to a discrete delta function in the frequency spectrum, whereas the spectrum for other frequencies will have the shape of the sinc function. When the signal is multiplied by some window function, it correponds to convolving the spectrum with the DFT of the window function. Thus, when truncated and windowed, the sinusoids in the signal adopt the spectral shape of the window, offset by their frequency. The spectral shape of the window function consist of a mainlobe and sidelobes that spread out over the spectrum. The width of the mainlobe and the maximum level of the neighboring sidelobes are determined by the shape of the window. These factors influence on each other so that a narrow mainlobe, corresponding to high frequency resolution, results in relativly large sidelobes, i.e. more spectral energy is leaked into other frequency bands. This combined with the relationship between frequency and time resolution, makes the DFT a simple and efficient tool, which albeit requires a compromize between time and frequency resolution.

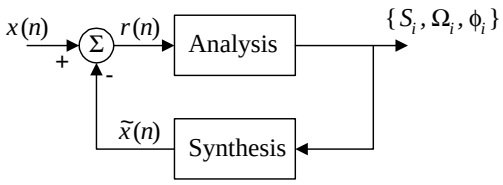

Figure 2.1: *The analysis-by-synthesis approach to sinusoidal modeling. The signal $x(n)$ is analysed and the parameters of the dominant sinusoid is found $\{S_i, \Omega_i, \phi_i\}$. These parameters are then used to synthesize a signal approximation $\tilde{x}(n)$ which is subtracted from the original signal to give a residual $r(n)$. The analysis is then performed on the residual iteratively each time estimating one new sinusoid.*

Peak picking is a method where the sinusoids in the signal are estimated by finding peaks in the spectrum. This can be done using a varity of heuristic methods, such as checking for negative zero-crossings in the derivative of the spectrum [4], or fitting sinusoid-shaped curves to the peaks [6]. Some of the drawbacks of these methods is that they depend on a high frequency resolution to properly discriminate between sinusoids at closely spaced frequencies. Sidelobes can also present a problem as they can be wrongly identified as sinusoids. When a suitable number of sinusoids has been identified, their amplitudes and phases can be found directly from the DFT.

Matching pursuit is a method which uses a analysis-by-synthesis scheme to identify the parameters of the signal [18] (see figure 2.1). This is done by iteratively analysing the signal $x(n)$, identifying the most powerful sinusoid and estimating its parameters, $\{S_i, \Omega_i, \phi_i\}$. The parameters are then used to synthesize a estimated version of the signal $\tilde{x}(n)$. The reconstructed signal is then subtracted from the original signal, creating a residual signal $r(n)$, which then is analysed to find the next most powerful sinusoid.

The model order (the number of sinusoids identified) can either be fixed or dynamic. When using dynamic model order, the residual, i.e. the modeling error, can be reduced to a arbitrary desired level in each signal segment.

2.4.2 Subspace based methods

The class of subspaces based parameter estimation methods can be divided into three different groups: subspace fitting methods, single-shift invariant methods, and orthogonal vector methods[9]. These methods all rely on the notion of signal and noise subspaces (see appendix C). Unified descriptions of subspace based signal analysis are given in e.g. [20] and [9].

Subspace fitting methods seek to find the signal model which matches the signal in the best possible way [9]. Thus, they seeks to minimize the following expression over all matrices $\boldsymbol{S}$ and all Vandermonde matrices $\boldsymbol{A}$, both of rank d [20]

$$\hat{\boldsymbol{A}},\ \hat{\boldsymbol{S}} = \arg\min_{\boldsymbol{A},\ \boldsymbol{S}} ||\boldsymbol{X} - \boldsymbol{A}\boldsymbol{S}||_F^2 \tag{2.11}$$

This is a nonlinear optimization problem and requires in general a multidimensional gradient search. However computationally expensive, the subspace fitting techniques can achieve excellent parameter estimates.

Examples of algorithms belonging to the class of subspace fitting methods include the "deterministic maximum likelihood method" [9], "method of direction estimation" (MODE) [9], [21], [22], and multiple invariance ESPRIT [23], [24].

Single shift-invariant methods are based on constructing two signal matrices which are shifted in time by one sample. The shift-invariance relation between these two matrices is then used to estimate the parameters of the signal, "hence ignoring any further (shift-invariant) structure that $\boldsymbol{A}$ or $\boldsymbol{S}$ might possess" [9].

Examples of algorithms belonging to the class of single shift-invariant methods include ESPRIT [25], [26], [27], [28],[29], HTLS [10], and the "toeplitz approximation method" (TAM) [30], [31]. The Unitary ESPRIT algorithm also belongs to this group of parameter estimation methods, and we will treat this in detail in the following chapters.

Orhtogonal vector methods are closely related to single shift-invariant methods, but are described in terms of finding vectors which are orthogonal to a vector from the noise subspace [9]. This is possible, because the noise subspace is orthogonal to the signal subspace (see e.g. Appendix C).

Examples of algorithms belonging to the class of orthogonal vector methods is Kumaresan-Tufts minumum norm method [32], Pisarenkos harmonic decomposition [9] and "multiple signal classification" (MUSIC) [33], [17].

Estimation of amplitudes and phases

The above mentioned subspace based methods all estimate the frequencies contained in the signal segment. When this has been done, the amplitude and phase of each sinusoid can be estimated using a least squares technique. A Vandermonde matrix $\boldsymbol{A}$ can be constructed according to (2.5). Then, the complex amplitudes found by the solution to the following minimization problem will yield the best estimates in the least squares sense.

$$\hat{\boldsymbol{s}} = \arg\min_{\boldsymbol{s}} ||\boldsymbol{W}\left(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s}\right)||_2^2, \tag{2.12}$$

where $\boldsymbol{W} = diag\big(w(k)\big)$ is a diagonal matrix of the window $w(k)$ defining the analyzed signal segment. This minimization problem has the following closed form solution [17]

$$\hat{\boldsymbol{s}} = (\boldsymbol{W}\boldsymbol{A})^{\dagger}\boldsymbol{W}\boldsymbol{x} \tag{2.13}$$

where $(\boldsymbol{W}\boldsymbol{A})^{\dagger}$ is the pseudoinverse of $(\boldsymbol{W}\boldsymbol{A})$.

2.5 Signal reconstruction

There exists a number of ways to reconstruct a signal which has been modeled by a sinusoidal model, all of which consists of either some form of interpolation of the parameters between the individual signal segments or a composition of reconstructed signal segments.

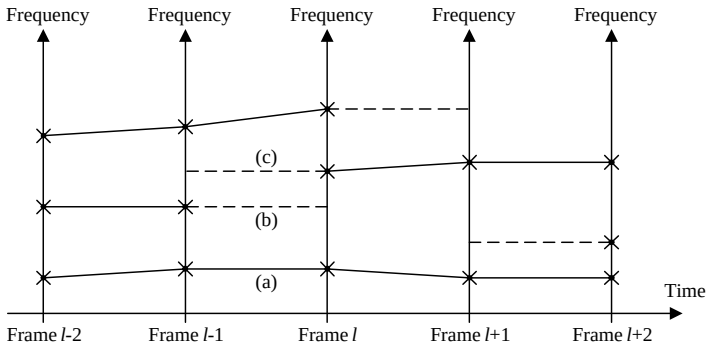

Figure 2.2: *In this figure, five frames from the sinusoidal coder are depicted. From frame $l-2$ the three frequencies, marked by "x", are interpolated to match the three frequencies in frame $l-1$ this is shown with a solid line as in (a). Before frame l, one of the frequency tracks does not match with a continuing frequency and therefore it "dies" (b). In the same frame, a frequency is also "born" because there were no previous match (c).*

Line tracking is a method which can be seen as having a bank of oscillators which are controlled by the signal parameters. To avoid discontinuities at the segment borders, the parameters are interpolated between frames for each oscillator. This involves the tracking of parameters between frames as shown on figure 2.2 — hence the name — line tracking [4]. The amplitudes are interpolated linearly from frame to frame and when a line is "born" or "dies" the amplitude is ramped up from or down to zero. The frequencies and phases between two frames have more degrees of freedom and therefore these parameters are interpolated by a cubic function. A heuristic method is used to determine which frequency lines to connect between frames [4]. One of the advantages of line tracking is that the parameters can be updated at a low rate, assuming that the signal satisfies the signal model well. However, the matching of the individual frequency lines can be difficult and erroneous tracking will often result in audible artifacts [4].

Overlap and add (OLA) is a method which recreates the signal on a block by block basis. The signal is segmented such that consecutive segments have a certain overlap. Then, a whole signal segment is recreated with fixed parameters and the segment is multiplied with a synthesis window. The windowed segments are then added together to one continuous signal [18]. It is a requirement that the multiplication by the synthesis windows does not affect the signal, and thus the overlapping windows must add up to unity

$$\sum_i w(k - ip) = 1, \tag{2.14}$$

where p denotes the frame stride. The synthesis window, $w(n)$, is often chosen to be triangular or Hann window, which for a segment overlap of 50% agree with (2.14) (see figure 2.3). One of the advantages of OLA is its simplicity; however, the parameters must be updated more often than when using e.g. line tracking.

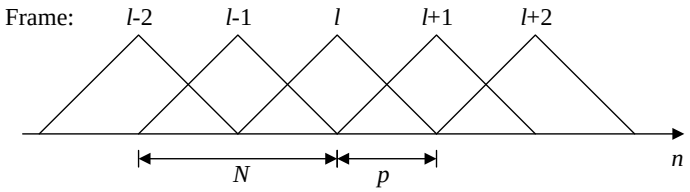

Figure 2.3: *The overlap and add method: Each reconstructed frame is multiplied by e.g. a triangular window of width N and added to the previous part of the signal. For each stride, p, a new frame is reconstructed.*

2.6 Summary

In this chapter we have set up the framework for the sinusoidal modeling of audio signals. We have described how an audio signal is segmented, how the signal parameters can be estimated, and how the signal can be reconstructed from those parameters. A signal vector and signal matrix model was introduced which will be used as a foundation for the Unitary ESPRIT parameter estimation methods presented in the sequel.

Chapter 3

Unitary ESPRIT

> "*A common mistake that people make when trying to design something completely foolproof is to underestimate the ingenuity of complete fools.*"
>
> Douglas Adams (1952 – 2001)

In this chapter: *We extensively describe the Unitary ESPRIT algorithm — a subspace based algorithm for estimating the frequencies in a sinusoidal model. We start by describing the ESPRIT algorithm which is the basis for the Unitary ESPRIT algorithm. Then we move on to describe the details of the Unitary ESPRIT algorithm which provides an increased estimation accuracy while having equal or lower computational complexity compared to the ESPRIT algorithm.*

3.1 ESPRIT

Estimation of signal parameters via rotational invariance techniques (ESPRIT) is a subspace based method for estimating parameters in a sinusoidal model. ESPRIT was developed by Roy and Paulray for estimating the direction of arrival (DOA) of a narrow band planar wavefront impinging on an array of sensors (cf. [25], [28], [26], [29]). However, that problem is similar to that of estimating parameters of complex sinusoids (cisoids) in noise [27].

In the following, the ESPRIT algorithm is described. The approach to the algorithm taken here to some extent follows that of Roy et al. [25]. Another common approach, known as the covariance method, is briefly outlined in appendix A. Although this dissertation is solely concerned with time series analysis, we will retain some of the DOA terminology since it is common for most of the litterature on ESPRIT and since the parallels between array signal processing and time series analysis can provide valueable insight.

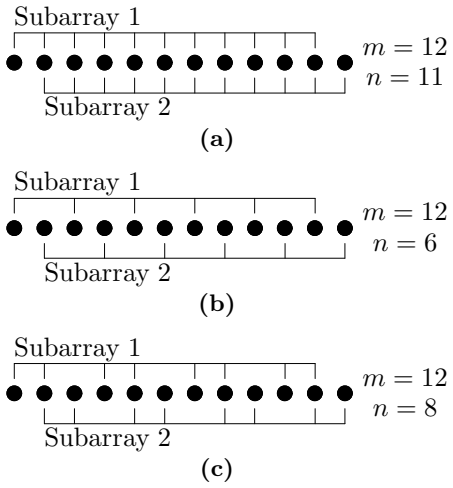

Figure 3.1: *Examples of possible choices of subarrays in ESPRIT for a uniformly sampled time series. The dots indicate the individual time samples and the lines show which samples are assigned to each subarray. The combined number of samples is denoted m and the number of samples in each subarray is denoted n.* ***(a)*** *Maximum overlap* ***(b)*** *Interleaved* ***(c)*** *Mixed.*

3.1.1 Signal model

The signal model used in the ESPRIT algorithm consists of a sum of cisoids plus noise, as described in section 2.2. The signal model for signal vector of length m is given by (2.4) repeated here

$$\boldsymbol{x}(k) = \boldsymbol{A}\boldsymbol{s}(k) + \boldsymbol{n}(k). \tag{2.4}$$

Written in terms of a signal matrix, the signal model is given by (2.7) repeated here

$$\boldsymbol{X}(k) = \boldsymbol{A}\boldsymbol{S}(k) + \boldsymbol{N}(k). \tag{2.7}$$

3.1.2 Subarrays

In ESPRIT, the parameters of the signal model are estimated based on measurements of a signal. The measurements are divided into two identical and possibly overlapping so-called subarrays[1] displaced by a fixed number of samples. Examples of how these subarrays can be chosen is shown in figure 3.1. Such two identical displaced subarrays are said to be translationally invariant when the signal is stationary. This induces a rotational invariance[2] of the underlying subspaces of the signals sampled at the two subarrays. Roy states: "The basic idea behind ESPRIT is to exploit the rotational invariance of the underlying signal subspaces induced by the translational invariance of the sensor array." [26] In the following, we show how this is used to estimate the signal parameters.

[1] This terminology originates from DOA estimation. In the context of DOA, a sensor array is employed which comprises a number of physical sensors such as microphones or radio antennas. In the context of time series analysis, the sensor array corresponds to the individual time samples. When the time series is uniformly sampled, it corresponds to a uniform linear array (ULA).

[2] When stating that two vector spaces are rotationally invariant we indicate that the basis vectors spanning the subspaces are rotated with respect to each other in the complex plane.

Let the number of samples assigned to each subarray be denoted n and the combined number of samples in the two subarrays be denoted m. Now we introduce a selection matrix

$$\boldsymbol{J} = \begin{bmatrix} \boldsymbol{J}_1 \\ \boldsymbol{J}_2 \end{bmatrix} \in \mathbb{R}^{2n \times m},$$

where $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$ are matrices which assign the desired array elements to each subarray. We denote the signal vectors at the two subarrays $\boldsymbol{x}_1(k), \boldsymbol{x}_2(k) \in \mathbb{R}^n$. Considering the vector, $\boldsymbol{J}\boldsymbol{x}(k) \in \mathbb{R}^{2n}$, which stacks these two vectors, we may write

$$\boldsymbol{J}\boldsymbol{x}(k) = \begin{bmatrix} \boldsymbol{J}_1 \\ \boldsymbol{J}_2 \end{bmatrix} \boldsymbol{x}(k) = \begin{bmatrix} \boldsymbol{x}_1(k) \\ \boldsymbol{x}_2(k) \end{bmatrix}. \tag{3.1}$$

As an example of such a selection matrix, consider the maximum overlap subarrays shown in figure 3.1.a. For this selection of subarrays we have

$$\boldsymbol{J}_1 = [\boldsymbol{I}_n | \boldsymbol{0}_n], \quad \boldsymbol{J}_2 = [\boldsymbol{0}_n | \boldsymbol{I}_n],$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and $\boldsymbol{0}_n = [0, \ldots, 0]^T \in \mathbb{R}^n$ is a vector of zeros. Choosing the subarrays to have maximum overlap will enable us to exploit all of the single shift translational invariance and this choice of subarrays will consequently provide the greatest estimation accuracy. However, choosing other subarrays will result in signal matrices for each subarray of less dimensions. Therefore, the choice of subarray configurations is a tradeoff between computational complexity and estimation accuracy.

Dividing the samples in the signal matrix into the two subarrays similar to (3.1) we may write

$$\boldsymbol{J}\boldsymbol{X}(k) = \begin{bmatrix} \boldsymbol{J}_1 \\ \boldsymbol{J}_2 \end{bmatrix} \boldsymbol{X}(k) = \begin{bmatrix} \boldsymbol{X}_1(k) \\ \boldsymbol{X}_2(k) \end{bmatrix}, \tag{3.2}$$

or written in terms of the signal model

$$\boldsymbol{J}\boldsymbol{X}(k) = \begin{bmatrix} \boldsymbol{J}_1\boldsymbol{A} \\ \boldsymbol{J}_2\boldsymbol{A} \end{bmatrix} \boldsymbol{S}(k) + \begin{bmatrix} \boldsymbol{N}_1(k) \\ \boldsymbol{N}_2(k) \end{bmatrix}. \tag{3.3}$$

3.1.3 Subspace invariance

Recall that the two subarrays are identical and displaced by a fixed number of samples. For simplicity of notation, in the sequel we assume the subarrays are displaced by one sample which can easily be extended to the general case. We may thus write

$$\boldsymbol{J}_1\boldsymbol{A}\boldsymbol{\Phi} = \boldsymbol{J}_2\boldsymbol{A}, \tag{3.4}$$

where $\boldsymbol{\Phi} = diag(e^{j\omega_1}, \ldots, e^{j\omega_d})$ is a diagonal matrix of the signal poles. We now make the following definition

$$\tilde{\boldsymbol{A}} = \boldsymbol{J}_1\boldsymbol{A}. \tag{3.5}$$

If we insert (3.4) and (3.5) in (3.3) we get

$$\boldsymbol{J}\boldsymbol{X}(k) = \begin{bmatrix} \tilde{\boldsymbol{A}} \\ \tilde{\boldsymbol{A}}\boldsymbol{\Phi} \end{bmatrix} \boldsymbol{S}(k) + \begin{bmatrix} \boldsymbol{N}_1(k) \\ \boldsymbol{N}_2(k) \end{bmatrix}. \tag{3.6}$$

By making the following definitions

$$\bar{\boldsymbol{A}} = \begin{bmatrix} \tilde{\boldsymbol{A}} \\ \tilde{\boldsymbol{A}}\boldsymbol{\Phi} \end{bmatrix} \in \mathbb{C}^{2n\times d}, \tag{3.7}$$

$$\tag{3.8}$$

$$\bar{\boldsymbol{N}}(k) = \begin{bmatrix} \boldsymbol{N}_1(k) \\ \boldsymbol{N}_2(k) \end{bmatrix} \in \mathbb{R}^{2n\times M}, \tag{3.9}$$

we may write (3.6) compactly as

$$\boldsymbol{J}\boldsymbol{X}(k) = \bar{\boldsymbol{A}}\boldsymbol{S}(k) + \bar{\boldsymbol{N}}(k). \tag{3.10}$$

This is the signal model for the two stacked subarray signal matrices. This model consists of two parts: the signal, $\bar{\boldsymbol{A}}\boldsymbol{S}(k)$, and the noise, $\bar{\boldsymbol{N}}(k)$. The vector space spanned by the signal matrix, $\boldsymbol{J}\boldsymbol{X}(k)$, is also said to consist of two parts: the signal subspace and the noise subspace. (For an introduction to the notion of signal and noise subspaces, see appendix C). Since the signal consists of linear combinations of the columns of $\bar{\boldsymbol{A}}$, the signal subspace is spanned by the columns of $\bar{\boldsymbol{A}}$, i.e. the signal subspace is given by the range of $\bar{\boldsymbol{A}}$: $\mathcal{R}(\bar{\boldsymbol{A}})$. Because $\bar{\boldsymbol{A}}$ has rank d, the signal subspace is d-dimensional. The noise subspace is the orthogonal complement of the signal subspace, i.e. it corresponds to the null space of $\bar{\boldsymbol{A}}^H$: $\mathcal{N}(\bar{\boldsymbol{A}}^H)$

Now, we introduce a matrix $\bar{\boldsymbol{E}} \in \mathbb{C}^{2n\times d}$ which spans the signal subspace of $\boldsymbol{J}\boldsymbol{X}(k)$, i.e. $\mathcal{R}\{\boldsymbol{E}\} = \mathcal{R}\{\bar{\boldsymbol{A}}\}$. Since $\bar{\boldsymbol{E}}$ and $\bar{\boldsymbol{A}}$ have the same column space, there exists a unique non-singular matrix, $\boldsymbol{T} \in \mathbb{C}^{d\times d}$, such that [17]

$$\bar{\boldsymbol{E}} = \bar{\boldsymbol{A}}\boldsymbol{T}. \tag{3.11}$$

Because of the structure consisting of the two subarrays, $\bar{\boldsymbol{E}}$ can be decomposed into two matrices, $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$, which span the signal subspaces of the two subarrays. Using (3.7) in (3.11) we may write

$$\bar{\boldsymbol{E}} = \begin{bmatrix} \boldsymbol{E}_1 \\ \boldsymbol{E}_2 \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{A}}\boldsymbol{T} \\ \tilde{\boldsymbol{A}}\boldsymbol{\Phi}\boldsymbol{T} \end{bmatrix}. \tag{3.12}$$

From this we see that $\mathcal{R}\{\boldsymbol{E}_1\} = \mathcal{R}\{\boldsymbol{E}_2\} = \mathcal{R}\{\tilde{\boldsymbol{A}}\}$. Based on the same argument as above, there exists a unique non-singular matrix $\boldsymbol{\Psi} \in \mathbb{C}^{d\times d}$ such that

$$\boldsymbol{E}_1\boldsymbol{\Psi} = \boldsymbol{E}_2. \tag{3.13}$$

Inserting (3.12) in (3.13) yields

$$\tilde{\boldsymbol{A}}\boldsymbol{T}\boldsymbol{\Psi} = \tilde{\boldsymbol{A}}\boldsymbol{\Phi}\boldsymbol{T} \quad \Rightarrow \quad \boldsymbol{\Phi} = \boldsymbol{T}\boldsymbol{\Psi}\boldsymbol{T}^{-1}. \tag{3.14}$$

Since $\boldsymbol{\Phi}$ is defined to be diagonal, this is recognized as the eigenvalue decomposition (EVD) of $\boldsymbol{\Psi}$. The signal poles are the diagonal elements of $\boldsymbol{\Phi}$ and they can therefore be computed by this EVD

$$z_i = \lambda_i(\boldsymbol{\Psi}), \tag{3.15}$$

where $\lambda_i(\boldsymbol{\Psi})$ denotes the ith eigenvalue of $\boldsymbol{\Psi}$. Note however, that since in general $\boldsymbol{\Psi}$ can be an arbitrary matrix, there is no constraints imposed on its eigenvalues. Consequently, the eigenvalues may lie anywhere in the complex plane. However, if the signal is consistent with the signal model, the signal poles will lie on or very close to the unit circle. For analyses on the estimation accuracy of the ESPRIT algorithm, see e.g. [34].

3.1.4 Subspace estimation

In practical situations, matrices spanning the signal subspaces of the two subarrays are not available and must be estimated from the signal matrix. The singular value decomposition (SVD) provides a robust and numerically stable means for estimating the signal subspaces [9]. For an introduction of the SVD and its properties, see appendix B.

If the SVD of $\boldsymbol{JX}(k)$ is given by

$$\boldsymbol{JX}(k) = \boldsymbol{U_{JX}}\boldsymbol{\Sigma_{JX}}\boldsymbol{V_{JX}}^H, \tag{3.16}$$

the first d columns of $\boldsymbol{U_{JX}}$ constitute an estimate of the signal subspace. We may write

$$\hat{\bar{\boldsymbol{E}}}_{\boldsymbol{JX}} = \begin{bmatrix}\hat{\boldsymbol{E}}_{\boldsymbol{JX}1} \\ \hat{\boldsymbol{E}}_{\boldsymbol{JX}2}\end{bmatrix} = [\boldsymbol{u_{JX}}_{:,1}, \ldots, \boldsymbol{u_{JX}}_{:,d}], \tag{3.17}$$

where $\boldsymbol{u_X}_{:,i}$ is the ith column of $\boldsymbol{U_X}$. This requires taking the SVD of the $2n \times M$ stacked signal matrix.

Another approach is based on taking the SVD of the signal matrix $\boldsymbol{X}(k)$ before partitioning it into the two subarrays. If the SVD of $\boldsymbol{X}(k)$ is given by

$$\boldsymbol{X}(k) = \boldsymbol{U_X}\boldsymbol{\Sigma_X}\boldsymbol{V_X}^H, \tag{3.18}$$

and we define the first d columns of the left singular vectors as

$$\hat{\boldsymbol{E}}_{\boldsymbol{X}} = [\boldsymbol{u_X}_{:,1}, \ldots, \boldsymbol{u_X}_{:,d}], \tag{3.19}$$

then the signal subspace is spanned by

$$\hat{\bar{\boldsymbol{E}}}_{\boldsymbol{X}} = \begin{bmatrix}\hat{\boldsymbol{E}}_{\boldsymbol{X}1} \\ \hat{\boldsymbol{E}}_{\boldsymbol{X}2}\end{bmatrix} = \boldsymbol{J}\hat{\boldsymbol{E}}_{\boldsymbol{X}}. \tag{3.20}$$

The computation of $\hat{\bar{\boldsymbol{E}}}_{\boldsymbol{X}}$ requires taking the SVD of an $m \times M$ matrix, which involves significantly fewer computations than the aforementioned method. This is prudent if the subarrays have a large overlap, such that $m << 2n$. Since this is often the case, this latter formulation is usually preferred.

3.1.5 Least squares and total least squares

When the analyzed signal fits the signal model perfectly, the signal subspaces of the two subarrays are equal. For real life signals, however, noise, non-stationarity, or insufficient model order can cause model mismatch and thus the subarray subspace estimates do not exactly span the same subspace. In other words, for real life signals most likely $\mathcal{R}(\hat{\boldsymbol{E}}_1) \neq \mathcal{R}(\hat{\boldsymbol{E}}_1)$. Thus, (3.13) is not directly solvable and must be solved approximately

$$\hat{\boldsymbol{E}}_1\boldsymbol{\Psi} \approx \hat{\boldsymbol{E}}_2. \tag{3.21}$$

A least squares estimate can be obtained by the following expression

$$\hat{\boldsymbol{\Psi}}_{LS_1} = \hat{\boldsymbol{E}}_1^{\dagger}\hat{\boldsymbol{E}}_2. \tag{3.22}$$

where $\hat{\boldsymbol{E}}_1^{\dagger} = (\hat{\boldsymbol{E}}_1^H \hat{\boldsymbol{E}}_1)^{-1} \hat{\boldsymbol{E}}_1^H$ is the pseudoinverse of $\hat{\boldsymbol{E}}_1$. This solution corresponds to projecting the space spanned by $\hat{\boldsymbol{E}}_2$ onto the column space of $\hat{\boldsymbol{E}}_1$. This can be considered as introducing the least possible pertubation of the space spanned by $\hat{\boldsymbol{E}}_2$ that renders the equation solvable. Similarly, a least squares solution can be found by projecting the space spanned by $\hat{\boldsymbol{E}}_1$ onto the column space of $\hat{\boldsymbol{E}}_2$,

$$\hat{\boldsymbol{\Psi}}_{LS_2}^{-1} = \hat{\boldsymbol{E}}_2^{\dagger} \hat{\boldsymbol{E}}_1. \tag{3.23}$$

However, since the inaccuracy can reasonably be attributed to both subspace estimates, a total least squares (TLS) estimate of $\boldsymbol{\Psi}$ will be more appropriate [25] — although, if the signal block is large, the difference is only marginal [9]. The TLS solution corresponds to projecting both $\hat{\boldsymbol{E}}_1$ and $\hat{\boldsymbol{E}}_2$ onto a subspace that "lies between" $\hat{\boldsymbol{E}}_1$ and $\hat{\boldsymbol{E}}_2$. Finding the TLS solution to an equation of the form of (3.21) is known as the multi dimensional TLS problem. It can be formulated as follows [23]: Given $\hat{\boldsymbol{E}}_1, \hat{\boldsymbol{E}}_2$, find $\hat{\boldsymbol{\Psi}}_{TLS}$ as well as $\Delta\hat{\boldsymbol{E}}_1$ and $\Delta\hat{\boldsymbol{E}}_2$ of minimum Frobenius norm[3] such that

$$\left(\hat{\boldsymbol{E}}_1 + \Delta\hat{\boldsymbol{E}}_1\right) \hat{\boldsymbol{\Psi}}_{TLS} = \hat{\boldsymbol{E}}_2 + \Delta\hat{\boldsymbol{E}}_2. \tag{3.24}$$

We start by bringing (3.24) onto the following form [35]

$$[\hat{\boldsymbol{E}}_1 + \Delta\hat{\boldsymbol{E}}_1 \quad \hat{\boldsymbol{E}}_2 + \Delta\hat{\boldsymbol{E}}_2] \begin{bmatrix} \hat{\boldsymbol{\Psi}}_{TLS} \\ -\boldsymbol{I} \end{bmatrix} = 0. \tag{3.25}$$

We proceed by selecting $\Delta\hat{\boldsymbol{E}}_1$ and $\Delta\hat{\boldsymbol{E}}_2$ of minimum Frobenius norm that reduce the rank of $[\hat{\boldsymbol{E}}_1 + \Delta\hat{\boldsymbol{E}}_1 \quad \hat{\boldsymbol{E}}_2 + \Delta\hat{\boldsymbol{E}}_2]$ to d such that a solution exists [35]. In other words, we must determine the best rank d approximant of $[\hat{\boldsymbol{E}}_1 \ \hat{\boldsymbol{E}}_2]$ in the Frobenius norm. This is found by making the d smallest singular values of $[\hat{\boldsymbol{E}}_1 \ \hat{\boldsymbol{E}}_2]$ zero [35]. Let the SVD of $[\hat{\boldsymbol{E}}_1 \ \hat{\boldsymbol{E}}_2]$ be given by

$$[\hat{\boldsymbol{E}}_1 \ \hat{\boldsymbol{E}}_2] = \boldsymbol{U}_E \boldsymbol{\Sigma}_E \boldsymbol{V}_E^H.$$

The SVD can be partitioned as [16]

$$\boldsymbol{U}_{\boldsymbol{E}} = \begin{bmatrix} \boldsymbol{U}_{\boldsymbol{E}1} & \boldsymbol{U}_{\boldsymbol{E}2} & \boldsymbol{U}_{\boldsymbol{E}3} \\ d & d & n-2d \end{bmatrix} n \tag{3.26}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{E}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{E}1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{\boldsymbol{E}2} \\ \boldsymbol{0} & \boldsymbol{0} \\ d & d \end{bmatrix} \begin{matrix} d \\ d \\ n-2d \end{matrix} \tag{3.27}$$

$$\boldsymbol{V}_{\boldsymbol{E}} = \begin{bmatrix} \boldsymbol{V}_{\boldsymbol{E}11} & \boldsymbol{V}_{\boldsymbol{E}12} \\ \boldsymbol{V}_{\boldsymbol{E}21} & \boldsymbol{V}_{\boldsymbol{E}22} \\ d & d \end{bmatrix} \begin{matrix} d \\ d \end{matrix}. \tag{3.28}$$

The best rank d approximant of $[\hat{\boldsymbol{E}}_1 \ \hat{\boldsymbol{E}}_2]$ is given by [35]

$$[\hat{\boldsymbol{E}}_1 + \Delta\hat{\boldsymbol{E}}_1 \quad \hat{\boldsymbol{E}}_2 + \Delta\hat{\boldsymbol{E}}_2] = \boldsymbol{U}_{\boldsymbol{E}1} \boldsymbol{\Sigma}_{\boldsymbol{E}1} \begin{bmatrix} \boldsymbol{V}_{\boldsymbol{E}11} \\ \boldsymbol{V}_{\boldsymbol{E}21} \end{bmatrix}^H. \tag{3.29}$$

[3] The Frobenius norm of a matrix is defined as the square root of the sum of the absolute square value of all matrix elements or alternatively as the trace of the matrix multiplied by its hermitian transposed: $||\boldsymbol{A}||_F^2 = \sum_{m,n} |a_{m,n}|^2 = tr(\boldsymbol{A}\boldsymbol{A}^H)$ [17]

Inserting (3.29) in (3.25) we have

$$\boldsymbol{U}_{\boldsymbol{E}1}\boldsymbol{\Sigma}_{\boldsymbol{E}1}\begin{bmatrix}\boldsymbol{V}_{\boldsymbol{E}11}\\ \boldsymbol{V}_{\boldsymbol{E}21}\end{bmatrix}^H\begin{bmatrix}\hat{\boldsymbol{\Psi}}_{TLS}\\ -\boldsymbol{I}\end{bmatrix}=0. \tag{3.30}$$

Since $\boldsymbol{U}_{\boldsymbol{E}1}\boldsymbol{\Sigma}_{\boldsymbol{E}1}$ is rank d, the solution to this equation is the null space of $\begin{bmatrix}\boldsymbol{V}_{\boldsymbol{E}11}\\ \boldsymbol{V}_{\boldsymbol{E}21}\end{bmatrix}^H$ which is equal to the orthogonal complement of the range of $\begin{bmatrix}\boldsymbol{V}_{\boldsymbol{E}11}\\ \boldsymbol{V}_{\boldsymbol{E}21}\end{bmatrix}$ [36]. This vector space is spanned by the right singular vectors corresponding to the d smallest singular values [35]. Consequently we have

$$\begin{bmatrix}\hat{\boldsymbol{\Psi}}_{TLS}\\ -\boldsymbol{I}\end{bmatrix}\in\mathcal{N}\left(\begin{bmatrix}\boldsymbol{V}_{\boldsymbol{E}11}\\ \boldsymbol{V}_{\boldsymbol{E}21}\end{bmatrix}^H\right)=\mathcal{R}\left(\begin{bmatrix}\boldsymbol{V}_{\boldsymbol{E}11}\\ \boldsymbol{V}_{\boldsymbol{E}21}\end{bmatrix}\right)^{\perp}=\mathcal{R}\left(\begin{bmatrix}\boldsymbol{V}_{\boldsymbol{E}12}\\ \boldsymbol{V}_{\boldsymbol{E}22}\end{bmatrix}\right). \tag{3.31}$$

From this we may find a solution for $\hat{\boldsymbol{\Psi}}_{TLS}$

$$\begin{bmatrix}\hat{\boldsymbol{\Psi}}_{TLS}\\ -\boldsymbol{I}\end{bmatrix}=\begin{bmatrix}\boldsymbol{V}_{\boldsymbol{E}12}\\ \boldsymbol{V}_{\boldsymbol{E}22}\end{bmatrix}, \tag{3.32}$$

which yields

$$\hat{\boldsymbol{\Psi}}_{TLS}=-\boldsymbol{V}_{\boldsymbol{E}12}\boldsymbol{V}_{\boldsymbol{E}22}^{-1}. \tag{3.33}$$

For further discussion of the multidimensional TLS problem such as the existence and uniqueness of the solution, see e.g. [35, sec. 3.2] and [16, sec. 12.3].

3.1.6 Summary of the ESPRIT algorithm

The TLS based ESPRIT algorithm can be summarized in the following steps, where the dominating computations for each step is included in the rightmost column

1.	Obtain an estimate of the signal subspaces for the two subarrays, $\hat{\bar{\boldsymbol{E}}}=\begin{bmatrix}\hat{\boldsymbol{E}}_1\\ \hat{\boldsymbol{E}}_2\end{bmatrix}$.	Real $(m\times M)$ SVD
2.	Solve the overdetermined system of equations $\hat{\boldsymbol{E}}_1\hat{\boldsymbol{\Psi}}\approx\hat{\boldsymbol{E}}_2$ by means of total least squares.	Real $(m\times 2d)$ SVD
3.	Compute the signal poles by the eigenvalue decomposition $\hat{z}_i=\lambda_i(\hat{\boldsymbol{\Psi}})$	Real $(d\times d)$ EVD

3.2 Unitary ESPRIT

The idea behind Unitary ESPRIT is to perform a forward-backward averaging of the signal matrix so that the signal poles are constrained to the unit circle. Also, the forward-backward averaging results in an improved estimation accuracy [13]. In addition to this, for complex signals, the algorithm has a lower computational complexity than standard ESPRIT because the special structure of the signal matrix employed can be exploited [13]. For real signals, the computational complexity of ESPRIT and Unitary ESPRIT is the same, as we will show in the following.

3.2.1 Forward-backward signal matrix

First, let us introduce the matrix $\boldsymbol{\Pi}_p$ which is a $p \times p$ matrix with ones on the main antidiagonal and zeros elsewhere

$$\boldsymbol{\Pi}_p = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix} \in \mathbb{R}^{p \times p}. \tag{3.34}$$

When a matrix of appropriate size is right multiplied by $\boldsymbol{\Pi}_p$ it corresponds to reversing the order of its columns. Left multiplication by $\boldsymbol{\Pi}_p$ corresponds to reversing the order of the rows.

Consider a signal matrix consisting of the Hankel structured signal matrix used in the ESPRIT algorithm augmented by the same signal matrix conjugated and with the order of the rows reversed.

$$\boldsymbol{Z}(k) = \left[\boldsymbol{X}(k) \quad \boldsymbol{\Pi}_m \boldsymbol{X}^*(k)\right]. \tag{3.35}$$

We denote this the forward-backward (FB) signal matrix. An intuitive justification for using this signal matrix in the ESPRIT algorithm is that the translational invariance between the two subarrays will be forced to apply equally both from the first subarray to the second and vice versa. Consequently, as we will show later, the estimates of the signal poles will be constrained to the unit circle[4] [13].

3.2.2 Constraints on selection matrices

In the Unitary ESPRIT algorithm, an added constraint on how to choose the selection matrices for the two subarrays is imposed. The two selection matrices must be centro symmetric with respect to each other [13].

Definition 3.1. *The matrices $\boldsymbol{F}, \boldsymbol{G} \in \mathbb{C}^{p \times q}$ are said to be centro symmetric with respect to each other if the following is true for each of their components*

$$f_{i,j} = g_{p+1-i,q+1-j} \qquad (1 \leq i \leq p, \quad 1 \leq j \leq q).$$

Equally we may write

$$\boldsymbol{\Pi}_p \boldsymbol{F} \boldsymbol{\Pi}_q = \boldsymbol{G}.$$

□

Thus, for the two selection matrices, $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$ we may write

$$\boldsymbol{\Pi}_n \boldsymbol{J}_1 \boldsymbol{\Pi}_m = \boldsymbol{J}_2, \tag{3.36}$$

or equivalently

$$\boldsymbol{\Pi}_n \boldsymbol{J}_1 = \boldsymbol{J}_2 \boldsymbol{\Pi}_m, \qquad \boldsymbol{\Pi}_n \boldsymbol{J}_2 = \boldsymbol{J}_1 \boldsymbol{\Pi}_m. \tag{3.37}$$

[4] Haardt et al. [13] has shown that the Unitary ESPRIT algorithm produces consistent estimates of the signal poles so that asymptotically all the estimated signal poles will be on the unit circle. "If, however, the number of snapshots N is too small or if there is only noise present, the eigenvalues of Ψ_{TLS} might fail to satisfy $|\phi_k| = 1 \; \forall k$ ..." [13] In that case, however, the eigenvalues will by symmetric with respect to the unit circle [13].

Because of the special structure of the matrix $\boldsymbol{A}$ from (2.5), given that the signal complies with the sinusoidal model, there must exist a unitary diagonal matrix, $\boldsymbol{\Lambda}$, such that [13]

$$\boldsymbol{\Pi}_m \boldsymbol{A}^* = \boldsymbol{A}\boldsymbol{\Lambda}. \tag{3.38}$$

To see that this is true, consider the following: Defining $\boldsymbol{\Lambda} = diag(\lambda_1, \dots \lambda_d)$, another way of stating (3.38) is that for each column of $\boldsymbol{A}$, $\boldsymbol{a}_{:,i}$, there exists a unique scalar of unit magnitude, λ_i, such that

$$\boldsymbol{\Pi}_m \boldsymbol{a}_{:,i}^* = \boldsymbol{a}_{:,i}\lambda_i. \tag{3.39}$$

Recall from (2.5) that each column of $\boldsymbol{A}$ can be written as

$$\boldsymbol{a}_{:,i} = [1 \; e^{jw\omega_i} \; e^{j2\omega_i} \; \cdots \; e^{j(m-1)\omega_i}]^T.$$

From this it can be seen that (3.39) holds with $\lambda_i = e^{j(m-1)\omega_i}$.

As a consequence of (3.38), the following relation holds

$$\boldsymbol{\Pi}_{2n} \bar{\boldsymbol{A}}^* = \bar{\boldsymbol{A}}\boldsymbol{\Lambda}. \tag{3.40}$$

This can be shown by inserting (3.37) and (3.38) in (3.41) and using the definition $\bar{\boldsymbol{A}} = \left[\begin{smallmatrix} \boldsymbol{J}_1 \\ \boldsymbol{J}_2 \end{smallmatrix}\right] \boldsymbol{A}$

$$\boldsymbol{\Pi}_{2n} \bar{\boldsymbol{A}}^* = \begin{bmatrix} \boldsymbol{\Pi}_n \boldsymbol{J}_2 \boldsymbol{A}^* \\ \boldsymbol{\Pi}_n \boldsymbol{J}_1 \boldsymbol{A}^* \end{bmatrix} = \begin{bmatrix} \boldsymbol{J}_1 \boldsymbol{\Pi}_m \boldsymbol{A}^* \\ \boldsymbol{J}_2 \boldsymbol{\Pi}_m \boldsymbol{A}^* \end{bmatrix} = \begin{bmatrix} \boldsymbol{J}_1 \boldsymbol{A}\boldsymbol{\Lambda} \\ \boldsymbol{J}_2 \boldsymbol{A}\boldsymbol{\Lambda} \end{bmatrix} = \bar{\boldsymbol{A}}\boldsymbol{\Lambda}.$$

The properties derived in this section are used in the derivation of some of the properties of the Unitary ESPRIT algorithm in the following.

3.2.3 FB signal matrix in the ESPRIT algorithm

Now, let us return to the FB signal matrix. If we split the signal matrix into the two subarrays and insert the signal model from equation 3.6 we get

$$\boldsymbol{J}\boldsymbol{Z}(k) = \left[\bar{\boldsymbol{A}}\boldsymbol{S}(k) \quad \boldsymbol{\Pi}_{2m}\bar{\boldsymbol{A}}^*\boldsymbol{S}^*(k)\right] + \left[\bar{\boldsymbol{N}} \quad \boldsymbol{\Pi}_{2m}\bar{\boldsymbol{N}}^*\right]. \tag{3.41}$$

Now, using the centrosymmetry of the sensor array from (3.41) we may write

$$\boldsymbol{J}\boldsymbol{Z}(k) = \bar{\boldsymbol{A}}\left[\boldsymbol{S}(\boldsymbol{k}) \quad \boldsymbol{\Lambda}\boldsymbol{S}^*(k)\right] + \left[\bar{\boldsymbol{N}} \quad \boldsymbol{\Pi}_{2m}\bar{\boldsymbol{N}}^*\right]. \tag{3.42}$$

From this, it is apparant that the signal subspace of $\boldsymbol{J}\boldsymbol{Z}(k)$ is spanned by $\bar{\boldsymbol{A}}$. Thus, the standard ESPRIT algorithm applies using this signal matrix with the only difference that the translational invariance of the subarrays are used both forwards and backwards which increases the estimation accuracy and restricts the signal poles to the unit circle [13].

3.2.4 Signal poles constrained to the unit circle

We now proceed to discuss how the signal poles are constrained to the unit circle in the Unitary ESPRIT algorithm. If we insert (3.7) in (3.41) we have

$$\boldsymbol{\Pi}_{2m}\begin{bmatrix}\tilde{\boldsymbol{A}}^* \\ \tilde{\boldsymbol{A}}^*\boldsymbol{\Phi}^*\end{bmatrix} = \begin{bmatrix}\tilde{\boldsymbol{A}} \\ \tilde{\boldsymbol{A}}\boldsymbol{\Phi}\end{bmatrix}\boldsymbol{\Lambda}. \tag{3.43}$$

By rearranging this we get the following matrix equation pair

$$\begin{bmatrix}\boldsymbol{\Pi}_m\tilde{\boldsymbol{A}}^*\boldsymbol{\Phi}^* \\ \boldsymbol{\Pi}_m\tilde{\boldsymbol{A}}^*\end{bmatrix} = \begin{bmatrix}\tilde{\boldsymbol{A}}\boldsymbol{\Lambda} \\ \tilde{\boldsymbol{A}}\boldsymbol{\Phi}\boldsymbol{\Lambda}\end{bmatrix}. \tag{3.44}$$

Inserting the bottom row equation in the top row equation and rearranging terms (note that $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$ are diagonal matrices) we have

$$\tilde{\boldsymbol{A}}\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^* = \tilde{\boldsymbol{A}}\boldsymbol{\Lambda}. \tag{3.45}$$

From this we see that $\boldsymbol{\Phi}\boldsymbol{\Phi}^* = \boldsymbol{I}$, i.e. $\boldsymbol{\Phi}$ is unitary. Since $\boldsymbol{\Phi}$ is a diagonal matrix of the signal poles, it is obvious that the signal poles are constrained to the unit circle.

3.2.5 Subspace estimation using the FB signal matrix

The first step in the Unitary ESPRIT algorithm consists of estimating the signal subspaces for the two subarrays. Using the FB signal matrix in the ESPRIT algorithm, this can be done by computing the singular value decomposition of $\boldsymbol{J}\boldsymbol{Z}(k)$. The computational complexity of this can be significantly reduced by exploiting the structure of the FB signal matrix [13]. To show this, we start by making the following definition.

Definition 3.2. *A matrix, $\boldsymbol{M} \in \mathbb{C}^{p\times q}$, is said to be centro hermitian if the following is true for each of its components [37]:*

$$m_{i,j} = m^*_{p+1-i,q+1-j} \qquad (1 \le i \le p, \quad 1 \le j \le q).$$

Equally we may write

$$\boldsymbol{\Pi}_p\boldsymbol{M}^*\boldsymbol{\Pi}_q = \boldsymbol{M}.$$

□

Thus, a matrix is centro hermitian if it is equal to its complex conjugate with the rows and columns in reverse order. The FB signal matrix, $\boldsymbol{Z}(k)$, is not itself centro hermitian. However, by reversing the order of the last M colums of $\boldsymbol{Z}(k)$ we obtain a centro hermitian matrix, $\boldsymbol{Z}_C(k)$

$$\boldsymbol{Z}_C(k) = \boldsymbol{Z}(k)\begin{bmatrix}\boldsymbol{I}_M & \\ & \boldsymbol{\Pi}_M\end{bmatrix} = \left[\boldsymbol{X}(k) \quad \boldsymbol{\Pi}_m\boldsymbol{X}^*(k)\boldsymbol{\Pi}_M\right]. \tag{3.46}$$

We note, that the column spaces of $\boldsymbol{Z}(k)$ and $\boldsymbol{Z}_C(k)$ are equal, $\mathcal{R}(\boldsymbol{Z}(k)) = \mathcal{R}(\boldsymbol{Z}_C(k))$, since the only difference between the matrices is a permutation of the columns. The symmetry of $\boldsymbol{Z}_C(k)$ can be exploited to decrease the complexity of the computation of the SVD of $\boldsymbol{Z}(k)$ as we will show in the following.

Isomorphism between centro hermitian and real matrices

There exists an isomorphism between the set of centro hermitian matrices and the set of real matrices of equal dimensions. To show this, we start by making the following definition.

Definition 3.3. *A matrix, $\boldsymbol{Q}$ is said to be column conjugate symmetric if it is equal to its complex conjugate with the order of the rows reversed [37]:*

$$\boldsymbol{\Pi Q}^* = \boldsymbol{Q}.$$

"Further we remark that if $\boldsymbol{Q}$ is column conjugate symmetric and $\boldsymbol{R}$ is an arbitrary real matrix (of appropriate size), then $\boldsymbol{QR}$ is column conjugate symmetric as well." [37]

□

Consider the following mapping where $\boldsymbol{T}_p \in \mathbb{C}^{p\times p}$ and $\boldsymbol{U}_q \in \mathbb{C}^{q\times q}$ are invertible column conjugate symmetric matrices.

$$\varphi : \boldsymbol{M} \mapsto \boldsymbol{T}_p^{-1}\boldsymbol{M}\boldsymbol{U}_q. \tag{3.47}$$

This can be shown to be a bijective[5] mapping that maps the set of all $p \times q$ (generally complex) centro hermitian matrices onto $\mathbb{R}^{p\times q}$, the set of all real $p \times q$ matrices [37]. To prove this it is sufficient to show that $\varphi(\boldsymbol{M})$ is real when $\boldsymbol{M}$ is centro hermitian [37].

$$\varphi(\boldsymbol{M}) = \boldsymbol{T}_p^{-1}\boldsymbol{M}\boldsymbol{U}_q. \tag{3.48}$$

Using that $\boldsymbol{M}$ is centro hermitian (Definition 3.2) we get

$$\varphi(\boldsymbol{M}) = \boldsymbol{T}_p^{-1}\boldsymbol{\Pi}_p\boldsymbol{M}^*\boldsymbol{\Pi}_p\boldsymbol{U}_q. \tag{3.49}$$

Then, using that $\boldsymbol{T}_p$ and $\boldsymbol{U}_q$ are column conjugate symmetric (Definition 3.3) yields

$$\varphi(\boldsymbol{M}) = \boldsymbol{T}_p^{*-1}\boldsymbol{M}^*\boldsymbol{U}_q^*. \tag{3.50}$$

Comparing (3.52) to (3.50) we see that $\varphi(\boldsymbol{M}) = \varphi(\boldsymbol{M})^*$ and consequently $\varphi(\boldsymbol{M})$ is real. Thus, φ maps the set of all centro hermitian matrices onto the set of all real matrices and we may say that every centro hermitian matrix is isomorph to a real matrix of the same dimensions.

The relations described above hold for any two column conjugate symmetric matrices, $\boldsymbol{T}_p$ and $\boldsymbol{U}_q$. However, a mapping which can be described very compactly is obtained when $\boldsymbol{T}_p$ and $\boldsymbol{U}_q$ are chosen on the following form

$$\boldsymbol{Q}_{2n} = \frac{1}{\sqrt{2}}\begin{bmatrix} \boldsymbol{I}_n & j\boldsymbol{I}_n \\ \boldsymbol{\Pi}_n & -j\boldsymbol{\Pi}_n \end{bmatrix}, \tag{3.51}$$

or

$$\boldsymbol{Q}_{2n+1} = \frac{1}{\sqrt{2}}\begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0} & j\boldsymbol{I}_n \\ \boldsymbol{0}^T & \sqrt{2} & \boldsymbol{0}^T \\ \boldsymbol{\Pi}_n & \boldsymbol{0} & -j\boldsymbol{\Pi}_n \end{bmatrix}, \tag{3.52}$$

for even or odd p, q respectively. In addition to being column conjugate symmetric, these matrices are also unitary. In the following we denote by φ_Q the mapping (3.49) using these unitary column conjugate symmetric matrices

$$\varphi_Q : \boldsymbol{M} \mapsto \boldsymbol{Q}_p^H\boldsymbol{M}\boldsymbol{Q}_q. \tag{3.53}$$

[5] A mapping is said to be bijective if it is both injective (one-to-one) and surjective (onto).

SVD of a centro hermitian matrix

The singular value decomposition of the centro hermitian matrix $\boldsymbol{M} \in \mathbb{C}^{p\times q}$ can be written as

$$\boldsymbol{M} = \boldsymbol{U}_{\boldsymbol{M}} \boldsymbol{\Sigma}_{\boldsymbol{M}} \boldsymbol{V}_{\boldsymbol{M}}^{H}. \tag{3.54}$$

Now, consider the SVD of the real matrix $\varphi_Q(\boldsymbol{M})$

$$\varphi_Q(\boldsymbol{M}) = \boldsymbol{Q}_p^H \boldsymbol{M} \boldsymbol{Q}_q = \boldsymbol{U}_{\varphi\boldsymbol{M}} \boldsymbol{\Sigma}_{\varphi\boldsymbol{M}} \boldsymbol{V}_{\varphi\boldsymbol{M}}^{H}. \tag{3.55}$$

Isolating $\boldsymbol{M}$ in (3.57) and equating with (3.56) we get an expression for the SVD of $\boldsymbol{M}$ formulated in terms of the SVD of $\varphi_Q(\boldsymbol{M})$

$$\boldsymbol{M} = \boldsymbol{U}_{\boldsymbol{M}} \boldsymbol{\Sigma}_{\boldsymbol{M}} \boldsymbol{V}_{\boldsymbol{M}}^{H} = (\boldsymbol{Q}_p \boldsymbol{U}_{\varphi\boldsymbol{M}})\, \boldsymbol{\Sigma}_{\varphi\boldsymbol{M}} \left(\boldsymbol{V}_{\varphi\boldsymbol{M}}^{H} \boldsymbol{Q}_q^{H}\right), \tag{3.56}$$

where, in the last expression, terms have been grouped to emphasize the relation between the SVD of $\boldsymbol{M}$ and $\varphi_Q(\boldsymbol{M})$ respectively. From this we see how the SVD of $\boldsymbol{M}$ can be derived from the SVD of $\varphi_Q(\boldsymbol{M})$

$$\begin{aligned} \boldsymbol{U}_{\boldsymbol{M}} &= \boldsymbol{Q}_p \boldsymbol{U}_{\varphi\boldsymbol{M}} && (3.57)\\ \boldsymbol{\Sigma}_{\boldsymbol{M}} &= \boldsymbol{\Sigma}_{\varphi\boldsymbol{M}} && (3.58)\\ \boldsymbol{V}_{\boldsymbol{M}} &= \boldsymbol{Q}_q \boldsymbol{V}_{\varphi\boldsymbol{M}}. && (3.59) \end{aligned}$$

Thus, the SVD of a complex centro hermitian matrix can be reduced to the SVD of a real matrix.

Efficient computation of $\varphi_Q(\boldsymbol{M})$

If we assume that p is odd and q is even and introduce $r = (p-1)/2$ and $s = q/2$, $\boldsymbol{M}$ can be written in the following form

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{M}_1 & \boldsymbol{\Pi}_r \boldsymbol{M}_2^* \boldsymbol{\Pi}_s \\ \boldsymbol{m}^T & \boldsymbol{m}^T \boldsymbol{\Pi}_s \\ \boldsymbol{M}_2 & \boldsymbol{\Pi}_r \boldsymbol{M}_1^* \boldsymbol{\Pi}_s \end{bmatrix} \qquad \begin{aligned} \boldsymbol{M}_1, \boldsymbol{M}_2 &\in \mathbb{C}^{r\times s} \\ \boldsymbol{m} &\in \mathbb{C}^{s}. \end{aligned} \tag{3.60}$$

This expression is valid for odd p, however, if p is even the center row should simply be dropped and $r = p/2$. In the following we will proceed with the assumption of odd p, noting that expressions for even p can easily be deduced. Now, since $\boldsymbol{M}$ is centro hermitian, it can be mapped into a real matrix by the following mapping

$$\varphi_Q(\boldsymbol{M}) = \boldsymbol{Q}_p \boldsymbol{M} \boldsymbol{Q}_q. \tag{3.61}$$

Inserting (3.53), (3.54), and (3.62) in (3.63) yields an expression for $\varphi_Q(\boldsymbol{M})$ in closed form

$$\begin{aligned} \varphi_Q(\boldsymbol{M}) &= \frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{0} & j\boldsymbol{I}_r \\ \boldsymbol{0}^T & \sqrt{2} & \boldsymbol{0}^T \\ \boldsymbol{\Pi}_r & \boldsymbol{0} & -j\boldsymbol{\Pi}_r \end{bmatrix} \begin{bmatrix} \boldsymbol{M}_1 & \boldsymbol{\Pi}_r \boldsymbol{M}_2^* \boldsymbol{\Pi}_s \\ \boldsymbol{m}^T & \boldsymbol{m}^T \boldsymbol{\Pi}_s \\ \boldsymbol{M}_2 & \boldsymbol{\Pi}_r \boldsymbol{M}_1^* \boldsymbol{\Pi}_s \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{I}_s & j\boldsymbol{I}_s \\ \boldsymbol{\Pi}_s & -j\boldsymbol{\Pi}_s \end{bmatrix} \\ &= \begin{bmatrix} Re\left(\boldsymbol{M}_1 + \boldsymbol{\Pi}_r \boldsymbol{M}_2^*\right) & -Im\left(\boldsymbol{M}_1 - \boldsymbol{\Pi}_r \boldsymbol{M}_2^*\right) \\ \sqrt{2} \cdot Re(\boldsymbol{m}^T) & -\sqrt{2} \cdot Im(\boldsymbol{m}^T) \\ Im\left(\boldsymbol{M}_1 + \boldsymbol{\Pi}_r \boldsymbol{M}_2^*\right) & Re\left(\boldsymbol{M}_1 - \boldsymbol{\Pi}_r \boldsymbol{M}_2^*\right) \end{bmatrix}. \end{aligned} \tag{3.62}$$

By inspection we verify that this is indeed a real matrix. In addition we note that for real $\boldsymbol{M}$ this is a block diagonal[6] matrix. Thus, the computation of the SVD of $\boldsymbol{M}$ is significantly simplified using this mapping both for complex and real $\boldsymbol{M}$.

[6] A block diagonal matrix $\boldsymbol{A}$ is a block matrix where, if we denote its block elements by $\boldsymbol{A}_{i,j}$, we may write $\boldsymbol{A}_{i,j} = \boldsymbol{0}$ if $i \neq j$.

SVD of a block diagonal matrix

The SVD of a block diagonal matrix can be found from the SVD of the individual blocks. To see this, consider two matrices, $\boldsymbol{F}$ and $\boldsymbol{G}$ of which the SVD is given by

$$\begin{aligned} \boldsymbol{F} &= \boldsymbol{U}_F \boldsymbol{\Sigma}_F \boldsymbol{V}_F^H, \\ \boldsymbol{G} &= \boldsymbol{U}_G \boldsymbol{\Sigma}_G \boldsymbol{V}_G^H. \end{aligned}$$

Then, it can be verified that the SVD of the 2×2 block diagonal matrix with $\boldsymbol{F}$ and $\boldsymbol{G}$ on the diagonal is given by

$$\begin{bmatrix} \boldsymbol{F} & \\ & \boldsymbol{G} \end{bmatrix} = \begin{bmatrix} \boldsymbol{U}_F & \\ & \boldsymbol{U}_G \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_F & \\ & \boldsymbol{\Sigma}_G \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_F & \\ & \boldsymbol{V}_G \end{bmatrix}^H. \tag{3.63}$$

Using this expression, however, most likely the singular values do not occur in non-decreasing order. This must subsequently be ensured by the appropriate column permutations.

Since the computational complexity of the SVD of a matrix of size $m \times n$ is $\mathcal{O}\big(\min(m,n)^3\big)$ [16], computing the two SVDs of the block diagonal elements as opposed to computing the SVD of the full matrix will reduce the computational complexity by a factor of four.

Estimation of signal subspaces

An estimate of the signal subspaces for the two subarrays can be found from the first d left singular vectors of $\boldsymbol{Z}_C(k)$. Since this is a centro hermitian matrix, the SVD of $\boldsymbol{Z}_C(k)$ can be found from the SVD of $\varphi_Q(\boldsymbol{Z}_C(k))$. Since $\boldsymbol{Z}_C(k)$ is real, $\varphi_Q(\boldsymbol{Z}_C(k))$ is a real block diagonal matrix according to (3.64). Let the SVD of $\boldsymbol{Z}_C(k)$ be given by

$$\boldsymbol{Z}_C(k) = \boldsymbol{U}_{\boldsymbol{Z}_C} \boldsymbol{\Sigma}_{\boldsymbol{Z}_C} \boldsymbol{V}_{\boldsymbol{Z}_C}^H, \tag{3.64}$$

and the SVD of $\varphi_Q(\boldsymbol{Z}_C(k))$ be given by

$$\varphi_Q(\boldsymbol{Z}_C(k)) = \boldsymbol{U}_{\varphi\boldsymbol{Z}_C} \boldsymbol{\Sigma}_{\varphi\boldsymbol{Z}_C} \boldsymbol{V}_{\varphi\boldsymbol{Z}_C}^H, \tag{3.65}$$

If we denote the first d columns of $\boldsymbol{U}_{\varphi\boldsymbol{Z}_C}$ by

$$\hat{\boldsymbol{E}}_{\varphi\boldsymbol{Z}_C} = [\boldsymbol{u}_{\varphi\boldsymbol{Z}_C:,1}, \ldots, \boldsymbol{u}_{\varphi\boldsymbol{Z}_C:,d}], \tag{3.66}$$

then, according to (3.59)

$$\hat{\boldsymbol{E}}_{\boldsymbol{Z}_C} = \boldsymbol{Q}_m \hat{\boldsymbol{E}}_{\varphi\boldsymbol{Z}_C}. \tag{3.67}$$

Finally, by an expression similar to (3.20), $\hat{\bar{\boldsymbol{E}}}_{\boldsymbol{Z}_C}$ is given by

$$\hat{\bar{\boldsymbol{E}}}_{\boldsymbol{Z}_C} = \begin{bmatrix} \hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 1} \\ \hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 2} \end{bmatrix} = \boldsymbol{J}\boldsymbol{Q}_m \hat{\boldsymbol{E}}_{\varphi\boldsymbol{Z}_C}. \tag{3.68}$$

Summary of subspace estimation

The efficient computation of a subspace estimate for the two subarrays in Unitary ESPRIT can be summarized in the following steps:

1. Compute the real block diagonal matrix, $\varphi_Q(\boldsymbol{Z}_C(k))$ using (3.64).
2. Compute the SVD of $\varphi_Q(\boldsymbol{Z}_C(k))$ by computing the SVD of its block elements, (3.65).
3. Extract the d left singular vectors corresponding to the d largest singular values, (3.68).
4. Compute an estimate of the signal subspaces for the two subarrays using (3.70).

3.2.6 Total least squares solution using the FB signal matrix

The second step in the Unitary ESPRIT algorithm consists of solving the overdetermined system of equations $\hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 1}\hat{\boldsymbol{\Psi}} \approx \hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 2}$ by means of least squares or total least squares. When using total least squares, this requires computing the singular value decomposition of the matrix $[\hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 1} \;\; \hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 2}]$. In the following we show how this computation can be significantly simplified by exploiting the special structure of this matrix.

Inserting $\boldsymbol{J} = \left[\begin{smallmatrix}\boldsymbol{J}_1\\ \boldsymbol{J}_2\end{smallmatrix}\right]$ in (3.70) yields

$$\hat{\bar{\boldsymbol{E}}}_{\boldsymbol{Z}_C} = \begin{bmatrix}\hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 1}\\ \hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 2}\end{bmatrix} = \begin{bmatrix}\boldsymbol{J}_1\\ \boldsymbol{J}_2\end{bmatrix}\boldsymbol{Q}_m \hat{\boldsymbol{E}}_{\varphi \boldsymbol{Z}_C}. \tag{3.69}$$

Using that the subarray selection matrices are centro symmetric with respect to each other (3.36) and the column conjugate symmetry property of $\boldsymbol{Q}_m$ (Definition 3.3) and noting that $\boldsymbol{E}_{\varphi \boldsymbol{Z}_C}$ is real we see

$$\begin{aligned}\hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 2} &= \boldsymbol{J}_2 \boldsymbol{Q}_m \hat{\boldsymbol{E}}_{\varphi \boldsymbol{Z}_C}\\ &= \boldsymbol{\Pi}_n \boldsymbol{J}_1 \boldsymbol{\Pi}_m \boldsymbol{\Pi}_m \boldsymbol{Q}_m^* \hat{\boldsymbol{E}}_{\varphi \boldsymbol{Z}_C}\\ &= \boldsymbol{\Pi}_n \boldsymbol{J}_1 \boldsymbol{Q}_m^* \boldsymbol{U}_{\varphi \boldsymbol{Z}_C} \qquad = \boldsymbol{\Pi}_n \hat{\boldsymbol{E}}^*_{\boldsymbol{Z}_C 1}.\end{aligned} \tag{3.70}$$

The TLS solution in the Unitary ESPRIT algorithm requires taking the SVD of the matrix $[\hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 1} \;\; \hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 2}]$. Using (3.72) we may write

$$[\hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 1} \;\; \hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 2}] = [\hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 1} \;\; \boldsymbol{\Pi}_n \hat{\boldsymbol{E}}^*_{\boldsymbol{Z}_C 1}]. \tag{3.71}$$

We see, that this matrix has the same special structure as $\boldsymbol{Z}(k)$. Consequently, it can be transformed to a centro hermitian matrix by reversing the order of the last d columns. We denote this centro hermitian matrix $\hat{\boldsymbol{C}}_{\boldsymbol{Z}_C}$

$$\hat{\boldsymbol{C}}_{\boldsymbol{Z}_C} = [\hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 1} \;\; \boldsymbol{\Pi}_n \hat{\boldsymbol{E}}^*_{\boldsymbol{Z}_C 1}] \begin{bmatrix}\boldsymbol{I}_M & \\ & \boldsymbol{\Pi}_M\end{bmatrix} = [\hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 1} \;\; \boldsymbol{\Pi}_n \hat{\boldsymbol{E}}^*_{\boldsymbol{Z}_C 1} \boldsymbol{\Pi}_M]. \tag{3.72}$$

Thus, the computation of its SVD can be simplified in the same way as we have shown for $\boldsymbol{Z}(k)$. If the SVD of $[\hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 1} \;\; \hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 2}]$ is given by

$$\hat{\boldsymbol{C}}_{\boldsymbol{Z}_C} = \boldsymbol{U}_{\boldsymbol{C}} \boldsymbol{\Sigma}_{\boldsymbol{C}} \boldsymbol{V}_{\boldsymbol{C}}^H, \tag{3.73}$$

and the SVD of $\varphi_Q(\hat{\boldsymbol{C}}_{\boldsymbol{Z}_C})$ is given by

$$\varphi_Q(\hat{\boldsymbol{C}}_{\boldsymbol{Z}_C}) = \boldsymbol{U}_{\varphi \boldsymbol{C}} \boldsymbol{\Sigma}_{\varphi \boldsymbol{C}} \boldsymbol{V}_{\varphi \boldsymbol{C}}^H, \tag{3.74}$$

then according to (3.61) and using (3.74), the right singular vectors of $[\hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 1} \;\; \hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 2}]$ are given by

$$\boldsymbol{V}_{\boldsymbol{C}} = \boldsymbol{Q}_{2d} \boldsymbol{V}_{\varphi \boldsymbol{C}} \begin{bmatrix}\boldsymbol{I}_M & \\ & \boldsymbol{\Pi}_M\end{bmatrix}. \tag{3.75}$$

Partitioning this similar to (3.28) we have

$$\boldsymbol{V}_{\boldsymbol{C}} = \begin{bmatrix}\boldsymbol{V}_{\boldsymbol{C}11} & \boldsymbol{V}_{\boldsymbol{C}12}\\ \boldsymbol{V}_{\boldsymbol{C}21} & \boldsymbol{V}_{\boldsymbol{C}22}\end{bmatrix}, \tag{3.76}$$

where the block elements are of dimension $d \times d$. Now, according to (3.33) we may obtain the TLS estimate as

$$\hat{\boldsymbol{\Psi}}_{TLS} = -\boldsymbol{V}_{\boldsymbol{C}12} \boldsymbol{V}_{\boldsymbol{C}22}^{-1}. \tag{3.77}$$

Signal poles are symmetric with respect to the unit circle

Properties of the TLS estimate derived in the previous section can be used to show that the estimated signal poles will be symmetric with respect to the unit circle. Notice that since $\boldsymbol{V}_{\varphi \boldsymbol{C}}$ is real, $\boldsymbol{Q}_{2d}\boldsymbol{V}_{\varphi \boldsymbol{C}}$ is column conjugate symmetric (Definition 3.3). Consequently it can be written on the form [13]

$$\boldsymbol{Q}_{2d}\boldsymbol{V}_{\varphi \boldsymbol{C}} = \begin{bmatrix} \boldsymbol{V}_A & \boldsymbol{V}_B \\ \boldsymbol{\Pi}_d \boldsymbol{V}_A^* & \boldsymbol{\Pi}_d \boldsymbol{V}_B^* \end{bmatrix}, \tag{3.78}$$

for some $\boldsymbol{V}_A,\ \boldsymbol{V}_B \in \mathbb{C}^{d\times d}$. If we insert (3.80) in (3.77) we have

$$\boldsymbol{V}_{\boldsymbol{C}} = \begin{bmatrix} \boldsymbol{V}_A & \boldsymbol{V}_B \\ \boldsymbol{\Pi}_d \boldsymbol{V}_A^* \boldsymbol{\Pi}_d & \boldsymbol{\Pi}_d \boldsymbol{V}_B^* \boldsymbol{\Pi}_d \end{bmatrix}. \tag{3.79}$$

Comparing this to (3.78) see that $\boldsymbol{V}_{\boldsymbol{C}12} = \boldsymbol{V}_{\boldsymbol{C}22}^*$ and thus the TLS solution can be found from

$$\hat{\boldsymbol{\Psi}}_{TLS} = -\boldsymbol{V}_{\boldsymbol{C}12}\boldsymbol{V}_{\boldsymbol{C}12}^{*-1} = \hat{\boldsymbol{\Psi}}_{TLS}^{*-1}. \tag{3.80}$$

Remember that the signal poles are given by the eigenvalue decomposition of $\hat{\boldsymbol{\Psi}}_{TLS}$. Since we have $\hat{\boldsymbol{\Psi}}_{TLS} = \hat{\boldsymbol{\Psi}}_{TLS}^{*-1}$, we see that the eigenvalues and thus the signal poles will be symmetric with respect to the unit circle [36]: if z_i is an eigenvalue then $1/z_i^*$ is also an eigenvalue.

Summary of total least squares solution

The efficient computation of the total least squares solution in Unitary ESPRIT can be summarized in the following steps

1. Compute the real matrix, $\varphi_Q(\hat{\boldsymbol{C}}_{\boldsymbol{Z}_C})$ using (3.64).
2. Compute the SVD of $\varphi_Q(\hat{\boldsymbol{C}}_{\boldsymbol{Z}_C})$.
3. Partition the right singular vectors according to (3.78).
4. Compute the TLS solution using (3.79).

3.2.7 Summary of the Unitary ESPRIT algorithm

The Unitary ESPRIT algorithm can be summarized in the following steps, where the dominating computations for each step is included in the rightmost column

1.	Obtain an estimate of the signal subspaces for the two subarrays as described in section 3.2.5	$2 \times$ Real $(m/2 \times M)$ SVD
2.	Solve the overdetermined system of equations $\hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 1}\hat{\boldsymbol{\Psi}} \approx \hat{\boldsymbol{E}}_{\boldsymbol{Z}_C 2}$ by means of total least squares as described in section 3.2.6	Real $(m \times 2d)$ SVD
3.	Compute the signal poles by the eigenvalue decomposition $\hat{z}_i = \lambda_i(\hat{\boldsymbol{\Psi}})$	Real $(d \times d)$ EVD

3.3 Summary

In this chapter, we have presented the Unitary ESPRIT algorithm which belongs to the class of subspace based single shift-invariance estimation methods. The data matrix used in the Unitary ESPRIT algorithm consists of a signal matrix combined with a row-reversed version of the same signal matrix, resulting in an algorithm which exploits the single shift-invariance property both forwards and backwards. This essentially doubles the data used, thus increasing the accuracy of the parameter estimation. By setting up the signal in a centro hermitian matrix we have shown, that the computational complexity of the algorihm can be reduced significantly.

Since the signal poles estimated by the Unitary ESPRIT algorithm are constrained to the unit circle, it is an accurate and efficient method for estimating parameters in a constant amplitude sinusoidal model. The purpose of this dissertation is to introduce a novel approach to estimating the most perceptually relevant parameters in a sinusoidal mdel using the Unitary ESPRIT algorithm. In order to do this, the perceptual relevance of the parameter estimates must be taken into account. This can be done by combining a psychoacoustic model with the Unitary ESPRIT algorithm, one of which we will treat in the sequel.

Chapter 4

Psychoacoustic Model

"It is the province of knowledge to speak and it is the privilege of wisdom to listen."

Oliver Wendell Holmes (1841 – 1935)

In this chapter: *We start by reviewing the psychoacoustic phenomena on which perceptual masking models are based. Then we thoroughly study a well known psychoacoustic model, namely the MPEG-1 Psychoacoustic Model 1.*

4.1 Psychoacoustics

Psychoacoustic models are based on the masking properties in the human auditory system. Masking is the process in which one sound can be rendered inaudible due to the presence of other sounds. In the following we briefly review the basic theories of human auditory perception. For a thorough introduction to psychoacoustics see e.g. [38].

4.1.1 Human auditory system

The human hearing, or the perception of audio signals, is facilitated by the human auditory system. Physiologically, the auditory system consist of three parts: the outer ear, the middle ear, and the inner ear [38, pp. 15–28].

The outer ear contributes to the spatial location of sounds by changing the spectral coloring dependent on the angle of arrival. From the outer ear the sound is transmitted to the middle ear through the ear drum.

The middle ear consist of three bones connected to each other, to the ear drum, and to the oval window in the inner ear. Through these connections, air vibrations are changed to mechanical vibrations which are transferred to the inner ear.

The inner ear (the cochlea) is shaped as a twirled up cone and filled with a liquid. The vibrations transferred to the inner ear set in motion a standing wave, which has peaks at specific locations correponding to the frequency contents of the vibration. Because of these standing wave patterns in the cochlea, it is in effect a frequency to location analyzer. Along the length of the cochlea, hair cells are excited by the vibrations. The hair cells stimulate nerve endings, which convey the information to the brain.

4.1.2 Absolute threshold of hearing

The absolute hearing threshold is "the minimum detectable level of a sound in the absence of any other external sounds" [38]. The absolute threshold depends on several factors such as the frequency characteristics of the middle ear and the phycical condition of the inner ear. The middle ear contributes to the frequency dependency of the sensitivity of the ear, because its transmission is most efficient in the range of 500 Hz to 4000 Hz. With age and especially with exposure to loud sounds, the cochlea can become damaged thus raising the threshold for specific frequencies.

By measuring the sensitivity of single tones for a large population, a model for the absolute threshold of hearing has been developed, which approximates the sensitivity for a young listener with good hearing. It has been found that the absolute threshold can be approximated with the following non-linear function [15] shown in figure 4.1

$$T_q(f) \;=\; 3.64\Big(\frac{f}{1000}\Big)^{-0.8} - 6.5e^{-0.6\left(\frac{f}{1000}-3.3\right)^2} + 10^{-3}\Big(\frac{f}{1000}\Big)^{4} \quad \text{(dB SPL)}, \tag{4.1}$$

where the frequency, f, is given in Hz and the threshold, $T_g(f)$, has the unit dB sound pressure level (SPL). The SPL is defined as the ratio between the pressure of a soundwave, p, and a reference amplitude, p_0, i.e. $L_{SPL} = 20\log_{10}(p/p_0)$. The reference level $p_0 = 20\ \mu\text{Pa} = 2 \cdot 10^{-5}\ \text{N/m}^2$ is defined so that the SPL is about 0 dB for the frequencies where the sensitivity is greatest.

4.1.3 Masking

The absolute threshold describes the hearing threshold in silence — when other sounds are present, the threshold changes. This is due to a phenomenon known as masking. "Masking is the amount (or the process) by which the threshold of audibility for one sound is raised by the presence of another (masking) sound" [38]. It is the concealment of one sensation, resulting from the presence of an other, often stronger, sensation.

Masking is often divided into two classes: non-simultaneous and simultaneous masking.

Non-simultanous masking occurs in two forms: post- and pre-masking (see figure 4.2). Post-masking has the most significant masking effect which can last up to 200 ms after the masking sound ends. The physiological cause of non-simultaneous masking is not completely understood, but it most probably stems from nerve activity which dies out slowly [38]. In some psychoacoustic models, such as the MPEG-1 Psychoacoustic Model 1, non-simultaneous masking is not included.

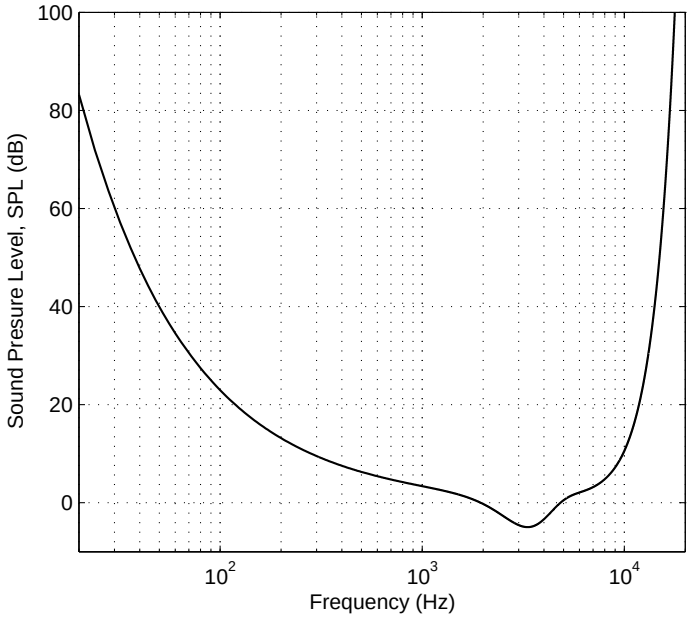

Figure 4.1: *The absolute threshold of hearing [38], i.e. the minimum detectable level of a sound in a quiet environment. The curve approximates the sensitivity of the hearing for a young listner with acute hearing.*

Simultaneous masking stems from the phenomenon of standing waves in the cochlea [38]. In figure 4.3 an example of the amplitude envelope of the standing wave along the basilar membrane in the cochlea is shown for a single frequency. As it can be seen, the standing wave amplitude envelope has a considerable peak at a specific position. At a higher frequency, the peak will be closer to the oval window and conversely further away from the oval window for a lower frequency. Figure 4.3 illustrates how the effects of a single tone is spread out in the cochlea. Because of this spreading, a single frequency is able to mask a signal of smaller amplitude located close to the masking frequency.

4.1.4 Critical bands

The spread of masking is related to the concept of critical bands. Experimental results show that the masking of a single tone is primarily affected by other sound components lying within a certain frequency dependent band, the so-called critical band. To explain this, it has been suggested that the auditory system behaves as if it consists of a non-linear distributed set of bandpass-filters known as the auditory filters. The properties of the critical bands have been investigated experimentally by numerous researchers and is well understood, although the underlying physiological explanation is not fully understood [38]. Experiments show that the critical bandwidth is almost constant (approximatly 100 Hz) at frequencies below 500 Hz. Above 500 Hz, the critical bandwidth increases to about 20% of the center frequency [15].

Although physiologically, the critial bands are continuously distributed, the auditory system is often considered as consisting of a discrete set of band-pass filters. Using 25 critical bandwidth

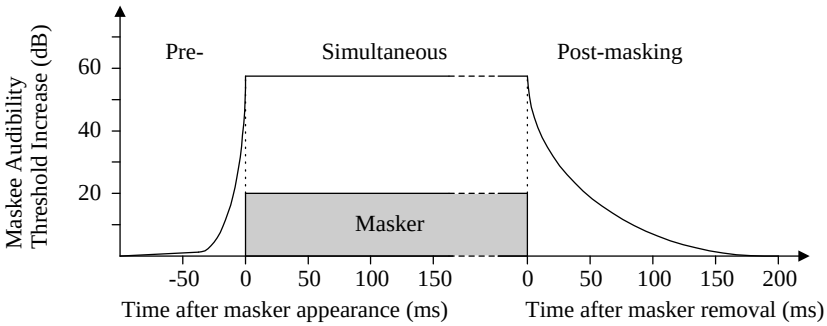

Figure 4.2: *An illustration of the masking phenomenon [15]: When a masker is present, the threshold of audibility for the masked sound is raised. Non-simultaneously masking can occur right before (<50 ms) or after (<200 ms) a masking sound.*

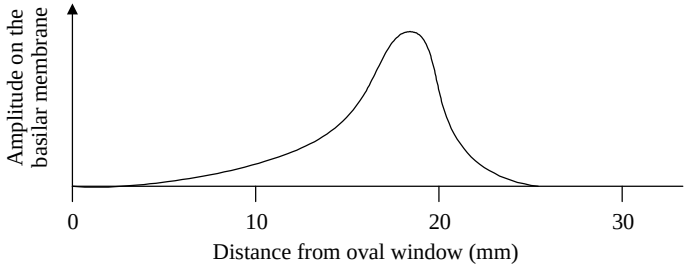

Figure 4.3: *The amplitude of the standing wave along the basilar membrane [38]. The wave is generated by a 1600 Hz tone, coming from the oval window located to the left. The figure illustrates how the effects of a single tone is spread out in the cochlea. Because of this spreading, a single frequency is able to mask a signal of smaller amplitude located close to the masking frequency.*

auditory filters which span the audio spectrum, the Bark scale has been defined in which one Bark corresponds to one critical band. The following expression approximates the relation between frequencies in Hz and the Bark scale [15]

$$z(f) \;=\; 13\tan^{-1}(7.6\cdot 10^{-4}f) + 3.5\tan^{-1}\left(\frac{f^2}{7500^2}\right) \quad \text{(Bark)}. \tag{4.2}$$

In figure 4.4, $z(f)$ is depicted along with the location of the center frequencies of the 25 critical bands.

4.1.5 Types of masking

The masking of one signal by another is a complex function of the signal spectra. For the purpose of modeling perceptual masking, we often distinguish between two simple types of masking: noise-maksing-tone and tone-masking-noise.

In figure 4.5 it is illustrated how a narrow-band noise signal masks the presence of a tone. The signal-to-mask ratio (SMR) describes the smallest difference between the intensity of the masking signal and the intensity of the masked signal [15]. Experiments show that the SMR is around 4

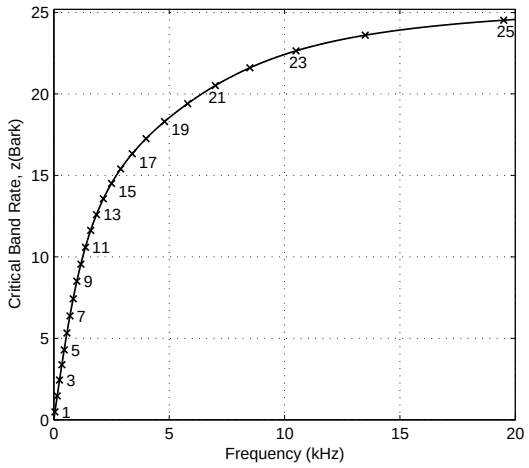

Figure 4.4: *The relation between frequency in Hz and the Bark scale (4.2). The x–marks, enumerated 1-25 in increments of 2, mark the center frequencies of the 25 critical bands, distributed linearly on the Bark scale.*

dB for a noise masker [15]. In figure 4.6 it is illustrated how a single tone can mask a narrow-band noise signal. Experiments show that the SMR is around 24 dB for a tonal masker [15]. Comparing the SMR of the tonal masker and the noise masker, it is obvious that noise has better masking abilities than a pure tone.

4.2 MPEG-1 Psychoacoustic Model 1

The MPEG-1 Psychoacoustic Model 1 is based on the psychoacoustic properties presented in the previous section. Here, we describe how the model is implemented using these properties. This section is based on the MPEG-1 Standard [39] and a tutorial by Painter et al. [15]. The aim of the MPEG-1 Psychoacoustic Model 1 is to estimate a global masking threshold for an arbitrary signal.

4.2.1 Spectral analysis and SPL normalization

The MPEG-1 Psychoacoustic Model 1 can opererate in two modes, which differ in the frequency resolution: it is based on either a 512 or a 1024 point DFT. The two modes in the MPEG-1 Psychoacoustic Model 1 are very similar and for the sake of simplicity, we only describe the algorithm based on a 512 point DFT which yields a frequency resolution of 86.13 Hz at a sample rate of 44.1 kHz.

Consider a signal, $s(k)$, which is assumed to have a maximum amplitude of ± 1. This is divided by the FFT-length, N, to achieve a 0 dB maximum after the DFT

$$x(k) = \frac{s(k)}{N}. \tag{4.3}$$

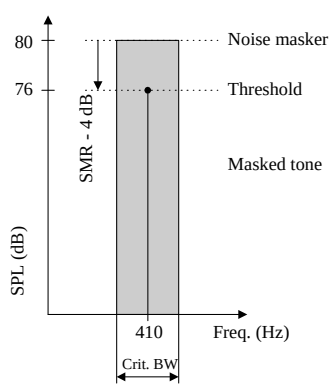

Figure 4.5: *Illustration of a noise masker and a tone at the threshold of detection. The pure tone, located at 410 Hz with a SPL of 76 dB, is just masked by the narrow-band noise with a bandwidth of 1 Bark and an overall intesity of 80 dB.*

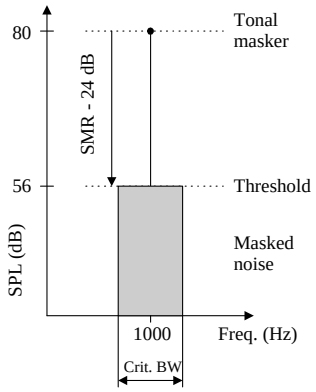

Figure 4.6: *Illustration of a tonal masker and narrow-band noise at the threshold of detection. The pure tone, located at 1000 Hz with an SPL of 80 dB, just masks a narrow-band noise signal with a bandwidth of 1 Bark and an overall intensity of 56 dB.*

The signal is then windowed by a Hann window, $w(k)$, and the power spectrum is calculated using the DFT. A power normalization term, $PN = 90.302$ dB is added, in order to set the maximum amplitude to $PN = 90.302$ dB SPL.

$$P(l) = PN + 10\log_{10}\left|\sum_{k=0}^{N-1} w(k)x(k)e^{-j(2\pi lk/N)}\right|^2, \quad 0 \le l \le \frac{N}{2}. \tag{4.4}$$

The term PN is added in order to normalize the power spectrum to correspond to a worst case sound pressure level. Since the sound pressure level at which the sound is played cannot be determined at the point of analysis, the signal is assumed to be played at a loudness where 0 dB SPL corresponds to ± 1 least significant bit. Because the psychoacoustic model is designed to work with 16 bit resolution, the maximum sound pressure level is set at 90.302 dB.

In figure 4.7, the SPL normalized power spectrum of a signal is showed. This signal will function as an example in the following. The example signal is taken from a segment of pop-music sampled at 44.1 kHz, and it contains both tonal and noise-like components. The signal is plotted on the Bark scale in order to best visualize the perceived spectrum of the signal.

4.2.2 Identification of tonal and noise maskers

When the power spectrum has been determined, tonal maskers and noise maskers are identified.

Tonal maskers are defined to exist at local maxima in the power spectrum, where the peak is more than 7 dB higher than the neighboring spectral components within a certain window. This window is defined as a frequency dependent number of spectral bins. Tonal maskers are identified at the following frequency bins

$$S_T = \left\{ l \;\middle|\; \begin{array}{c} P(l) > P(l \pm 1), \\ P(l) > P(l \pm \Delta_l) + 7 \text{ dB} \end{array} \right\}, \tag{4.5}$$

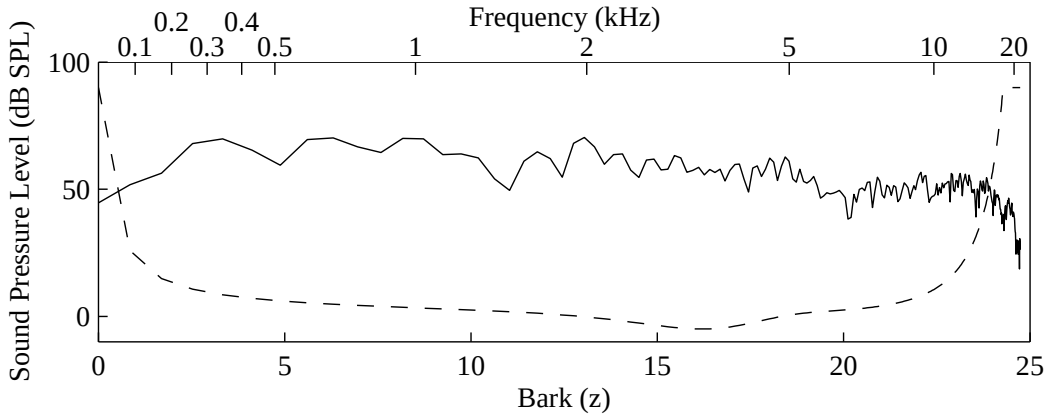

Figure 4.7: *The power spectrum of an example signal segment normalized to dB sound pressure level. The signal, sampled at 44.1 kHz, is a segment of pop-music, containing both tonal and noise-like components. The absolute threshold of hearing is plotted as a dashed line. The frequency axis is on the Bark scale.*

where the distance, Δ_k, is defined as,

$$\Delta_l \in \begin{cases} 2 & 2 < l < 63 \quad (0.17\text{–}5.5 \text{ kHz}) \\ [2,3] & 63 \leq l < 127 \quad (5.5\text{–}11 \text{ kHz}). \\ [2,6] & 127 \leq l < 256 \quad (11\text{–}20 \text{ kHz}) \end{cases} \tag{4.6}$$

The power of the tonal maskers $P_{TM}(k)$ is calculated from the sum of the frequency bins of the tonal masker and the two neighboring bins

$$P_{TM}(l) = 10 \log_{10} \sum_{i=-1}^{1} 10^{0.1P(i+l)} \text{ (dB)}. \tag{4.7}$$

The tonal maskers identified in the example signal is shown in figure 4.8, indicated by the symbol 'x'.

Noise maskers are computed form the part of the spectrum in which no tonal maskers reside. One noise masker is computed for each critical band, and the location of the noise maskers is defined as the geometric mean $\bar{l}$ of the frequency bins in the critical band,

$$\bar{l} = \left(\prod_{l=a}^{b} l \right)^{1/(b-a+1)}. \tag{4.8}$$

Where a and b respectively denote the lower and upper spectral bins associated with the critical band. The power of the noise maskers $P_{NM}(\bar{l})$ is computed from the spectral bins which are not within the neighboring window of a tonal masker

$$\begin{aligned} P_{NM}(\bar{l}) &= 10 \log_{10} \sum_{l=a}^{b} 10^{0.1P(l)} \text{ (dB)} \\ &\forall\ l \notin \{S_T, S_T \pm 1, S_T \pm \Delta_k\}. \end{aligned} \tag{4.9}$$

The noise maskers identified in the example signal is shown in figure 4.8, indicated by the symbol 'o'.

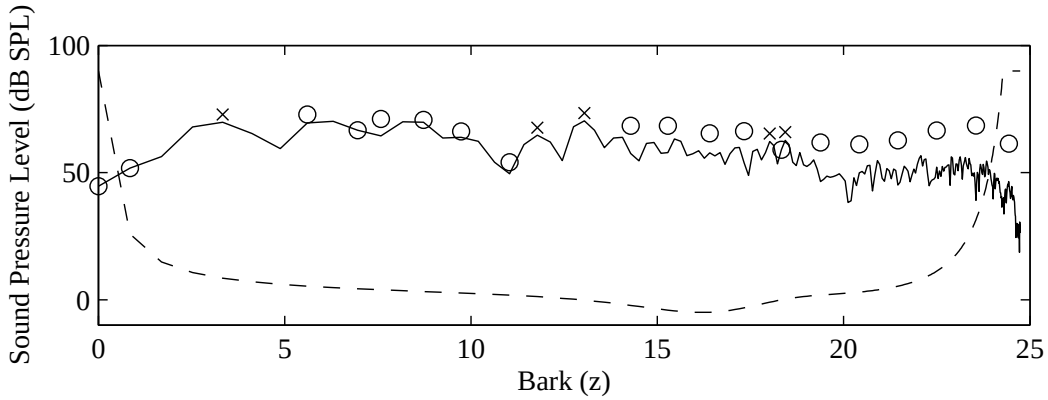

Figure 4.8: *Identified tonal and noise maskers for an example signal segment. Tonal maskers are indicated by the symbol 'x' and noise maskers are indicated by the symbol 'o'.*

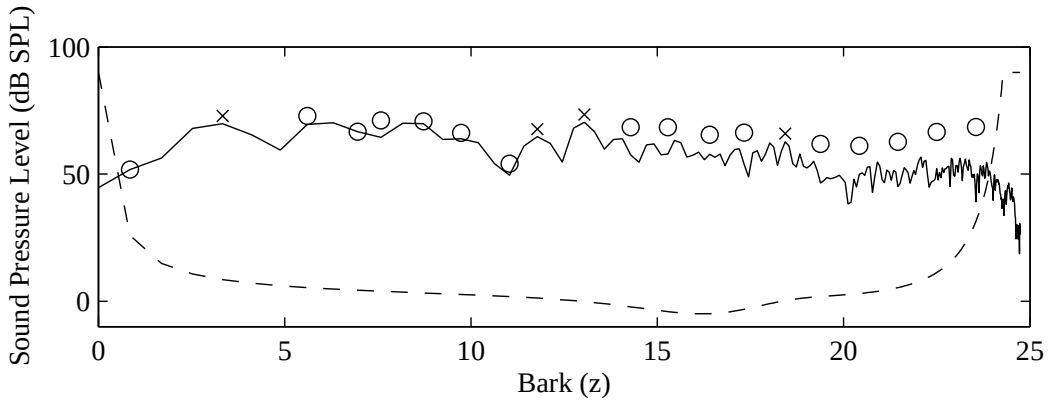

Figure 4.9: *Decimated and reorganized tonal and noise maskers. Tonal maskers are indicated by the symbol 'x' and noise maskers are indicated by the symbol 'o'. Compared to figure 4.8 the removal of maskers is exemplified at around 17 bark, where a tonal and noise masker is removed.*

4.2.3 Decimation and reorganization of maskers

In order to reduce computational complexity and avoid redundant maskers, some of the maskers are removed under certain circumstances. Firstly, maskers below the absolute threshold of hearing are discarded. Thus, any masker that does not satisfy

$$P_{TM,NM}(l) \geq T_q(l) \tag{4.10}$$

is removed. Secondly, to avoid clustering of maskers, a sliding window of one-half Bark length is moved across the maskers. If two or more maskers are present at any time within the one-half Bark window, only the most powerful masker is retained. For the example signal, the result of this is shown in figure 4.9.

Next, the higher frequencies are also subsampled to reduce the spectral resolution at these frequencies. In the critical bands 18–22, the frequency bins are decimated by 2:1 and in the critical bands 23–25 by 4:1. The total of 256 frequency bins is hereby reduced to 106 bins.

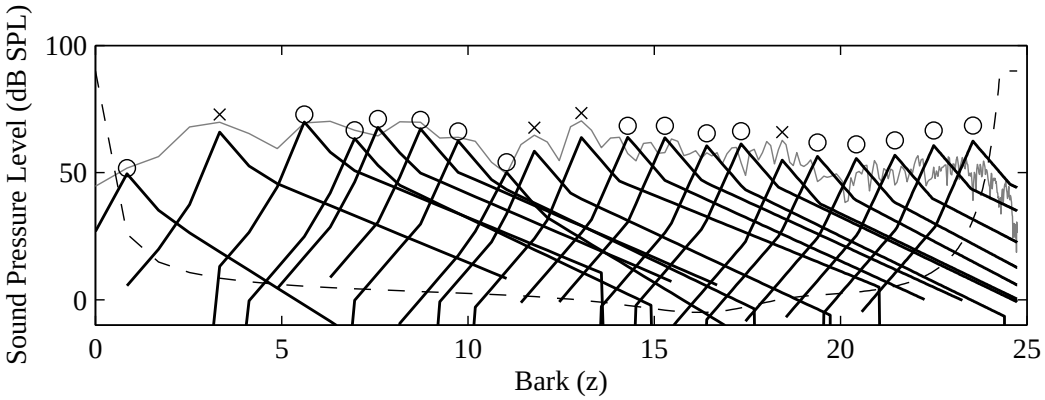

Figure 4.10: *The individual masking thresholds for tonal and noise maskers. This figure shows the spreading of the individual maskers. Notice how the noise maskers "o" have a higher masking threshold than tonal maskers "x".*

4.2.4 Calculation of individual masking thresholds

Before the global masking threshold can be computed, the individual contributions from each masker is computed. This is done by, for each masker denoted by i, computing the spread of masking in the adjacent frequency bins denoted by l. The threshold of masking for the tonal maskers is given by,

$$T_{TM}(l,i) = P_{TM}(i) - 0.275z(i) + SF(l,i) - 6.025 \text{ (dB SPL)}. \tag{4.11}$$

Where $P_{TM}(i)$ is the SPL of the tonal masker located at position i. $z(i)$ is the Bark scale indexed by the frequency bin i. $SF(l,i)$ describes the spread of masking from masker bin i to maskee bin l. The spread of masking is approximated by,

$$SF(l,i) = \begin{cases} 17\Delta_z - 0.4P_{TM}(i) + 11, & -3 \leq \Delta_z < -1 \\ \big(0.4P_{TM}(i) + 6\big)\Delta_z, & -1 \leq \Delta_z < 0 \\ -17\Delta_z, & 0 \leq \Delta_z < 1 \\ \big(0.15P_{TM}(i) - 17\big)\Delta_z - 0.15P_{TM}(i), & 1 \leq \Delta_z < 8 \end{cases} \text{ (dB SPL)}. \tag{4.12}$$

The separation between the masker and maskee is defined by Δ_z as $z(l) - z(i)$. The individual thresholds of masking for noise maskers is described similarly as

$$T_{NM}(l,i) = P_{NM}(i) - 0.175z(i) + SF(l,i) - 2.025 \text{ (dB SPL)}. \tag{4.13}$$

For the example signal, the individual masking thresholds for the identified tonal and noise maskers are shown in figure 4.10.

4.2.5 Calculation of global masking threshold

Finally, the masking threshold of the individual maskers are combined to yield the global masking threshold. For each frequency bin, all the individual maskers and the absolute threshold of hearing are added together on a linear intensity scale and converted back to the dB SPL scale.

$$T_g(l) = 10\log_{10}\left(10^{0.1T_q(l)} + \sum_i 10^{0.1T_{TM}(l,i)} + \sum_i 10^{0.1T_{NM}(l,i)}\right) \text{ (dB SPL)}. \tag{4.14}$$

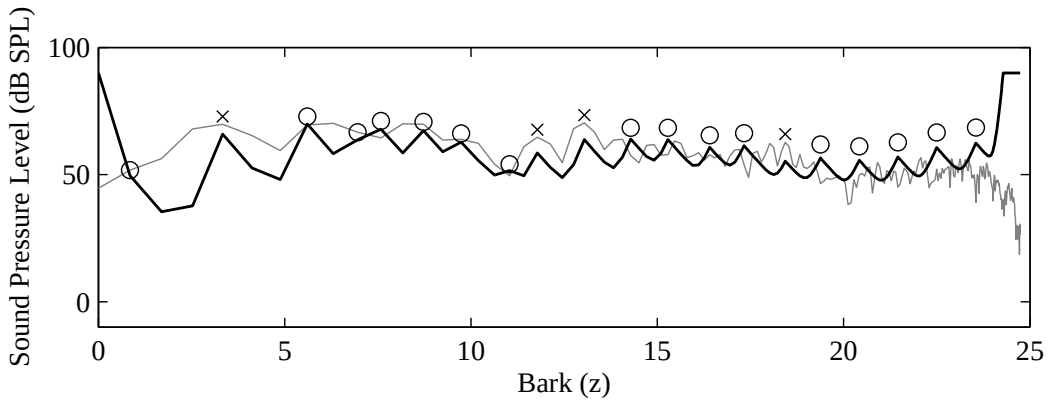

Figure 4.11: *The final gobal masking threshold for the example signal segment as estimated by the MPEG-1 Psychoacoustic Model 1.*

This concludes the computation of the global masking threshold as defined in the MPEG-1 Psychoacoustic Model 1. For the example signal, the global masking threshold is shown in figure 4.11.

4.3 Summary

In this chapter we have investigated the physiological principles which acount for the masking effects in the human auditory system. These masking effects, combined with the absolute threshold of hearing, are the building blocks of the MPEG-1 Psychoacoustic model 1. Using this psychoacoustic model, we can calculate the global masking threshold for an arbitrary signal segment. This threshold describes the frequency dependent sensitivity of the ear when a masking sound is present and describes in effect the signal-to-mask ratio. Since the most perceptually relevant frequency components are the components with the greatest signal-to-mask ratio, the global masking threshold can be used to find the most perceptually relevant components of the signal segment. The next step is to introduce the global masking threshold in the Unitary ESPRIT algorithm, which we will treat in the sequel.

Chapter 5

Perceptual Unitary ESPRIT

“I don’t think necessity is the mother of invention — invention, in my opinion, arises directly from idleness, possibly also from laziness. To save oneself trouble.”

Agatha Christie (1890 – 1976)

In this chapter: *We present a novel approach to estimating the most perceptually relevant parameters in a sinusoidal audio model: We propose a method which incorporates the perceptual model from the MPEG-1 standard in the Unitary ESPRIT algorithm. We start by discussing how the perceptual distortion of a signal can be measured, based on information from a psychoacoustic model. Then, we introduce methods for incorporating perceptual knowledge in the estimation of signal frequencies by means of Unitary ESPRIT as well as in the estimation of amplitudes and phases.*

5.1 Perceptual distortion

To include a perceptual model in the Unitary ESPRIT algorihm such that perceptually relevant signal parameter estimates can be made, we must modify the algorithm such that the perceptual distiortion of the signal is minimized. Thus, we must use a distortion measure which takes the psychoacoustic properties of the human auditory system into account.

5.1.1 Perceptual distortion measure

Let us define the distortion of a signal estimate as the difference between the original signal, $x(k)$, and the modeled signal, $\hat{x}(k)$

$$\Delta x(k) = x(k) - \hat{x}(k). \tag{5.1}$$

When applying an analysis window, $w(k)$, which defines the signal segment to be analyzed, taking the discrete Fourier tranform yields

$$E(\omega) = \sum_{k=-\infty}^{\infty} w(k)\Delta x(k)e^{-j\omega k}, \tag{5.2}$$

where ω is the normalized frequency in radians per sample. Now, we define a perceptual distortion measure, D, as a weighted integral of the signal distortion in the frequency domain [40]

$$D = \frac{1}{2\pi} \int_{2\pi} H^2(\omega) \big| E(\omega) \big|^2 d\omega, \tag{5.3}$$

where $H^2(\omega)$ is non-negative and real for all ω. $H^2(\omega)$ "is a weighting function representing the sensitivity of the human auditory system which we will generally select to be the inverse of the masking threshold" [40].

If we define $h(k)$ as the inverse Fourier transform of $H(\omega)$ we may equally express the perceptual distortion measure, D, in terms of the following infinite convolution sum [1]

$$D = \sum_{k=-\infty}^{\infty} \big| h(k) * w(k) \Delta x(k) \big|^2, \tag{5.4}$$

where $*$ denotes the convolution operation. This infinite summation is recognized as the squared vector ℓ_2 norm of the convolution sequence and we may thus write

$$D = \big|\big| h(k) * w(k) \Delta x(k) \big|\big|_2^2. \tag{5.5}$$

Alternatively we may express the perceptual distortion measure in terms of a matrix vector multiplication

$$D = \big|\big| \boldsymbol{H} \boldsymbol{W} \boldsymbol{\Delta x} \big|\big|_2^2, \tag{5.6}$$

where $\boldsymbol{H}$ is an infinite Toeplitz filter matrix constructed from the filter impulse response $h(k)$. $\boldsymbol{W}$ is a diagonal matrix with the window, $w(k)$, on the main diagonal, and $\boldsymbol{\Delta x}$ is the distortion signal vector.

The ℓ_2 norm of the distortion signal, $||\Delta x(k)||_2$, corresponds to the energy of the distortion. However, the signal energy is not necessarily in correspondence with the perceived signal distortion. An intuitive understanding of the perceptual distortion measure, as shown in (5.5) and (5.6), is to see it as a weighted signal norm. When the signal distortion, $\Delta x(k)$, is windowed and convolved with the perceptual weighting filter, $h(k)$, it is transformed into a domain in which the ℓ_2 norm is in better correspondence with the perceived signal distortion.

5.1.2 Perceptual weighting filter design

In the following, we consider methods in which the perceptual weighting filter is approximated by a finite impulse response (FIR) filter derived from the psychacoustic model described in chapter 4. Thus, we need to design an FIR filter based on the estimated masking threshold curve for a signal segment. A variety of methods for designing an FIR filter with arbitrary frequency response exist, such as the frequency sampling method and the Parks-McClellan method [41]. As an example, we here describe the frequency sampling method.

We start with a specification of the desired frequency response of the filter given by a set desired filter magnitudes at a corresponding set of frequencies. This is given by the inverse of the estimated masking threshold.

$$H^2(\omega_l) = \frac{1}{T_g(l)}, \tag{5.7}$$

where ω_l is the frequency corresponding to the lth frequency bin in the global masking threshold $T_g(l)$. We interpolate this desired frequency response onto a dense evenly spaced frequency grid of length $Q/2$. Let the frequencies in this grid be denoted by

$$\omega_i = \frac{2\pi i}{Q}, \qquad i = 0, 1, \ldots, \frac{Q}{2} - 1. \tag{5.8}$$

The frequency response and the filter coefficients of an FIR filter are related by the discrete Fourier transform (DFT)

$$H(\omega) = \sum_{k=0}^{Q-1} h(k) e^{-j\omega k}, \tag{5.9}$$

and thus, the response specified at the frequency grid is related to the filter coefficients by

$$H(\omega_i) = \sum_{k=0}^{Q-1} h(k) e^{-j2\pi ik/Q}. \tag{5.10}$$

Isolating $h(k)$ gives an expression for the filter coefficients in terms of $H(\omega_i)$

$$h(k) = \frac{1}{Q} \sum_{i=0}^{Q-1} H(\omega_i) e^{j2\pi ik/Q}, \qquad k = 0, 1, \ldots, Q-1. \tag{5.11}$$

Note that this is simply the inverse discrete Fourier transform (IDFT) of $H(\omega_i)$. Finally, the filter coefficients are windowed to give a filter of the desired impulse response length, q

$$h_q(k) = w_q(k) h(k), \tag{5.12}$$

where $w_q(k)$ is a window function of length q. An example of an FIR filter computed from a masking threshold curve for a sample signal segment is shown in figure 5.1. Because of the linear frequency resolution of an FIR filter (as compared to the frequency resolution of the human auditory system), the filter fits the desired frequency response best at high frequencies.

5.2 Signal prefiltering

The Unitary ESPRIT algorithm gives estimates of the d most powerful frequency components in a signal. However, the most powerful components are not necessarily the most perceptually relevant. To ensure that the signal parameters estimated by the Unitary ESPRIT are the most perceptually relevant, we propose to prefilter the signal matrix using the weighting filter $h_q(k)$ derived from the perceptual masking model, $T_g(l)$.

Prefiltering is not trivial since the filtering process must not disturb the underlying signal model on which the Unitary ESPRIT algorithm relies. This means that the rank and the rotational invariance properties of the original signal matrix must be retained in the prefiltered signal matrix.

Several methods for signal matrix prefiltering for use in other subspace based parameter estimation methods have been proposed (cf. [42],[43],[1]). In the following we show how the signal model is affected by different filtering operations and based on this we propose suitable prefiltering methods for the Unitary ESPRIT algorithm.

We identify two different strategies for prefiltering the signal prior to its use in the Unitary ESPRIT algorithm.

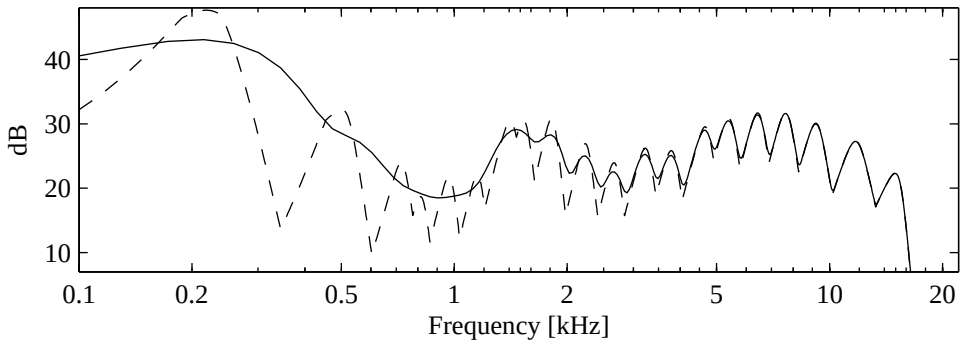

Figure 5.1: *Example of an FIR filter designed to approximate the inverse of a perceptual masking curve. The dashed line indicates the desired frequency response (the inverse of the masking curve), and the full line shows the frequency of an FIR filter designed by the frequency sampling method to match the desired frequency response. The filter impulse repsonse length is $q = 256$, and the window function used in the design is a Kaiser windows with the parameter $\beta = 10$. The masking curve is estimated by the MPEG-1 Psychoacoustic Model 1 as described in section 4.2 for an example signal segment of length 512 from a piece of pop music sampled at 44.1 kHz. The signal segment is the same as used as an example in section 4.2.*

1. The signal is filtered prior to being arranged in the signal matrix.
2. The signal is arranged in the signal matrix, and the individual rows of the signal matrix are filtered seperately.

In the following, we denote these two strategies "signal vector prefiltering" and "signal matrix prefiltering" respectively. Both strategies offer useful results as we will show in the following.

5.2.1 Signal vector prefiltering

One method we may use for prefiltering is to filter the signal prior to arranging it in the signal matrix. To understand how such a prefiltering affects the signal model, consider the noise free sinusoidal signal model

$$x(k) = \sum_{i=1}^{d} s_i e^{j\omega_i k}. \tag{5.13}$$

This model corresponds to an autoregressive moving average (ARMA) model with d poles and $d-1$ zeros. Taking the Z-transform of equation 5.13 yields

$$X(z) = \sum_{i=1}^{d} \frac{s_i}{1 - z_i z^{-1}} = \frac{\sum_{i=0}^{d-1} b_i z^{-i}}{1 + \sum_{i=1}^{d} a_i z^{-i}}, \tag{5.14}$$

where $z_i = s_i e^{jk\omega_i}$ are the signal poles, and b_i and a_i are the moving average (MA) and autoregressive (AR) parameters respectively. Returning to the time domain, the signal can be described by the following difference equation

$$x(k) = \sum_{i=0}^{d-1} b_i \delta(k-i) - \sum_{i=1}^{d} a_i x(k-i). \tag{5.15}$$

Now, consider filtering the signal with an FIR filter of order q

$$H(z) = \sum_{i=0}^{q} h_i z^{-i}. \tag{5.16}$$

Denoting the filtered signal in the z-domain $Y(z)$, we have

$$\begin{aligned} Y(z) &= H(z)X(z) && (5.17)\\ &= \sum_{i=0}^{q} h_i z^{-i} \frac{\sum_{i=0}^{d-1} b_i z^{-i}}{1 + \sum_{i=1}^{d} a_i z^{-i}} && (5.18)\\ &= \frac{\sum_{i=0}^{d+q-1} c_i z^{-i}}{1 + \sum_{i=1}^{d} a_i z^{-i}}, && (5.19) \end{aligned}$$

where c_i are the new MA coefficients of the filtered signal. Naturally, the poles of the filtered signal are equal to those of the original signal, but the number of zeros has increased to $d+q-1$. Taking the inverse z-transform, the filtered signal can be described by the following difference equation

$$y(k) = \sum_{i=0}^{d+q-1} c_i \delta(k-i) - \sum_{i=1}^{d} a_i y(k-i). \tag{5.20}$$

With zero initial conditions, $y(-1) = \cdots = y(-d) = 0$, we may write the forward recursion of the difference equation in matrix form as [43]

$$\begin{bmatrix} y(0) \\ y(1) \\ \vdots \end{bmatrix} = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{d+q-1} \\ 0 \\ \vdots \end{bmatrix} - \begin{bmatrix} & & & 0 \\ & & \iddots & y(0) \\ & \iddots & y(0) & y(1) \\ & & & \vdots \\ y(0) & y(1) & \cdots & y(d-1) \\ y(1) & y(2) & \cdots & y(d) \\ \vdots & \vdots & & \vdots \end{bmatrix} \begin{bmatrix} a_d \\ a_{d-1} \\ \vdots \\ a_1 \end{bmatrix}. \tag{5.21}$$

Examining this set of equations gives a lot of information on $y(n)$. If we look at the part of the matrix equation starting from the row yielding $y(q+d)$ we have

$$\begin{bmatrix} y(d+q) \\ y(d+q+1) \\ \vdots \end{bmatrix} = - \begin{bmatrix} y(q) & y(q+1) & \cdots & y(q+d-1) \\ y(q+1) & y(q+2) & \cdots & y(q+d) \\ \vdots & \vdots & & \vdots \end{bmatrix} \begin{bmatrix} a_d \\ a_{d-1} \\ \vdots \\ a_1 \end{bmatrix}. \tag{5.22}$$

The matrix in this equation is a Hankel structured matrix consisting of the filtered signal, $y(k)$, starting at $y(q)$. This filtered signal matrix has rank d since there are d independent columns. Notice, that extending the signal matrix with extra columns on the right does not increase its rank, since these extra columns will be linear combinations of the existing columns. This is not the case if samples of $y(k)$ for $k < q$ are included in the signal matrix. Thus, in order to retain the rank property of the existing signal, the first q rows of the filtered signal matrix must be truncated. This corresponds to truncating the part of the signal matrix which is affected by the choice of initial conditions for the FIR filter.

Since we must truncate the first q rows of the signal matrix, we possibly discard useful information. The prefiltered signal matrix constructed using this method will not have the same dimensions as the original signal matrix of size $m \times M$, although it does retain the rank and rotational invariance properties of the original signal matrix. This will affect the estimation accuracy of the unitary ESPRIT algorithm. To overcome this, we could use a signal block of $N + q$ samples. After filtering and discarding the first q samples of the filtered signal, a signal matrix of size $m \times M$ where $N = m+M+1$ can be formed. However, this will affect the stationarity assumptions, since now a signal block of size $N + q$ must be assumed stationary.

Example 5.1. *Consider a signal, $x(k)$, consisting of two cosines with unit amplitude at frequencies $\omega_1 = 0.3\pi$ and $\omega_2 = 0.7\pi$ with random phase in additive white gaussian noise, $n(k)$, with a variance of $\sigma_n = 0.1$.*

$$x(k) = cos(\omega_1 k + p_1) + cos(\omega_2 k + p_2) + n(k),$$

where p_1 and p_2 are random variables distributed evenly on the interval $[0, 2\pi)$. Since each cosine can be written as a sum of two complex sinusoids, the signal, $x(k)$, consists of 4 cisoids plus noise.

Consider now a fourth order low pass filter with a cut-off frequency at $\omega_n = 0.3\pi$ where the $q = 5$ filter coefficients are given by $h(n) = \{0.0201, 0.2309, 0.4981, 0.2309, 0.0201\}$. The frequency response of the filter can be seen in figure 5.2. Using this, the filtered signal will be dominated by the cosine at frequency $\omega_1 = 0.3\pi$. Thus, using the prefiltered Unitary ESPRIT algorithm, assuming a signal order of 2, we should be able to estimate ω_1. We now filter a signal block of

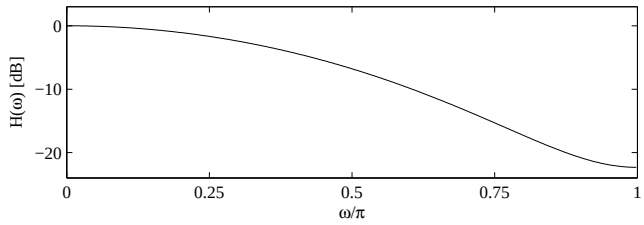

Figure 5.2: *Frequency response of the fourth order low-pass filter used in example 5.1.*

length $N = 50$ by the filter $h(k)$. Then, we discard the first q samples of the filtered signal, and arrange the remaining samples in a Hankel structured matrix $\boldsymbol{X}$ of size 25×21. Finally, we use the Unitary ESPRIT algorithm as described in section 3.2.7 on the matrix $\boldsymbol{X}$ to estimate the frequency ω_1. We wish to examine the statistical properties of the frequency estimation using this method, i.e. the mean value and the variance of the estimated frequency. Doing 100 Monte Carlo[1] *runs we get the following estimates of the mean value and variance of the estimated frequency, ω_1.*

$$\begin{aligned} \hat{m}_{\omega_1} &= 0.300\pi \\ \hat{\sigma}_{\omega_1} &= 3.58 \cdot 10^{-6} \end{aligned}$$

We see that using the method described, the frequency ω_1 was estimated without significant bias and with a relatively low variance.

□

[1] A Monte Carlo method can be defined as any method which solves a problem numerically by generating a suitable number of random inputs, generating corresponding outcomes, and observing some statistical property of these. Monte Carlo methods are useful for finding solutions to problems which are too complex to solve analytically.

5.2.2 Signal matrix prefiltering

The second method we may use to prefilter the signal is to first arrange it in the signal matrix and then filter each row of the signal matrix individually.

It is important to ensure that the prefiltering does not affect the underlying signal model — otherwise the Unitary ESPRIT algorithm will not apply to the filtered signal matrix. That is, the filtered signal matrix must retain the rank and rotational invariance properties of the original signal matrix.

Filtering the FB signal matrix

In the Unitary ESPRIT algorithm, we estimate the signal subspaces of the subarrays based on the column space of the FB signal matrix, i.e. the left singular vectors. By right multiplying the FB signal matrix by a full rank filter matrix, the rank property of the FB signal matrix is retained, since multiplication by a full rank matrix does not change the rank [17, p. 250]. Also, the column space of the FB signal matrix and the filtered FB signal matrix will be equal [43], and thus it will retain the rotational invariance property of the original signal matrix [43].

We propose to use a block diagonal filter matrix, such that the filtering corresponds to prefiltering the signal matrix $\boldsymbol{X}(k)$ prior to arranging it in the FB signal matrix, $\boldsymbol{Z}(k)$.

$$\boldsymbol{H}_{\boldsymbol{Z}} = \begin{bmatrix} \boldsymbol{H} & \\ & \boldsymbol{H} \end{bmatrix}, \tag{5.23}$$

where $\boldsymbol{H}$ is an $M \times M$ filter matrix, the specific structure of which we will treat in the sequel. Thus, we may write the filtered FB signal matrix in the following form

$$\boldsymbol{Z}(k)\boldsymbol{H}_{\boldsymbol{Z}} = \begin{bmatrix} \boldsymbol{X}(k)\boldsymbol{H} & \boldsymbol{\Pi}_m \boldsymbol{X}^*(k)\boldsymbol{H} \end{bmatrix}. \tag{5.24}$$

By writing out the convolution sequence, it can be verified that $\boldsymbol{X}(k)\boldsymbol{H}$ indeed is a matrix where each row of $\boldsymbol{X}$ is convolved by the filter in $\boldsymbol{H}$.

To see that the column space of the FB signal matrix and the filtered FB signal matrix are equal, consider the following: Let the rank of the FB signal matrix be denoted by r, $rank\big(\boldsymbol{Z}(k)\big) = rank\big(\boldsymbol{Z}(k)\boldsymbol{H}_{\boldsymbol{Z}}\big) = r$, and let the SVD of the FB signal matrix $\boldsymbol{Z}(k)$ be given by

$$\boldsymbol{Z}(k) = \boldsymbol{U}_{\boldsymbol{Z}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\boldsymbol{V}_{\boldsymbol{Z}}^H = [\boldsymbol{U}_{\boldsymbol{Z}1}\ \boldsymbol{U}_{\boldsymbol{Z}2}] \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{Z}1} & \\ & \boldsymbol{\Sigma}_{\boldsymbol{Z}2} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_{\boldsymbol{Z}1}^H \\ \boldsymbol{V}_{\boldsymbol{Z}2}^H \end{bmatrix}, \tag{5.25}$$

where $\boldsymbol{U}_{\boldsymbol{Z}1}$ contains the first r columns of $\boldsymbol{U}_{\boldsymbol{Z}}$ and thus spans the range of the FB signal matrix. Similarly, let the SVD of the filtered FB signal matrix $\boldsymbol{Z}(k)\boldsymbol{H}$ be denoted by

$$\boldsymbol{Z}(k)\boldsymbol{H}_{\boldsymbol{Z}} = \boldsymbol{U}_{\boldsymbol{ZH}}\boldsymbol{\Sigma}_{\boldsymbol{ZH}}\boldsymbol{V}_{\boldsymbol{ZH}}^H = [\boldsymbol{U}_{\boldsymbol{ZH}1}\ \boldsymbol{U}_{\boldsymbol{ZH}2}] \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{ZH}1} & \\ & \boldsymbol{\Sigma}_{\boldsymbol{ZH}2} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_{\boldsymbol{ZH}1}^H \\ \boldsymbol{V}_{\boldsymbol{ZH}2}^H \end{bmatrix}. \tag{5.26}$$

Since both matrices, $\boldsymbol{U}_{\boldsymbol{Z}}$ and $\boldsymbol{U}_{\boldsymbol{ZH}}$, are unitary, there must exist a unitary matrix, $\boldsymbol{\Omega}$, such that [17]

$$\boldsymbol{U}_{\boldsymbol{ZH}} = \boldsymbol{U}_{\boldsymbol{Z}}\boldsymbol{\Omega}, \tag{5.27}$$

$$[\boldsymbol{U}_{\boldsymbol{ZH}1}\ \boldsymbol{U}_{\boldsymbol{ZH}2}] = [\boldsymbol{U}_{\boldsymbol{Z}1}\ \boldsymbol{U}_{\boldsymbol{Z}2}] \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix}. \tag{5.28}$$

From (5.25) and (5.26) we have [43]

$$\boldsymbol{U}_{\boldsymbol{Z}1}\boldsymbol{\Sigma}_{\boldsymbol{Z}1}\boldsymbol{V}_{\boldsymbol{Z}1}^{H}\boldsymbol{H} = \boldsymbol{U}_{\boldsymbol{ZH}1}\boldsymbol{\Sigma}_{\boldsymbol{ZH}1}\boldsymbol{V}_{\boldsymbol{ZH}1}^{H}. \tag{5.29}$$

Isolating $\boldsymbol{U}_{\boldsymbol{ZH}1}$ yields

$$\boldsymbol{U}_{\boldsymbol{ZH}1} = \boldsymbol{U}_{\boldsymbol{Z}1}\boldsymbol{\Sigma}_{\boldsymbol{Z}1}\boldsymbol{V}_{\boldsymbol{Z}1}^{H}\boldsymbol{H}\boldsymbol{V}_{\boldsymbol{ZH}1}^{H}\boldsymbol{\Sigma}_{\boldsymbol{ZH}1}^{-1}. \tag{5.30}$$

By comparing (5.30) and (5.28) we see that the following relations must hold

$$\boldsymbol{\Omega}_{11} = \boldsymbol{\Sigma}_{\boldsymbol{Z}1}\boldsymbol{V}_{\boldsymbol{Z}1}^{H}\boldsymbol{H}\boldsymbol{V}_{\boldsymbol{ZH}1}^{H}\boldsymbol{\Sigma}_{\boldsymbol{ZH}1}^{-1}, \tag{5.31}$$

$$\boldsymbol{\Omega}_{12} = \boldsymbol{0}. \tag{5.32}$$

Now, since $\boldsymbol{\Omega}$ is unitary, such that $\boldsymbol{\Omega}\boldsymbol{\Omega}^{H} = \boldsymbol{\Omega}^{H}\boldsymbol{\Omega} = \boldsymbol{I}$, it is straightforward to show that $\boldsymbol{\Omega}_{21} = \boldsymbol{0}$ and that $\boldsymbol{\Omega}_{11}$ and $\boldsymbol{\Omega}_{22}$ are unitary. Thus, for the left singular vectors of the FB signal matrix and the filtered FB signal matrix we may write

$$\boldsymbol{U}_{\boldsymbol{ZH}} = \boldsymbol{U}_{\boldsymbol{Z}}\begin{bmatrix} \boldsymbol{\Omega}_{11} & \\ & \boldsymbol{\Omega}_{22} \end{bmatrix}, \tag{5.33}$$

where $\boldsymbol{\Omega}_{11}$ and $\boldsymbol{\Omega}_{22}$ are unitary and $\boldsymbol{\Omega}_{11}$ is of size $r \times r$. This proves that the range of the FB signal matrix and the range of the filtered FB signal matrix are equal, $\mathcal{R}\big(\boldsymbol{Z}(k)\big) = \mathcal{R}\big(\boldsymbol{Z}(k)\boldsymbol{H}_{\boldsymbol{Z}}\big)$, and thus, the rotational invariance property is retained.

Filter matrix structure

Now, we proceed to discuss the specific structure of the filter matrix. For use in other subspace based parameter estimation techniques, filter matrices have been proposed which realize the discrete convolution with zero padding [43] or the circular convolution [1], [42] of the filter with each row of the signal matrix. For the discrete convolution with zero padding, the filter matrix can be written on the form

$$\boldsymbol{H}_T = \begin{bmatrix} h_\tau & \cdots & h_q & 0 & \cdots & 0 \\ \vdots & \ddots & & \ddots & \ddots & \vdots \\ h_1 & & \ddots & & \ddots & 0 \\ 0 & \ddots & & \ddots & & h_q \\ \vdots & \ddots & \ddots & & \ddots & \vdots \\ 0 & \dots & 0 & h_1 & \dots & h_\tau \end{bmatrix}. \tag{5.34}$$

We note that this is a Toeplitz[2] structured matrix, and in the following, we refer to $\boldsymbol{H}_T$ as the toeplitz filter matrix.

[2] A matrix, $\boldsymbol{A}$, is said to be Toeplitz structured when it has constant diagonals. In other words, $a_{i,j}$ depends only on $i - j$.

For the circular convolution, the filter matrix has the form

$$\boldsymbol{H}_C = \begin{bmatrix} h_\tau & \cdots & h_q & 0 & \cdots & 0 & h_1 & \cdots & h_{\tau-1} \\ \vdots & \ddots & & \ddots & \ddots & & \ddots & \ddots & \vdots \\ h_1 & & \ddots & & \ddots & \ddots & & \ddots & h_1 \\ 0 & \ddots & & \ddots & & \ddots & \ddots & & 0 \\ \vdots & \ddots & \ddots & & \ddots & & \ddots & \ddots & \vdots \\ 0 & & \ddots & \ddots & & \ddots & & \ddots & 0 \\ h_q & \ddots & & \ddots & \ddots & & \ddots & & h_q \\ \vdots & \ddots & \ddots & & \ddots & \ddots & & \ddots & \vdots \\ h_{\tau+1} & \cdots & h_q & 0 & \cdots & 0 & h_1 & \cdots & h_\tau \end{bmatrix} \tag{5.35}$$

We note that this, in addition to being Toeplitz, is a circulant[3] matrix, and in the following, we refer to $\boldsymbol{H}_C$ as the circulant filter matrix. In the two filter matrix expressions, $\tau = (q+1)/2$. Note, that if the filter coefficients are symmetric around h_τ, the filter matrices are symmetric in addition to being toeplitz and are thus centro hermitian (or centro symmetric since $\boldsymbol{H}$ is real).

Using the circulant filter matrix corresponds to circularly convolving each row of the signal matrix with the filter. This, in turn, correponds to a point wise multiplication in the frequency domain. It has been argued that this leads a better performance than the toeplitz filter matrix [1]. However other studies show that the toeplitz filter matrix leads to better results because the zero padding reduces the filter end–effects[43]. The difference in results might be attributed to the different types of signals, which where analysed in these studies. In [1] the signals of interest were audio signals, while in [43] the signals were nuclear magnetic resonanse (NMR) recordings. The difference in other factors such as signal data lengths and prefilter order, can also have an influence on the results obtained. In the sequel we study both the toeplitz and the circulant filter matrix.

Example 5.2. *Consider the signal and filter from (Example 5.1). Now, we arrange the signal in a matrix $\boldsymbol{X}$ of size 25×26. Two experiments are now conducted, in which the filter coefficients are arranged in a filter matrix $\boldsymbol{H}$ of size 26×26 according to (5.34) and (5.35) respectively. Then, we use the Unitary ESPRIT algorithm as described in section 3.2.7 on the matrix product $\boldsymbol{XH}$. Doing 100 Monte Carlo runs we get the following estimates of the mean value and variance of the estimated frequency, ω_2*

Circulant filter matrix			*Toeplitz filter matrix*		
$\hat{m}_{\omega_1}$	$=$	0.300π	$\hat{m}_{\omega_1}$	$=$	0.300π
$\hat{\sigma}_{\omega_1}$	$=$	$2.88 \cdot 10^{-6}$	$\hat{\sigma}_{\omega_1}$	$=$	$2.93 \cdot 10^{-6}$.

As in (Example 5.1), the frequency is estimated with no significant bias and with a comparable variance.

□

[3] An $n \times n$ matrix, $\boldsymbol{A}$, is said to be circulant when each column is equal to the previous column rotated downward by one element. In other words, $a_{i,j}$ depends only on $(i-j)$ modulo n.

Example 5.3. *Let us repeat the two experiments from (Example 5.1) and (Example 5.2). This time we vary the signal to noise ratio (SNR), and estimate the variance of the frequency estimates obtained using the three methods. The results are shown in figure 5.3 and 5.4. As we see in figure 5.4, the variance on the estimates is lowest for the signal vector prefiltering method and highest for the signal matrix prefiltering with toeplitz filter matrix at SNRs below approximatly* -10 *dB.*

□

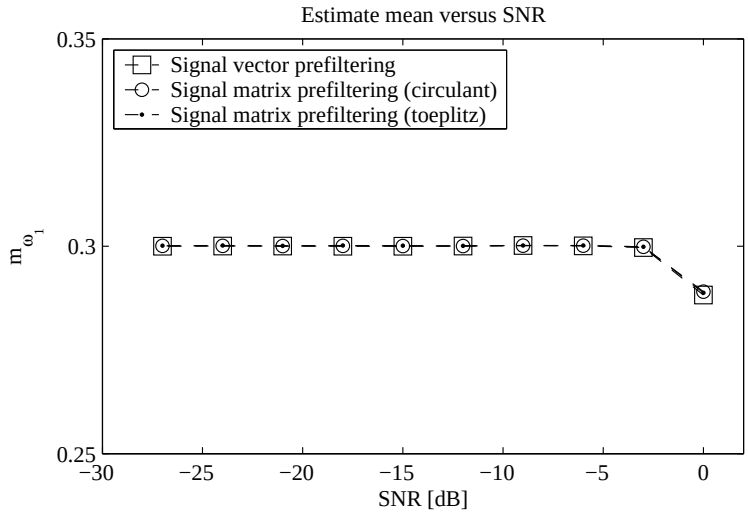

Figure 5.3: *Example 5.3: Estimates of the mean of frequency estimates using signal vector prefiltering and signal matrix prefiltering with circulant and toeplitz filter matrices.*

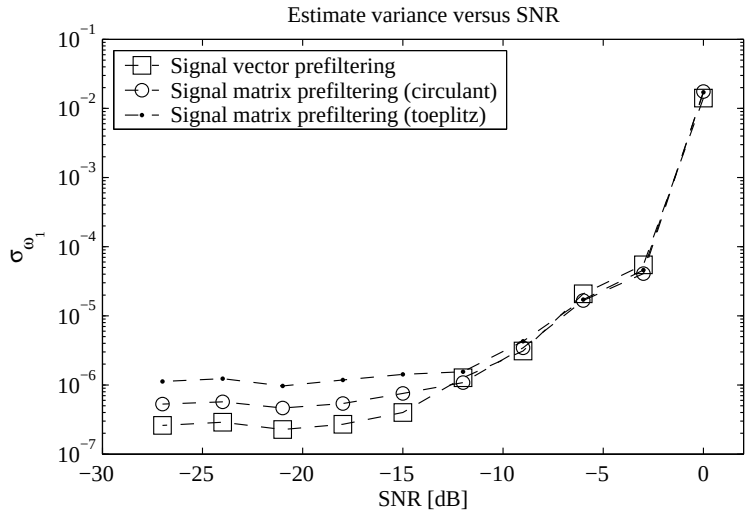

Figure 5.4: *Example 5.3: Estimates of the variance of frequency estimates using signal vector prefiltering and signal matrix prefiltering with circulant and toeplitz filter matrices.*

Summary of the signal prefiltering algorithm

The prefiltering of the signal matrix in the Perceptual Unitary ESPRIT can be summarized in the following steps

1. Compute the FIR filter coefficients $h_q(k)$ for the perceptually weighted filter using e.g. the method described in section 5.1.2.
2. Do one of the following:
   (a) Filter the signal by $h_q(k)$, discard the first q samples, and anrrange the remaining samples in the FB signal matrix.
   (b) Arrange the signal in a Hankel structured signal matrix and post multiply by either the toeplitz or the circulant filter matrix. Then, form the FB signal matrix.
3. Use the FB signal matrix in the Unitary ESPRIT algorithm.

5.3 Selection of perceptually relevant amplitudes and phases

When the signal poles have been estimated using the Perceptual Unitary ESPRIT algorithm, the complex amplitudes of each cisoid can be found by solving the following weighted least squares problem [1]

$$\hat{\boldsymbol{s}} = \arg\min_{\boldsymbol{s}} ||\boldsymbol{HW}\left(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s}\right)||_2^2, \tag{5.36}$$

where $\boldsymbol{H}$ is the perceptual weighting filter matrix given by either (5.34) or (5.35), $\boldsymbol{W} = diag\big(w(k)\big)$ is a diagonal matrix of the analysis window $w(k)$, and the Vandermonde matrix $\boldsymbol{A}$ is given by (2.5). This minimization problem has the following closed form solution [17]

$$\hat{\boldsymbol{s}} = (\boldsymbol{HWA})^{\dagger}\boldsymbol{HWx} \tag{5.37}$$

where $(\boldsymbol{HWA})^{\dagger}$ is the pseudoinverse of $(\boldsymbol{HWA})$.

5.4 Summary of the Perceptual Unitary ESPRIT algorithm

We are now able to sum up the steps for the Perceptual Unitary ESPRIT algorithm.

Analysis of signal segment

The analysis of a signal segment for a sinusoidal model using the Perceptual Unitary ESPRIT algorithm can then be summarized in the following steps

1. Compute the global masking threshold, as described in section 4.2.
2. Compute the coefficients of the perceptual weighting filter e.g. as described in section 5.1.2.
3. Filter the signal as described in section 5.2.1.
4. Compute the signal poles using the Unitary ESPRIT algorithm as described in section 3.2.7
5. Compute the complex amplitudes using (5.37).

Synthesis of signal segment

Synthesis of a modeled signal using the overlap-and-add method can be summarized in the following steps where l is the signal segment index and i indexes the individual sinusoids

1. For each signal pole $\hat{z}_{l,i}$ and complex amplitude $\hat{s}_{l,i}$ in each signal segment l, compute a time sequence for the segment of length N

$$\hat{x}_{l,i}(k) = \hat{s}_{l,i}\hat{z}_{l,i}^{k}. \tag{5.38}$$

2. Sum up all the time sequences in each segment and multiply by the synthesis window $w(k)$

$$\hat{x}_{l}(k) = w(k)\sum_{i}\hat{x}_{l,i}(k). \tag{5.39}$$

3. Reconstruct the signal with overlap and add.

$$\hat{x}(k) = \sum_{l}\hat{x}_{l}(k - lp). \tag{5.40}$$

5.5 Summary

In this chapter we have described how a psychoacoustic model can be incorporated in the Unitary ESPRIT algorithm by means of prefiltering the signal prior to applying the Unitary ESPRIT algorithm. We have identified three different techniques for filtering the signal matrix prior to parameter estimation: signal vector prefiltering and signal matrix prefitering with a toeplitz or circulant filter matrix. Finally, we have summarized the final Perceptual Unitary ESPRIT algorithm on which we will conduct a series of experiments in the following chapter.

Chapter 6

Experimental results

"You cannot acquire experience by making experiments. You cannot create experience. You must undergo it."

Albert Camus (1913 – 1960)

In this chapter: *We perform a series of experiments with the proposed Perceptual Unitary ESPRIT algorithm for a wide range of deterministic and real speech and audio signals. To examine the effects of the psychoacoustic model, we compare the Perceptual Unitary ESPRIT algorithm with the Unitary ESPRIT algorithm. Then, we relate the proposed algorithm to the P-ESM algorithm introduced by Jensen et al. [1].*

6.1 Experiments

In the preceeding chapters we have presented the theoretical framework for a novel algorithm: Perceptual Unitary ESPRIT. Now, through a series of objective and subjective experiments, we seek to evaluate the performance of the proposed algorithm under various conditions. We introduce a perceptually weighted signal-to-noise ratio, which we use to compare the Perceptual Unitary ESPRIT algorithm with P-ESM.

Comparisons between Perceptual Unitary ESPRIT and Unitary ESPRIT: We examine how the inclusion of a psychoacoustic model in the Unitary ESPRIT algorithm influences the parameter estimation. Through three case studies, we show how the Perceptual Unitary ESPRIT algorithm finds the most perceptually relevant sinusoidal signal components as opposed to the Unitary ESPRIT algorithm which finds the most powerful signal components. Then we examine how the two algorithms choose frequencies over time for two different audio signals. Finally, we study how often different frequencies are chosen for the two algorithms.

Perceptual Unitary ESPRIT: We examine how the signal matrix height-to-width ratio affects the parameter estimation in the Perceptual Unitary ESPRIT algorithm. For the three different proposed signal prefiltering methods, we study how the choice of model order relates to the perceptual signal-to-noise ratio for different types of signal segments.

Comparisons between Perceptual Unitary ESPRIT and P-ESM: We show that an increased estimation accuracy is achieved the by using Unitary ESPRIT as opposed to HTLS which is the subspace based parameter estimation technique used in P-ESM. Then we compare the Perceptual Unitary ESPRIT algorithm with P-ESM both for stationary signal segments and for transient segments.

6.1.1 Test signals

All the test signals used are sampled at 44.1 kHz and quantized in 16 bit resolution, corresponding to CD-audio quality. The signals are imported into MATLAB and processed in floating point precision. For an overview of the different signals see appendix D.

6.1.2 Perceptual signal-to-noise ratio

In order to objectively be able to compare the perceptual quality of a signals, we define the following perceptually weighted signal-to-noise ratio for a signal vector $\boldsymbol{x}$ as the squared ℓ_2 norm of the windowed and prefiltered signal vector divided by the squared ℓ_2 norm of the windowed and prefiltered signal distortion, $\boldsymbol{\Delta x} = \boldsymbol{x} - \tilde{\boldsymbol{x}}$

$$PSNR = 10 \log_{10} \frac{\big|\big|\boldsymbol{HWx}\big|\big|_2^2}{\big|\big|\boldsymbol{HW\Delta x}\big|\big|_2^2}, \tag{6.1}$$

where $\boldsymbol{W} = diag\big(w(k)\big)$ is a diagonal matrix which defines the signal window and $\boldsymbol{H}$ is the perceptual weighting filter matrix. Notice, that the denominator of this expression corresponds to the perceptual distortion measure defined in section 5.1.1.

6.2 Comparisons between Perceptual Unitary ESPRIT and Unitary ESPRIT

The Perceptual Unitary ESPRIT algorithm aims at finding the most perceptually relevant signal parameters in a sinusoidal signal model as opposed to Unitary ESPRIT which models the most powerful signal components. Through a series of experiments, we investigate the effects of including a psychoacoustic model in the Unitary ESPRIT algorithm.

This page has been left blank intentionally.

6.2.1 Three sinusoids

We wish to show how the inclusion of the psychoacoustic model in the Unitary ESPRIT algorithm affects the parameter estimation for a signal consisting of a sum of sinusoids.

Test signal

We generate a synthetic signal segment of length 1024 at a sample rate of 44.1 kHz, consisting of three sinusoids, with the frequencies $f_1 = 1.2$ kHz, $f_2 = 1.4$ kHz, and $f_3 = 20$ kHz. The amplitudes are chosen to be $a_1 = 0.03$, $a_2 = 0.032$, and $a_3 = 0.1$, and the phases are set to zero. A small amount of white gaussian noise with a standard deviation of $\sigma_n = 1 \cdot 10^{-4}$ is added to the signal.

Procedure

Using the Perceptual Unitary ESPRIT and the Unitary ESPRIT algorithms, we estimate the signal parameters for the test signal block, assuming a signal order of $d = \{2, 4, 6\}$.

Results

The power spectrum, normalized to dB sound pressure level, is shown in figure 6.1 (a and b) as a solid line. The frequency dependent perceptual weighting is shown in (a through h) as a dashed line. The perceptual weighting curve is not the inverse of the global masking threshold from the psychoacoustic model, but rather the frequency response of the weigting filter derived from the perceptual masking curve found using the approach described in chapter 5.

Figure 6.1 is divided into two columns where the first column shows the spectrum of the input signal (a) and the spectra of the estimated signals (c), (e), and (g) of the Perceptual Unitary ESPRIT algorithm. The second column shows the same original signal spectrum in (b) while the reconstructed signal spectra of the Unitary ESPRIT is shown in (d), (f), and (h). For the second row (c) and (d) the model order, d, is set to 2 so that only one sinusoid is modeled. For the other rows the model order is 4 and 6 respectively.

The Unitary ESPRIT algorithm (the right column), for a model order $d = 2$ detects the most powerful signal component, namely the sinusoid at 20 kHz. Next, the second most powerful sinusoid at 1.4 kHz is found, an finally, the third most powerful sinusoid at 1.2 kHz is detected.

The Perceptual Unitary ESPRIT algorithm (the left colum), first identifies the perceptually most important sinusoid at 1.2 kHz. This is the most perceptually relevant sinusoidal component, since the component at 1.4 kHz is partially masked by this, and the component at 20 kHz is below the absolute hearing threshold. Next, the second most perceptually relevant sinusoid, namely that at 1.4 kHz is detected. Finally, with a model order of 6, all three sinusoids are estimated.

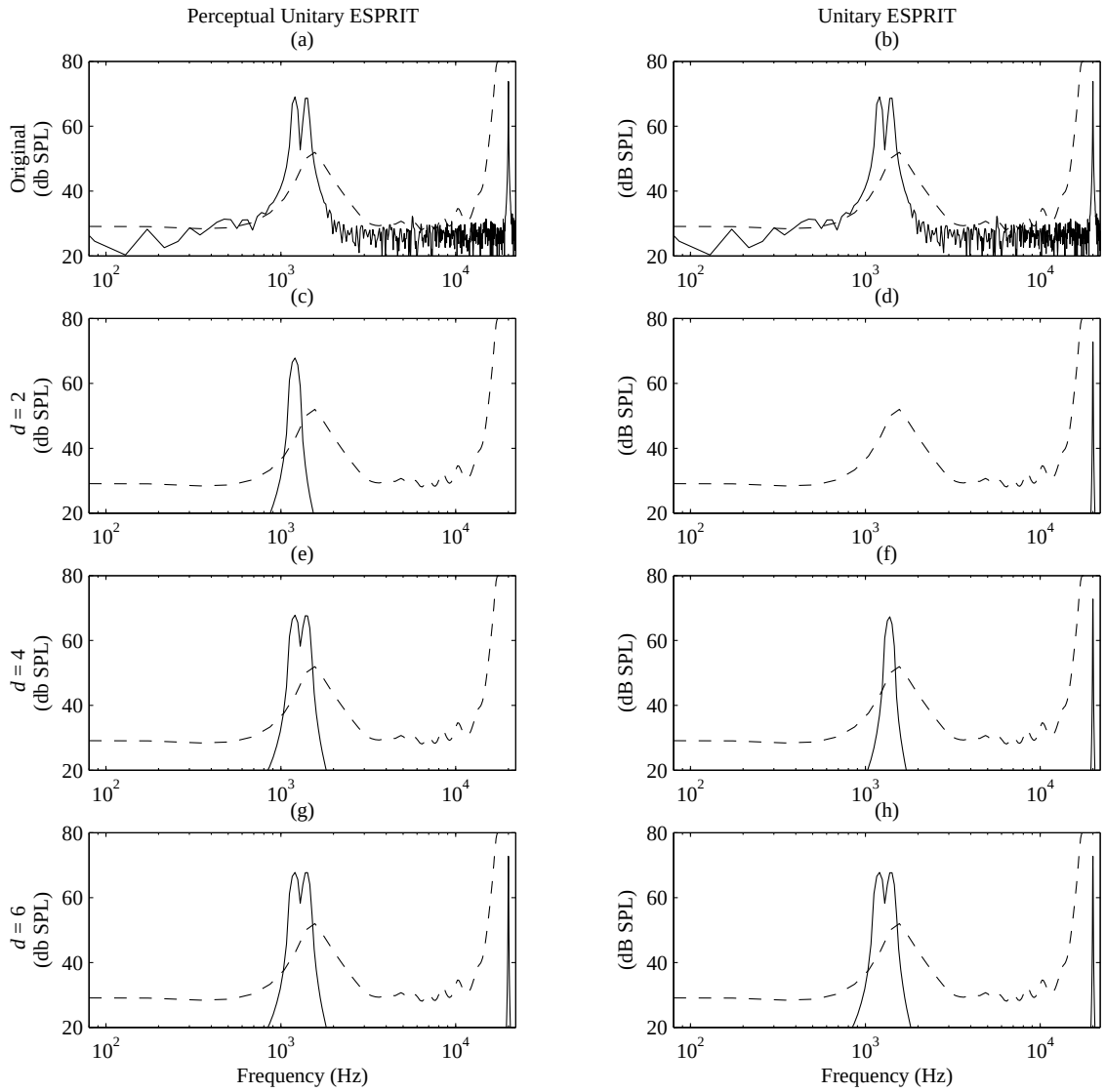

Figure 6.1: *Modeling of three sinusoids for different model orders, using Perceptual Unitary ESPRIT (left column) and Unitary ESPRIT (right column). (a) and (b) shows the power spectrum of the synthesized signal (solid) along with the inverse frequency response of the perceptual filter (dashed). (c) and (d) shows the modeled signal for* $d = 2$*. (e) and (f) modeling with* $d = 4$*. (g) and (h) modeling with* $d = 6$*.*

6.2.2 Three frequency chirps

Tonal signals, such as voiced regions in speech signals or instrument sounds such as trumpets or violins, can be modeled as a sum of slowly varying harmonically related sinusoids. This experiment aims at showing how the inclusion of the psychoacoustic model in the Unitary ESPRIT algorithm affects the parameter estimation for these types of signals.

Test signal

We generate a synthetic signal segment of length 1024 samples at a sample rate of 44.1 kHz, consisting of three frequency chips with the initial frequencies, $f_1 = 2$ kHz, $f_2 = 2.5$ kHz, and $f_3 = 3$ kHz. To simulate the slowly varying pitch-change of real audio signals, the frequencies are increased by 5% during the segment, which causes a "spreading" in the frequency domain. The amplitudes are chosen to be $a_1 = 0.06$, $a_2 = 0.12$, and $a_3 = 0.03$. The phases are all set to zero. As in the previous experiment, a small amount of white gaussian noise with a standard deviation of $\sigma_n = 1 \cdot 10^{-4}$ is added to the signal.

Procedure

Using the Perceptual Unitary ESPRIT and the Unitary ESPRIT algorithms, we estimate the signal parameters for the test signal segment, assuming a signal order of $d = \{2, 4, 6\}$.

Results

The power spectrum of the original signal, and the inverse frequency response of the masking filter is shown in the first row of figure 6.2. The next rows shows how the two algorithms model the signal for model orders 2, 4, and 6 respectively.

The Unitary ESPRIT algorithm (the right column) models the most powerful chirp (d), when the model order is 2. With a model order of 4, the algorithm uses two sinusoids to model the most powerful chirp signal, and with a model order of 6, two of the chirps are modeled.

The Perceptual Unitary ESPRIT algorithm (the left column), with a model order of 2, also models the most powerful chirp. Increasing the model order to 4 and 6 respectively, the Perceptual Unitary ESPRIT algorithm uses one sinusoid to model each of the chirp signals.

This example shows how the Perceptual Unitary ESPRIT algorithm models the chirps as individual sinusoids — thus, the frequency variations of the chirp signals are considered less perceptually important. The Unitary ESPRIT, which seeks to model the most powerful signal components, aims at modeling the frequency variation of the most powerful chirp signal, before the second most powerful chirp signal is modeled.

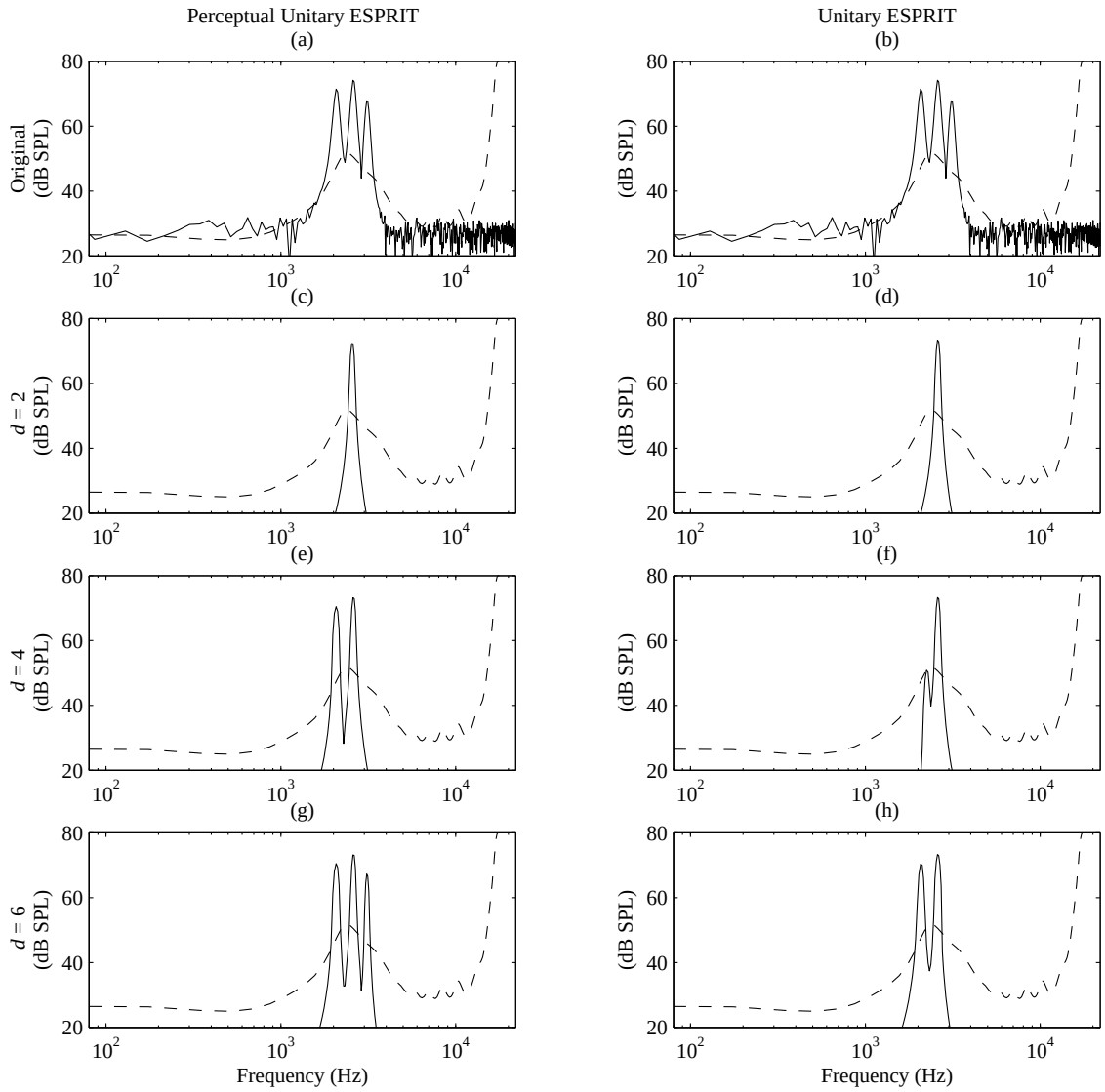

Figure 6.2: *Modeling of three sinusoids with a linear chirp, using Perceptual Unitary ESPRIT (left column) and Unitary ESPRIT (right column). (a) and (b) shows the power spectrum of the synthesized signal (solid) along with the inverse frequency response of the perceptual filter (dashed). (c) and (d) shows the modeled signal for* $d = 2$*. (e) and (f) modeling with* $d = 4$*. (g) and (h) modeling with* $d = 6$*.*

6.2.3 Noise-like signal segment

Noise-like signal segments are not well modeled by a sinusoidal model; however, we wish to examine how the inclusion of the perceptual model in the Unitary ESPRIT algorithm affects the signal parameter estimation for such a signal segment.

Test signal

A signal segment of lengh 1024 samples is extracted from a wave file containing male speech: "spme50_1_short". The signal segment is chosen from the "s" sound in the spoken word "distance". This signal possesses an noise-like quality which is especially dificult to model with the sinusoidal model.

Procedure

Using the Perceptual Unitary ESPRIT and the Unitary ESPRIT algorithms, we estimate the signal parameters for the test signal block, assuming a signal order of $d = \{2, 4, 16\}$.

Results

The power spectrum of the original signal, and the inverse frequency response of the masking filter is shown in the first row of figure 6.3. The next rows shows how the two algorithms models the signal for model orders 2, 4, and 16 respectively.

The Unitary ESPRIT algorithm (the right column) for model orders 2, 4, and 16, chooses the most powerful frequencies which are all in the range 3–8 kHz.

The Perceptual Unitary ESPRIT algorithm (the left column) mainly estimates the low frequency components in the range 100 Hz – 4 kHz, which by this algorithm are considered most perceptually relevant.

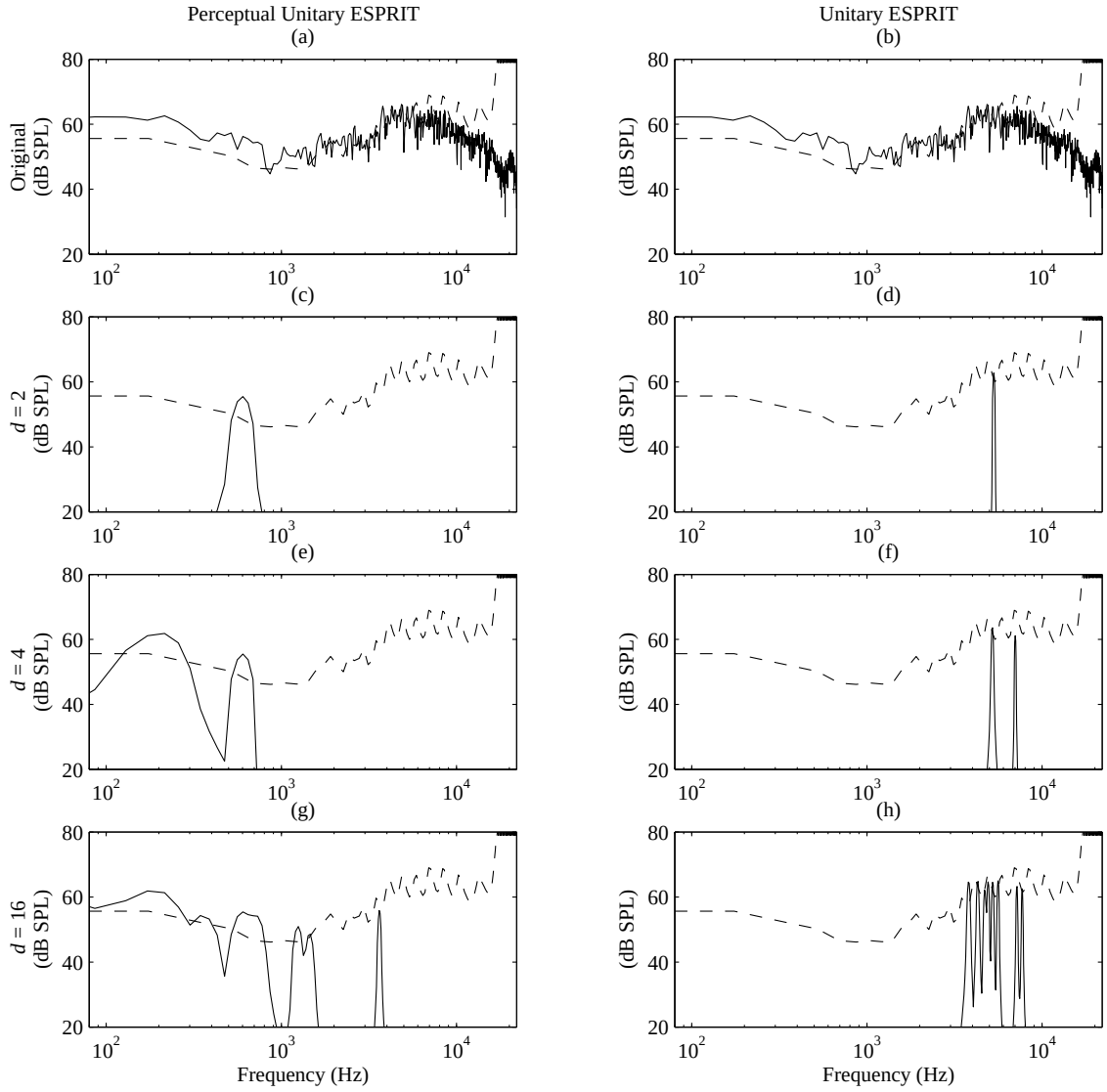

Figure 6.3: *Modeling of a noise-like segment from an unvoiced part of a monologue spoken by a male speaker, using Perceptual Unitary ESPRIT (left column) and Unitary ESPRIT (right column). (a) and (b) show the power spectrum of the origitnal signal (solid) along with the inverse frequency response of the perceptual filter (dashed). (c) and (d) shows the modeled signal for* $d = 2$*. (e) and (f) modeling with* $d = 4$*. (g) and (h) modeling with* $d = 16$*.*

6.2.4 Frequency distribution for a speech signal

From the previous experiments, it is evident that the inclusion of the psychoacoustic model in the Unitary ESPRIT algorithm significantly changes the parameter estimates. Now, we examine this further, by looking at the frequency estimates as a function of time for the Unitary ESPRIT and the Perceptual Unitary ESPRIT.

Test signal

The test signal is a 1.2 seconds sample at 44.1 kHz of a female speaker pronouncing the words: "To administer medicine" (spfe49_1_short).

Procedure

For consecutive 50% overlapping segments of 1024 samples, the signal parameters are estimated using Unitary ESPRIT and Perceptual Unitary ESPRIT using a model order of $d = 50$.

Results

The results of the parameter estimation is shown in figure 6.4. In (a) the waveform of the test signal is shown, (b) is the frequency estimates using the Perceptual Unitary ESPRIT, and (c) is the frequency estimates using Unitary ESPRIT. The individual estimated frequencies are represented by dots, and thus only the estimated frequencies and not their amplitudes and phases are shown.

We notice that the frequency estimates using the Perceptual Unitary ESPRIT algorithm are distributed more evenly than the Unitary ESPRIT. Especially for the unvoiced periods of the signal at around $t = 0.1$ s, $t = 0.6$ s, and $t = 1.1$ s, the Unitary ESPRIT algorithm (c) uses all its available sinusoids on the high energy regions at $6 - 12$ kHz. In those regions, the Perceptual Unitary ESPRIT algorithm (b) has a better modeling of the lower frequency bands. It can also be seen that for the voiced periods both of the algorithms are able to model the harmonics of the pitch frequency of the voice, which can be seen as a series of horizontal lines of dots where the first line starts at around 200 Hz at e.g. $t = 0.3$ s. In the voiced regions, the Perceptual Unitary ESPRIT is noted to include more high frequency components than the Unitary ESPRIT.

Informal listenting tests confirm these results. When the speech signal is modeled by the Perceptual Unitary ESPRIT and reconstructed using OLA, it is perceived to have a wider bandwidth in the voiced regions than for the Unitary ESPRIT algorithm. However, the unvoiced regions sound less "crisp" when using the Perceptual Unitary ESPRIT algorithm.

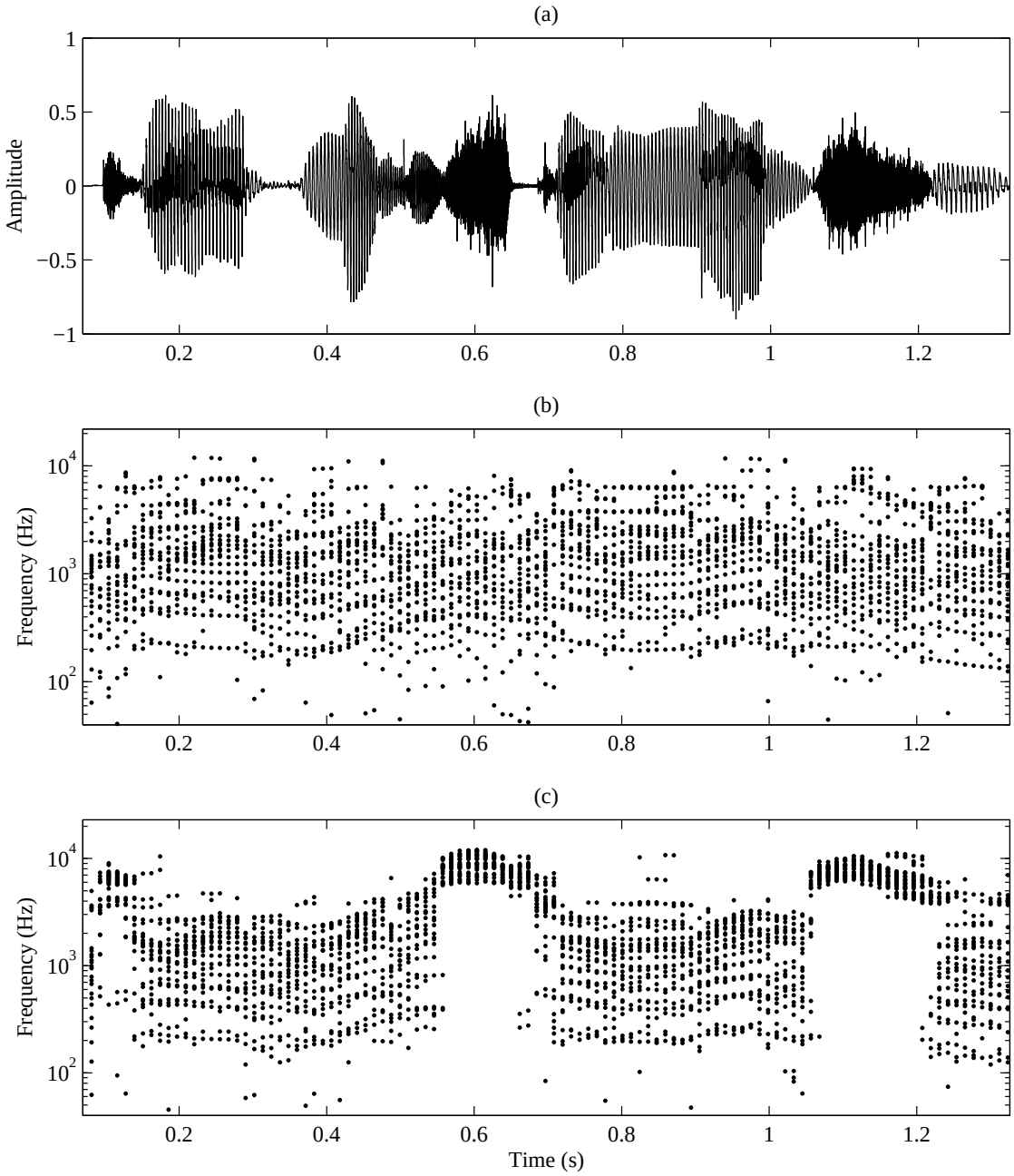

Figure 6.4: *Female speaker pronouncing "To administer medicine" from the wave-file "spfe49_1_short". (a) shows the wave form of the signal. (b) shows the distribution of frequency estimates over time using the Perceptual Unitary ESPRIT algorithm. (c) shows the distribution of frequency estimates over time using the Unitary ESPRIT algorithm.*

6.2.5 Frequency distribution for an audio signal

Similar to the previous experiment, we here examine the frequency estimates as a function of time for the Unitary ESPRIT and the Perceptual Unitary ESPRIT, this time with a tonal audio test signal: a male bass singer.

Test signal

The test signal is a 1.2 seconds sample at 44.1 kHz of a male bass singer (bass47_1_short).

Procedure

For consecutive 50% overlapping segments of 1024 samples, the signal parameters are estimated using Unitary ESPRIT and Perceptual Unitary ESPRIT using a model order of $d = 50$.

Results

The results of the parameter estimation is shown in figure 6.5. In (a) the waveform of the test signal is shown, (b) is the frequency estimates using the Perceptual Unitary ESPRIT, and (c) is the frequency estimates using Unitary ESPRIT.

Throughout the signal, the bass singer maintains a voiced tone. Both the Unitary ESPRIT and the Perceptual Unitary ESPRIT can be seen to model the pitch frequency and the harmonics well. However, the Unitary ESPRIT mainly models the low to mid range frequencies where most of the energy is contained. The Perceptual Unitary ESPRIT also models some of the high frequency components in the signal.

Informal listening tests confirm this: The audio signal of the bass singer modeled by the Perceptual Unitary ESPRIT algorithm and reconstructed using OLA is perceived to have a wider bandwidth and thus has a more natural sounding quality than when the signal is modeled by Unitary ESPRIT.

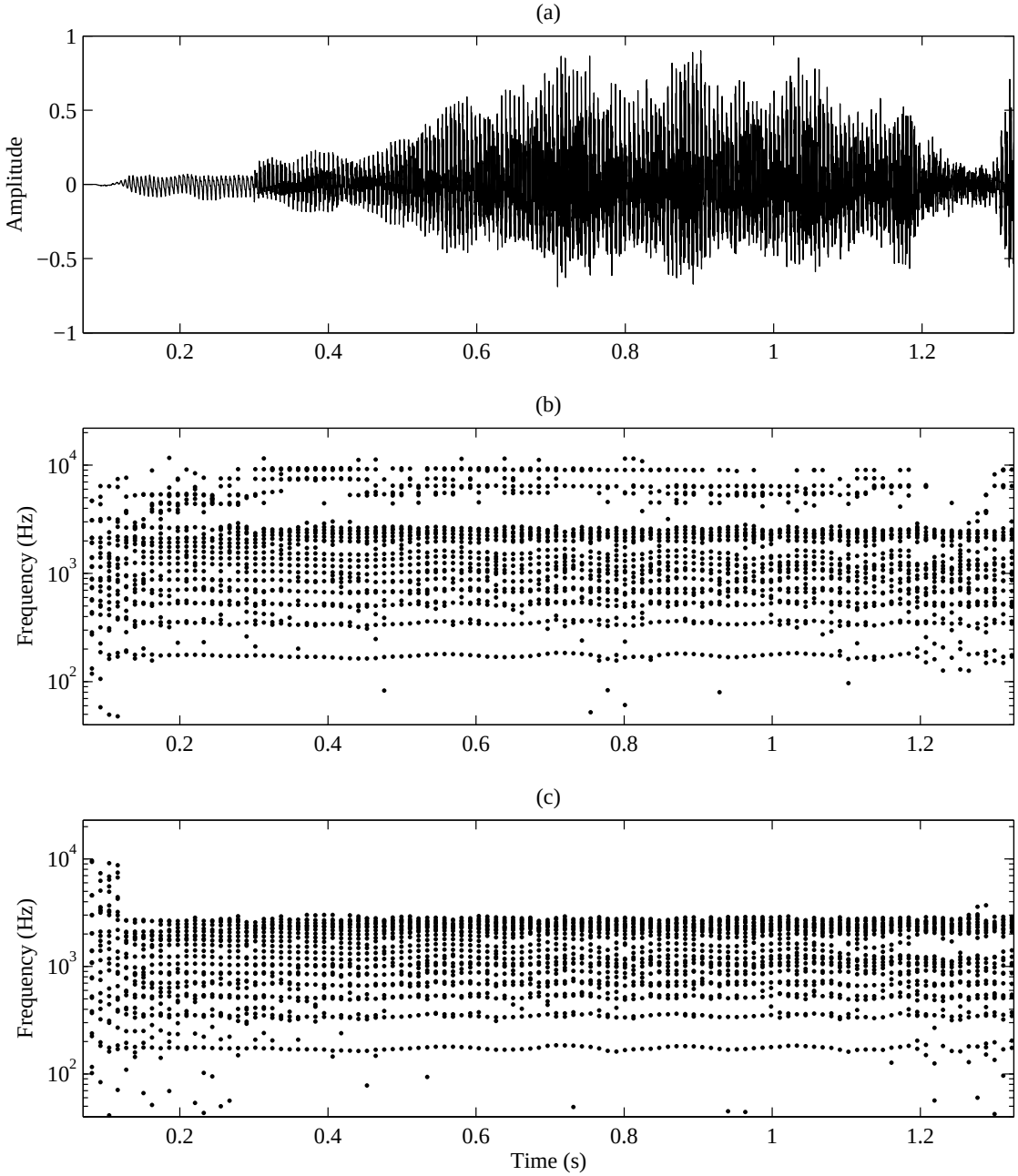

Figure 6.5: *Male bass singer, first 1.3 seconds of "bass47_1_short". (a) shows the wave form of the signal. (b) shows the distribution of frequency estimates over time using the Perceptual Unitary ESPRIT. (c) shows the distribution of frequency estimates over time using the Unitary ESPRIT.*

6.2.6 Frequency histograms

It has been argued that parameter estimation for sinusoidal modeling of audio signals can well be performed on a subsampled signal, since vast majority of the estimated frequencies lie in the low frequency range [1]. By resampling the signal prior to performing the parameter estimates the computational complexity of the parameter estimation algorithm is reduced significanly due to the reduction of the dimesions of the signal matrices employed. Here, we investigate the average distribution of estimated frequencies for the Perceptual Unitary ESPRIT algorithm, based on a wide range of audio signals.

Test signals

Eight different audio signals are used covering a wide variety of different types of audio signals: speech, song, tonal instruments, percussion, and band music (see appendix D)

Procedure

All the audio signals are modeled using Unitary ESPRIT and Perceptual Unitary ESPRIT for consecutive segments of length 1024 with 50% overlap and a model order of $d = 50$. A total of 3921 signal segments is used. Histograms of the frequencies estimated in all the signal segments are then created.

Results

In figure 6.6 the histograms of the frequency distribution for (a) the Perceptual Unitary ESPRIT algorithm and (b) the Unitary ESPRIT algorithm is shown. We notice that the frequency estimation of the Unitary ESPRIT algorithm is dominated by the often powerful lower frequencies. The high-frequency regions we saw modeled in the unvoiced regions of speech signals, were only characteristic for this signal type. For most other signals, the frequencies are primarily distributed in the region below 10 kHz.

The Perceptual Unitary ESPRIT algorithm has a different distribution. The lowest frequencies, below 1 kHz, are chosen more often than with the Unitary ESPRIT algorithm. Also it can be seen that some of the higher frequencies, are more frequently represented than with the Unitary ESPRIT algorithm.

It is, however, evident that for both algorithms the dominant part of the frequencies are in the region below 10 kHz. This corresponds to the observations made by Jensen et al. [1].

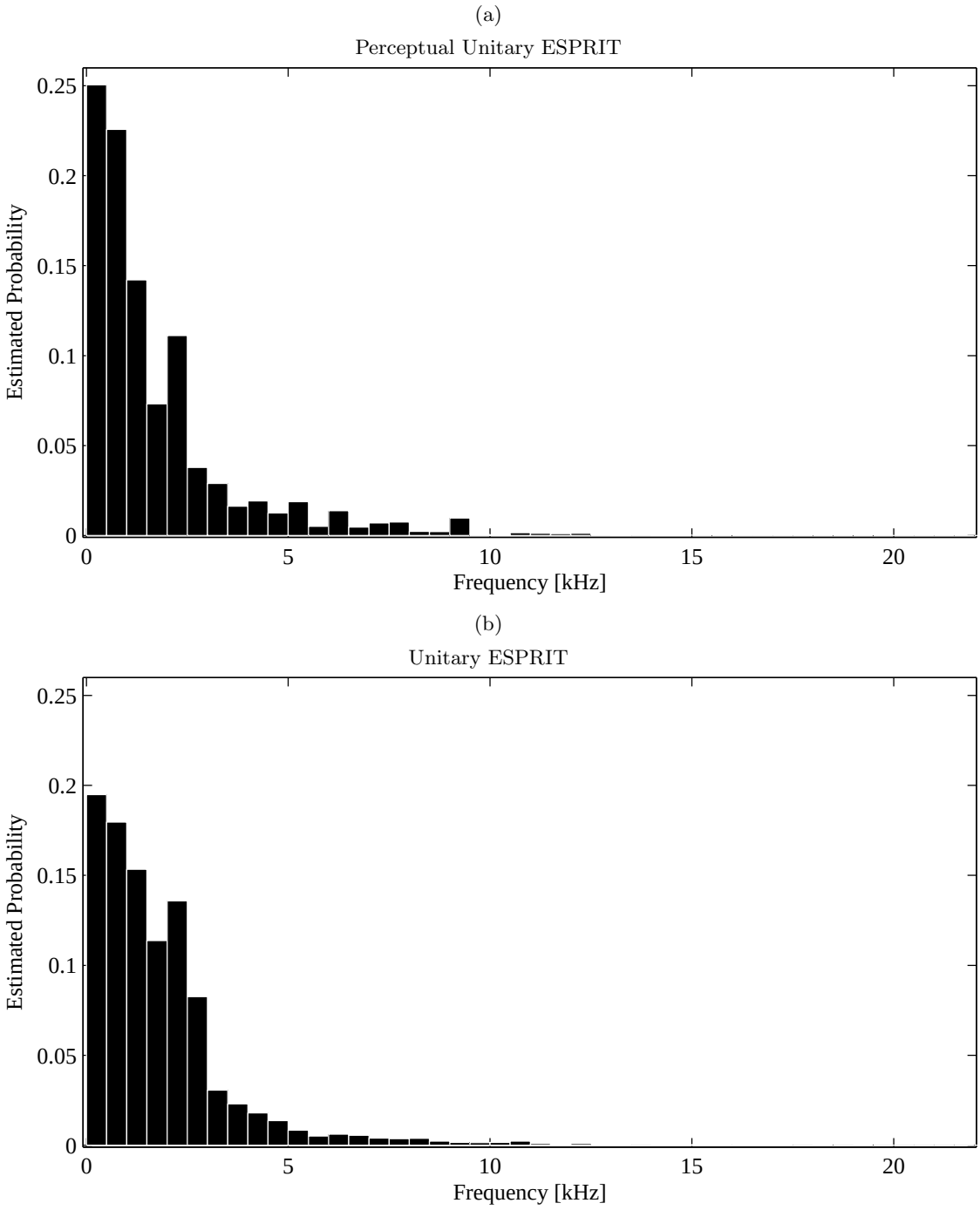

Figure 6.6: *Histograms of the distribution of the frequencies estimated by (a) the Perceptual Unitary ESPRIT algorithm and (b) the Unitary ESPRIT algorithm. The data is collected from the modeling of eight wave-files, and normalized to an estimated probability for a bin size of 500 Hz.*

6.3 Perceptual Unitary ESPRIT

In the following we investigate how the characteristics of the proposed Perceptual Unitary ESPRIT algorithm. We examine how the dimensions of the signal matrix should be chosen and we look into the differences between the three different proposed signal prefiltering techniques.

6.3.1 Height-to-width ratio of data matrix

One factor which influences the estimation accuracy of the Perceptual Unitary ESPRIT algorithm is the proper selection of the signal matrix heigth-to-width ratio. For the P-ESM algorithm, Jensen et al. found that the most optimum ratio, in a perceptual sense, between the height and width of the signal matrix was approximately 1/3 [1], i.e the best results were obtained with a "fat" signal matrix.

Test signals

Eight different audio signals are used covering a wide variety of different types of audio signals: speech, song, tonal instruments, percussion, and band music (see appendix D)

Procedure

A total of 16 segments of length 1024 samles are chosen by random — two from each of the test files. For different signal matrix dimensions $0.3 < m/N < 0.7$ the Perceptual Unitary ESPRIT algorithm using circulant matrix prefiltering is used to model each of the segments with a model order of $d = 50$. Then, the signal segments are reconstructed from the parameters and windowed by a Hann window. Finally, the PSNR is computed for each reconstructed signal segment, and the PSNR is averaged over the 16 signal segments for each m/N ratio.

Results

The result of the experiment is shown in figure 6.7. The PSNR is greatest for a m/N ratio around 0.55. We notice that since the width of the FB signal matrix used in the Perceptual Unitary ESPRIT is $2M$, the best results are obtained with a "fat" signal matrix as in the P-ESM algorithm.

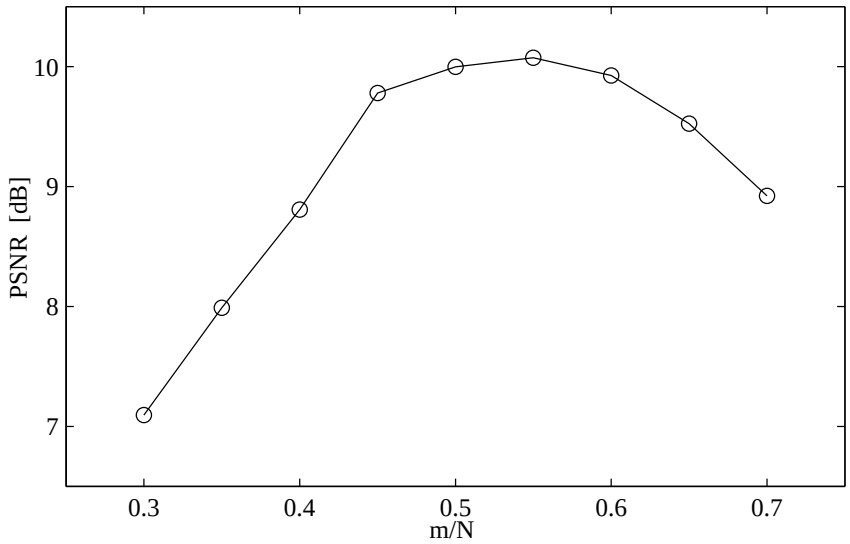

Figure 6.7: *The perceptual SNR as a function of the ratio between m and N using the Perceptual Unitary ESPRIT algorithm. Each data point is averaged over 16 signal segments chosen from 8 different wave-files.*

6.3.2 Type of prefiltering and model order

In the previous chapter we introduced three different methods in which the prefiltering of the signal matrix could be performed in the Perceptual Unitary ESPRIT algorithm: signal vector prefiltering and signal matrix prefiltering with a toeplitz or circulant filter matrix. Now, we examine how these different prefiltering methods affect the PSNR in the modeling of two different type of signal segments at a range of different model orders.

Test signals

Two different types of signal segments of length 1024 samples are selected: tonal segments from a voiced part of male speech (spme50_1_short), and noise-like segments from a musical piece consisting of cymbals and a choir (sicas13_orig). For each segment type, 13 consecutive segments are used.

Procedure

For each of the two different segment types the following is performed: The 13 consecutive signal segments are modeled using the Perceptual Unitary ESPRIT algorithm with model orders ranging from $d = 4$ to $d = 50$. This is done for each of the three different prefiltering methods. In addition to this, we also include results for the Unitary ESPRIT algorithm. The signal segments are reconstructed from the estimated parameters and windowed by a Hann window. Finally, the PSNR is computed for each reconstructed signal segment and the PSNR is averaged over the 13 consecutive segments.

Results

The results are shown in figure 6.8 for (a) the tonal segment and (b) the noise-like segment. We notice that the PSNR is generally higher for the tonal segment since a tonal segment is better modeled by the sinusoidal model. Naturally, the Perceptual Unitary ESPRIT algorithm provides a higher PSNR than the Unitary ESPRIT, since it is in fact designed to minimize the perceptual distortion. The results for the three different prefiltering methods are seen to be almost equal, increasing with the model order. Thus, we see that the three proposed prefiltering methods provide almost equal performance as measured by the PSNR.

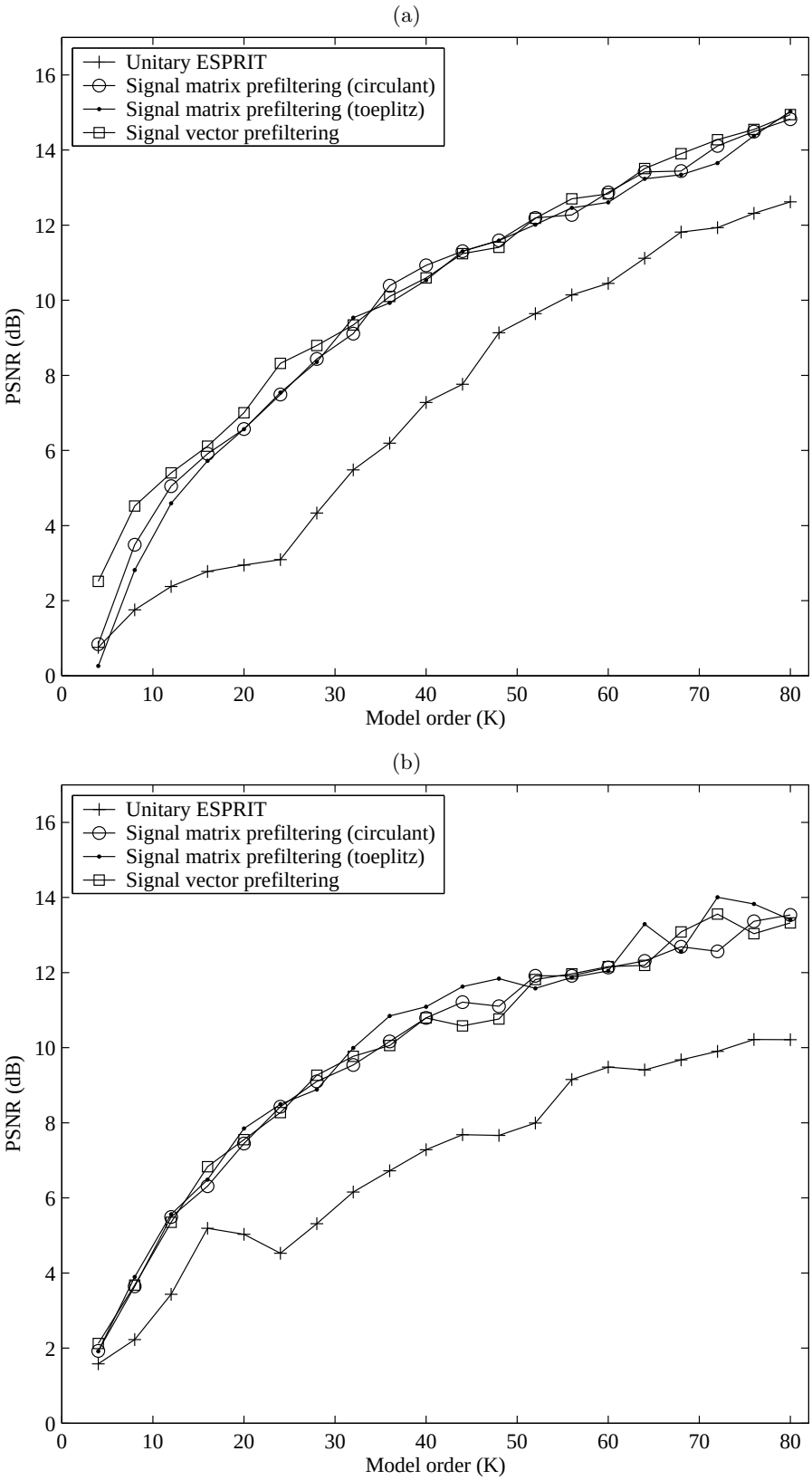

Figure 6.8: *The average PSNR for 13 consecutive signal segments from (a) a tonal part from male speech (spme50_1_short) and (b) a noise-like segment from a symphony orchestra (sicas13_orig). Each segment is modeled with four different methods to illustrate the influence of the perceptual filter, and how this was applied.*

6.4 Comparisons between Perceptual Unitary ESPRIT and P-ESM

Recently, Jensen et al. have proposed an algorithm for estimating perceptually relevant parameters in an exponentially damped sinusoidal model [1]. The method proposed by Jensen et al. is denoted P-ESM and is based on incorporating a psychoacoustic model in a subspace parameter estimation method known as HTLS.

The idea behind our development of the Perceptual Unitary ESPRIT algorithm is to overcome two drawbacks of the method proposed by Jensen et al., namely to achieve an increased estimation accuracy at an equal computational cost and to employ a signal model consisting of constant amplitude sinusoids as opposed to the exponentially damped sinusoids used in the P-ESM algorithm. Thus, the main differences between P-ESM and Perceptual Unitary ESPRIT is that the former provides more degrees of freedom due to the use of exponentially damped sinusoids, whereas the latter provides a more accurate parameter estimation.

In the following, we compare the Perceptual Unitary ESPRIT algorithm with the P-ESM algorithm. In our implementation of the P-ESM algorithm we use the same psychoacoustic model as in the Perceptual Unitary ESPRIT algorithm such that the two algorithms can be readily compared. In the following we show the main differences between the two algorithms through a few selected examples.

6.4.1 Parameter estimation accuracy

Since the Unitary ESPRIT algorithm exploits the signal data twice in the FB signal matrix, it provides an increased estimation accuracy over algorithms such as HTLS [13] — especially with regards to seperating closely spaced sinusoids. Here, we show that this is indeed true.

Test signal

We generate a signal segment of length 1024 samples consisting of two closely spaced frequencies: $\omega_1 = 0.1$ rad/sample and $\omega_2 = \omega_1 + \Delta\omega$ with unit amplitude. A small amount of white gaussian noise with a standard deviation of $\sigma_n = 0.1$ is added to the signal.

Procedure

For a wide range of $\Delta\omega$, the parameters of the synthesized signal segment are estimated using Unitary ESPRIT and HTLS assuming a model order of $d = 4$ corresponding to two sinusoids.

Reults

The results are shown in figure 6.9. We see that both HTLS and Unitary ESPRIT are able to seperate the two sinusoids correctly when they are properly spaced, however both algorithms break down when the two frequencies are very close. It is obvious from the figure, however, that the Unitary ESPRIT provides a better frequency seperation than the HTLS algorithm.

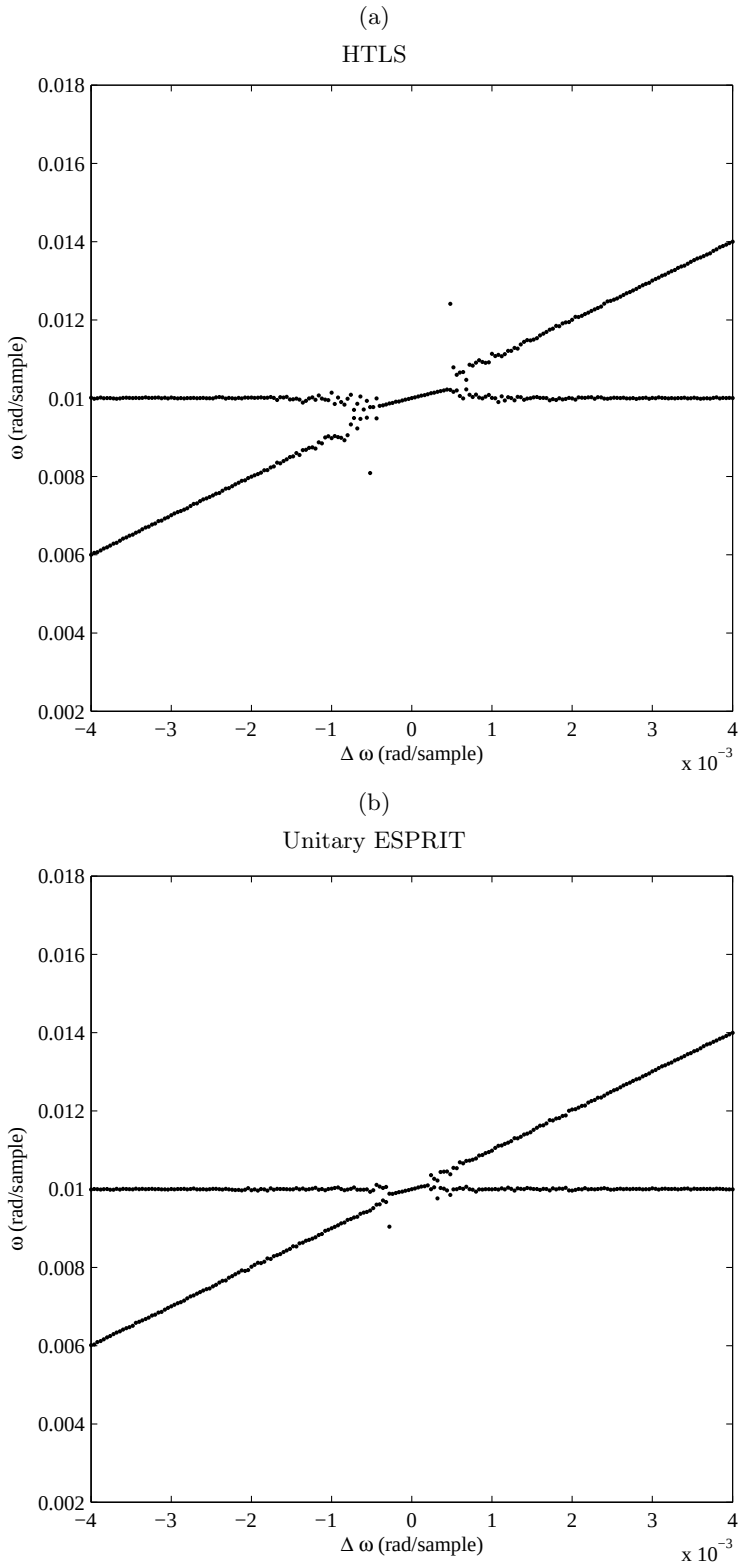

Figure 6.9: *The frequency estimation, using (a) HTLS and (b) Unitary ESPRIT for two closely spaced sinusoids, with varying frequency spacing. One sinusoid maintains a constant normalized frequency of* 0.01 *rad/sample. The normalized frequency of the other sinusoid is offset with by* $\Delta\omega$ *varying from -0.004 to 0.004 rad/sample.*

6.4.2 Pre-echo

A general problem with sinusoidal models is that when the underlying parameters of the signal change abruptly within one signal segment an artefact known as pre-echo can occur. Pre-echo occurs because the underlying assumption of constant parameters within each frame is violated. Here, we give an example of a signal which induces a pre-echo when modeled.

Test signal

The test signal is a 0.04 seconds sample at 44.1 kHz of the attack transient of a glockenspiel (gspi35_2_short).

Procedure

The test signal is divided in to 50% overlapping segments of length 1024 samples corresponding to approximatly 23 miliseconds. Each segment is modeled using HTLS, Unitary ESPRIT, P-ESM, and Perceptual Unitary ESPRIT with a model order of $d = 50$. Then, the signals are reconstructed using OLA.

Reults

The results are seen in figure 6.10, where the smearing of the sharp attack transient is evident for all four parameter estimation methods although to a larger extent in the perceptually based algorithms. Since the glockenspiel signal has a simple harmonic structure, the HTLS and Unitary ESPRIT algorithm can use the remaining sinusoids to model the transient. For the Perceptual Unitary ESPRIT algorithm and P-ESM, much of the high frequency contents of the signal which is responsible for the transient is removed, since it is not considered perceptually relevant. Therefore these algorithms cannot accurately model the signal since the transient is in effect removed by the perceptual weighting filter. Since both the perceptual model and the sinusoidal model itself is based on the assumption of constant signal parameters, it is obvious that a transient signal is not modeled well. For this reason, sinusoidal audio models are often used in hybrid audio coders which also include explicit models for transient and noise-like signals.

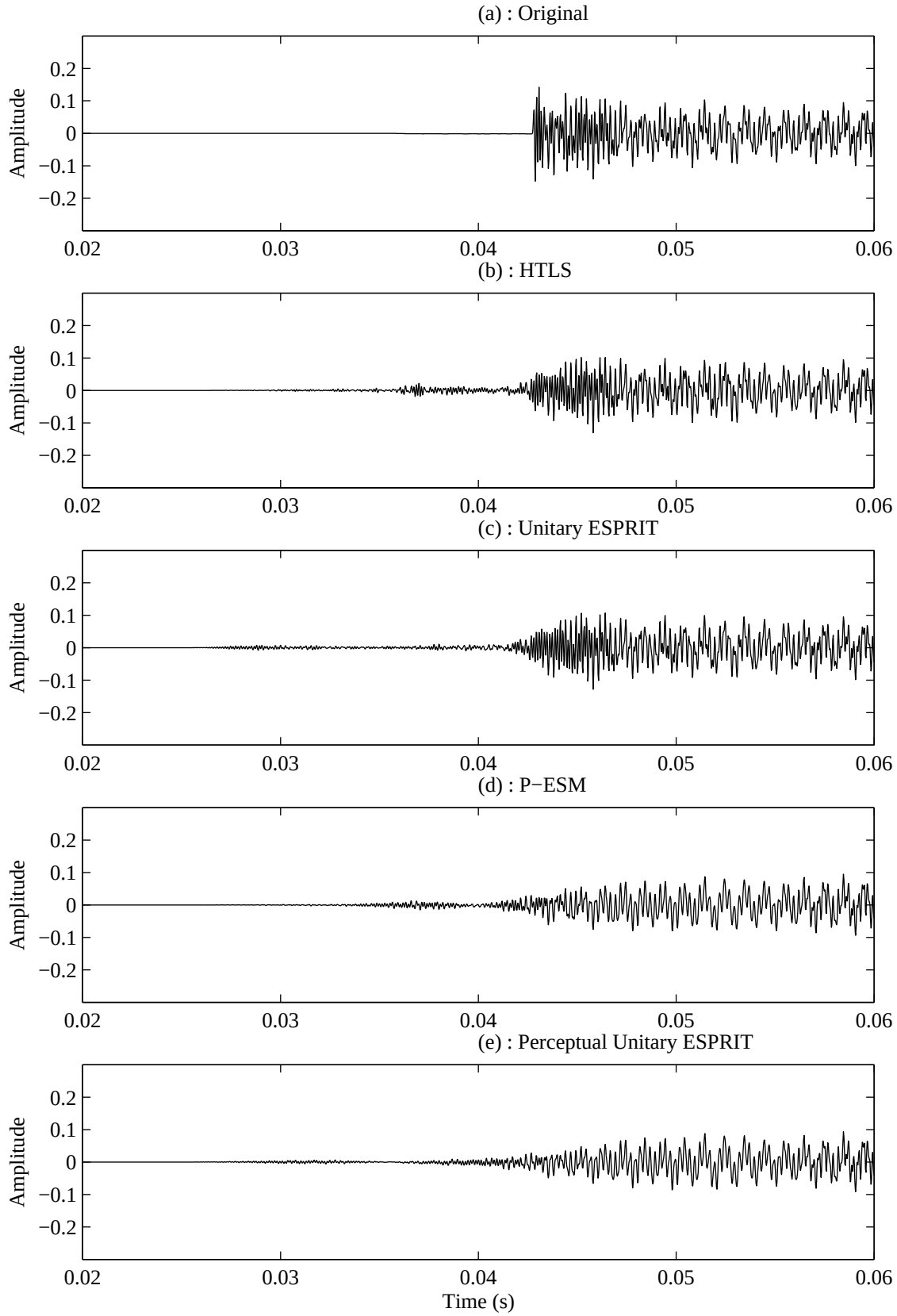

Figure 6.10: *Waveform of glockenspiel: In (a) the original signal is shown. The reconstructed signal is shown when modeled by (b) HTLS, (c) Unitary ESPRIT, (d) P-ESM and (e) Perceptual Unitary ESPRIT*

6.4.3 Transient and stationary segments

In order to get a better picture of how the two perceptual algorithms, Perceptual Unitary ESPRIT and P-ESM, differ with respect to modeling transient signal segments and stationary signal segments, the PSNR is computed for the glockenspiel signal used in section 6.4.2. This signal consists of a sharp attack transient followed by a stationary part. Here, we compare the two algorithms for different segment sizes by evaluating the PSNR for each segment.

Test signal

The test signal is a 0.04 seconds sample at 44.1 kHz of the attack transient of a glockenspiel (gspi35_2_short). In the quiet period the signal is set to zero.

Procedure

The signal is segmented into 50% overlapping blocks of length 1024, 512, and 256 samples respectively. The perceptual filter length used is 1/4 of the segment length, for all three segment lengths. Each signal segment is modeled by the Perceptual Unitary ESPRIT algorithm (with circulant matrix prefiltering) and by P-ESM using a model order of $d = 50$, and the perceptual signal-to-noise ratio is computed for each signal segment.

Reults

The results are shown in figure 6.11, where the segmentation is indicated by the greytoned Hann windows. For all three segment sizes, the PSNR is set to zero for all segments where the signal is zero. For segment sizes 1024 and 512, examining the transient part of the signal, we see that the P-ESM algorithm has the highest PSNR, whereas in the stationary part of the signal, the PSNR is almost equal for the two algorithms. This is due to the exponential window used in the P-ESM which enables some transient modeling to take place. For the 256 samples segments, the PSNR is highest for the Perceptual Unitary ESPRIT algorithm. When only 256 samples are used to estimate the signal parameters, the superior frequency resolution of the Perceptual Unitary ESPRIT algorithm has a significant influence on the results.

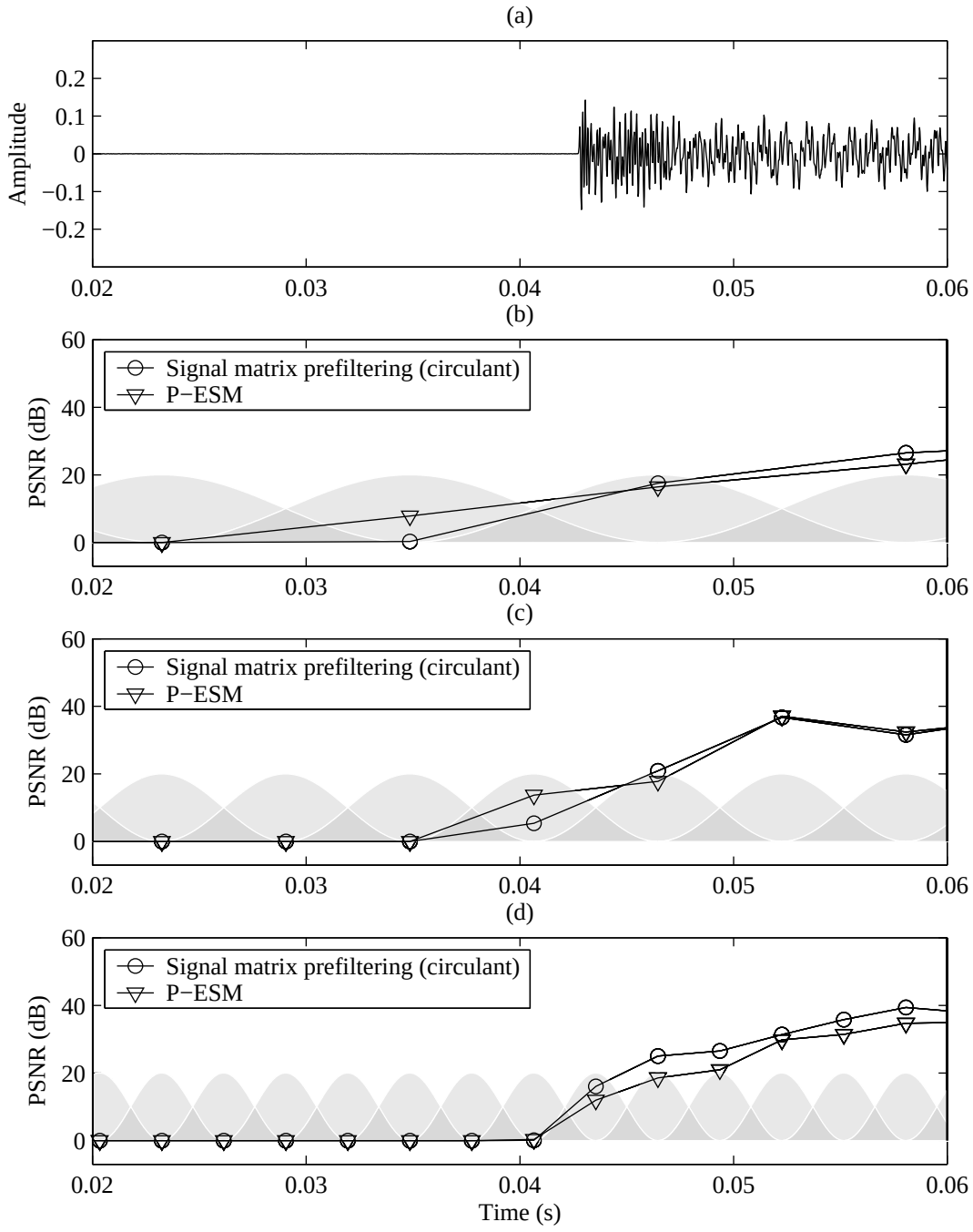

Figure 6.11: *Waveform of glockenspiel, (a). PSNR computed for each block, before reconstruction. (b) shows the PSNR for the two perceptual algorihms, with a segment size of* 1024 *samples. (c) shows the PSNR for segments* 512 *samples in length. (d) shows the PSNR for segments* 256 *samples in length. The analysis windows are represented by the graytoned Hann windows.*

6.4.4 Deterministic transient signal

Finally, we repeat the previous experiment, this time using a deterministic signal consisting of a single sinusoid.

Test signal

We generate a signal consisting of the sudden onset of a 1 kHz tone. The signal is generated at a sample rate of 44.1 kHz.

Procedure

The signal is segmented into 50% overlapping blocks of length 1024, 512, and 256 samples respectively. Each signal segment is modeled by the Perceptual Unitary ESPRIT algorithm (with circulant matrix prefiltering) and by P-ESM using a model order of $d = 2$. Then, the perceptual signal-to-noise ratio is computed for each signal segment.

Reults

The results are shown in figure 6.12. Concerning the segments in which the transient occurs, we notice a slightly higher PSNR for the P-ESM algorithm when using a large segment length, 1024, however for the short segment length, 256, the PSNR is slightly higher for the Perceptual Unitary ESPRIT algorithm. This corresponds well with the results obtained in the previous experiment for the attack transient of a glockenspiel. For the stationary part, consisting of just one sinusoid, there is no significant difference between the Perceptual Unitary ESPRIT algorithm and P-ESM. Thus, when only one sinusoid is to be estimated, there is not much difference in the accuracy of the two methods.

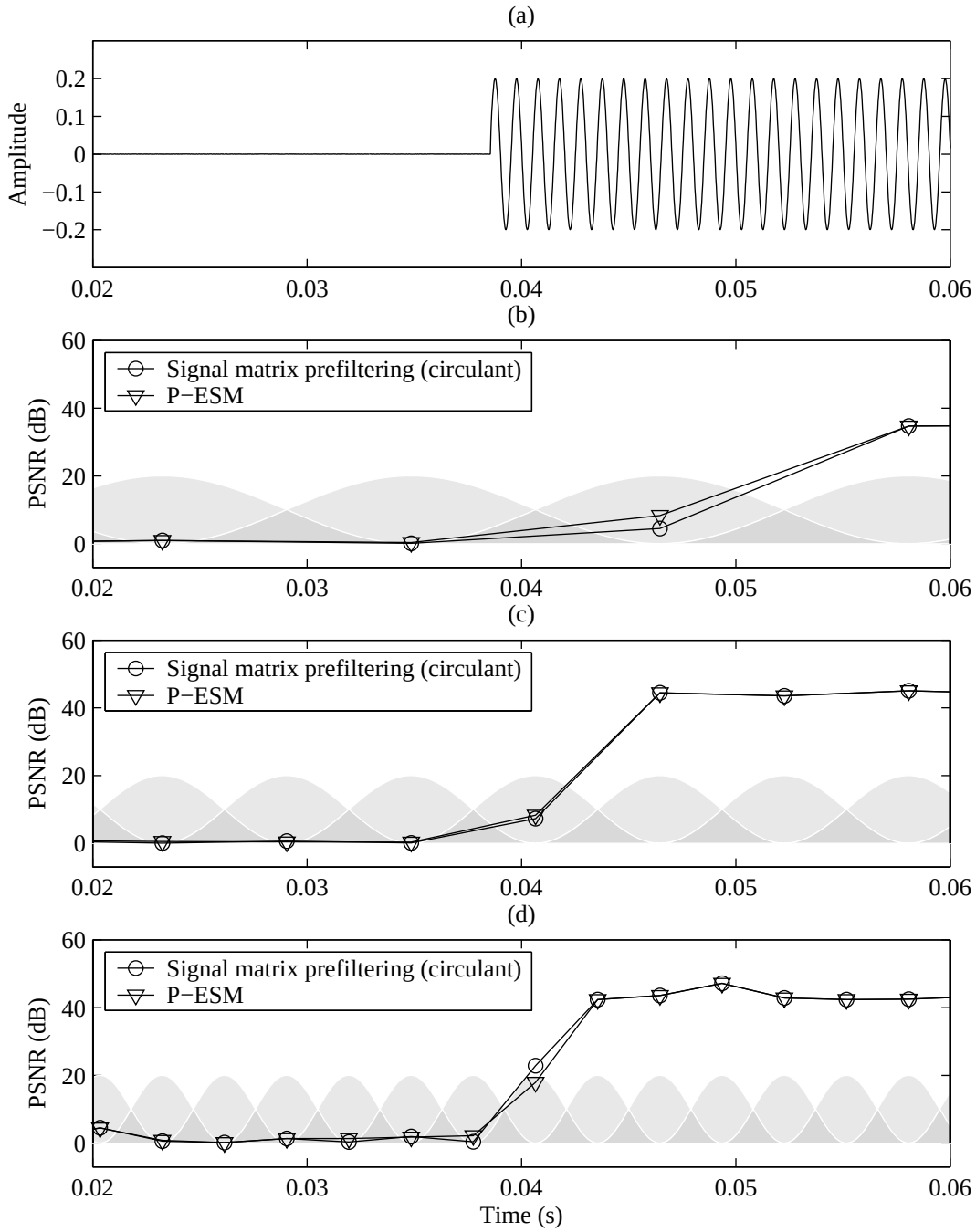

Figure 6.12: *Waveform of a sudden onset of a single sinusoid (a). PSNR computed for each block, before reconstruction. (b) shows the PSNR for the two perceptual algorihms, with a segment size of* 1024 *samples. (c) shows the PSNR for segments* 512 *samples in length. (d) shows the PSNR for segments* 256 *samples in length. The analysis windows are represented by the graytoned Hann windows.*

Chapter 7

Discussion and conclusions

> “*I may not have gone where I intended to go, but I think I have ended up where I intended to be.*”
>
> Douglas Adams (1952 – 2001)

In this chapter: *We summarize the results obtained in this work and discuss strengths and weaknesses of the proposed algorithm.*

In this dissertation we have proposed a novel algorithm for estimating perceptually relevant parameters for constant amplitude sinusoidal audio modeling. The proposed algorithm combines a psychoacoustic model with the Unitary ESPRIT algorithm.

The main goal of this work is to alleviate two drawbacks of a similar algorithm proposed by Jensen et al., namely the P-ESM algorithm, in which a subspace based parameter estimation method known as HTLS is combined with a psychoacoustic model for the purpose of estimating perceptually relevant parameters in an exponentially damped sinusoidal signal model. The two identified drawbacks are the following: 1) Exponentially damped sinusoids are used — in stationary signal segments, where an exponential damping factor is of little use, a constant amplitude sinusoidal model provides a more compact representation. 2) It is computationally complex — other subspace based parameter estimation methods provide better estimation accuracy at a comparable computational cost.

Concerning the modeling of signal segments which can rightly be considered stationary, there is not much difference between the results obtained by the Perceptual Unitary ESPRIT and the P-ESM for long signal segments of e.g. 1024 samples. For such signal segments, the Perceptual Unitary ESPRIT algorithm provides an equally good signal model using one parameter less than the P-ESM algorithm. For small signal segments of e.g. 256 samples the signal pole estimates will naturally be less precise because they are found from a relatively small amount of data. Here, the Perceptual Unitary ESPRIT provides superior estimates compared with the P-ESM algorithm, due to its increased frequency resolution.

With regards to modeling of transient signal segments neither the Perceptual Unitary ESPRIT nor the P-ESM show excellent results. This is due to the fact that the assumption of stationarity, on which both the signal model and the psychoacoustic model rely, is not valid. For long segment

sizes, such as segments of 1024 samples as used in some experiments in this work, the effects of an exponential damping factor is however quite evident. In transient signal segments, where the exponential damping is of great value in modeling the signal, the P-ESM clearly gives better results. However, the situation is reversed for short segments, such as segments of 256 samples. Here, the increased frequency estimation accuracy of the Perceptual Unitary ESPRIT algorithm is significant.

The computational cost of the algorithm proposed in this dissertation is comparable to that of the P-ESM algorithm; however, the data used for the parameter estimation is essentially doubled resulting in an increased estimation accuracy.

The objective of the work has been to combine a perceptual model with the Unitary ESPRIT algorithm. Three prefiltering schemes have been identified: the signal vector prefiltering and the signal matrix prefiltering with a toeplitz or circulant filter matrix. Each of these methods have been implemented and analysed. Experimental results, conducted on a wide range of signals have not shown any preference to one particular method. Therefore, further studies of the effects of the method of prefiltering should be conducted.

In conclusion, the Perceptual Unitary ESPRIT algorithm constitutes a robust, accurate, and efficient method for estimating perceptually relevant parameters for constant amplitude sinusoidal audio modeling.

Appendix A

ESPRIT: The covariance method

Here, we present an alternative approach to the ESPRIT algorithm, based on covariance matrices of the signal. Consider a signal consisting of d complex sinusoids in additive noise

$$x(k) = \sum_{i=1}^{d} s_i e^{j\omega_i} + n(k). \tag{A.1}$$

Let us define the vectors $\boldsymbol{x}(k)$, $\boldsymbol{y}(k)$, and $\boldsymbol{n}(k)$ [27]

$$\begin{aligned} \boldsymbol{x}(k) &= [x(k), \ldots, x(k+m-1)]^T & \text{(A.2)} \\ \boldsymbol{y}(k) &= [x(k+1), \ldots, x(k+m)]^T, & \text{(A.3)} \\ \boldsymbol{n}(k) &= [n(k), \ldots, n(k+m-1)]^T, & \text{(A.4)} \end{aligned}$$

where $m > d$. We may then write [27]

$$\begin{aligned} \boldsymbol{x}(k) &= \boldsymbol{A}\boldsymbol{s}(k) + \boldsymbol{n}(k), & \text{(A.5)} \\ \boldsymbol{y}(k) &= \boldsymbol{A}\boldsymbol{\Phi}\boldsymbol{s}(k) + \boldsymbol{n}(k+1), & \text{(A.6)} \end{aligned}$$

where $\boldsymbol{s} = [s_1, \ldots, s_d]^T$ is a vector of complex amplitudes, $\boldsymbol{\Phi} = diag(e^{j\omega_1}, \ldots, e^{j\omega_d})$ is a diagonal matrix of the phase lags of the d frequencies, and $\boldsymbol{A}$ is a Vandermonde matrix where each column corresponds to an individual complex sinusoid [27]. In the case where the noise is white Gaussian with variance σ_n^2, the autocovariance matrix of the signal vector, $\boldsymbol{x}(k)$, can be written as

$$\boldsymbol{R}_{xx} = E\{\boldsymbol{x}(k)\boldsymbol{x}^H(k)\} = \boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^H + \sigma_n^2\boldsymbol{I}, \tag{A.7}$$

where $\boldsymbol{S} = E\{\boldsymbol{s}(k)\boldsymbol{s}^H(k)\} \in \mathbb{C}^{d\times d}$ is the covariance matrix of the complex amplitudes of the sinusoids. Similarly, the crosscovariance matrix of $\boldsymbol{x}(k)$ and $\boldsymbol{y}(k)$ can be written as

$$\boldsymbol{R}_{xy} = E\{\boldsymbol{x}(k)\boldsymbol{y}^H(k)\} = \boldsymbol{A}\boldsymbol{S}\boldsymbol{\Phi}\boldsymbol{A}^H + \sigma_n^2\boldsymbol{Z}, \tag{A.8}$$

where $\boldsymbol{Z}$ is a square matrix with ones on the first subdiagonal and zeros elsewhere. Then, we define

$$\begin{aligned} \boldsymbol{C}_{xx} &= \boldsymbol{R}_{xx} - \sigma_n^2\boldsymbol{I} = \boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^H, & \text{(A.9)} \\ \boldsymbol{C}_{xy} &= \boldsymbol{R}_{xy} - \sigma_n^2\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{S}\boldsymbol{\Phi}\boldsymbol{A}^H, & \text{(A.10)} \end{aligned}$$

Now, consider the matrix pencil

$$\boldsymbol{C}_{xx} - \lambda\boldsymbol{C}_{xy} = \boldsymbol{A}\boldsymbol{S}(\boldsymbol{I} - \lambda\boldsymbol{\Phi}^H)\boldsymbol{A}^H. \tag{A.11}$$

By inspection, we can see that when $\lambda = e^{j\omega_i}$, the ith row of $\boldsymbol{I} - \lambda \boldsymbol{\Phi}^H$ is zero and the matrix pencil will decrease in rank. Thus, it is obvious that $\lambda = e^{j\omega_i}$ is a generalized eigenvalue of the matrix pair $\{\boldsymbol{C}_{xx}, \boldsymbol{C}_{xy}\}$. Thus, the signal frequencies can be determined from the generalized eigenvalue decomposition of the matrix pair $\{\boldsymbol{C}_{xx}, \boldsymbol{C}_{xy}\}$: The d generalized eigenvalues of the matrix pair closest to the unit circle will correspond to the signal poles.

Appendix B

Singular value decomposition

The singular value decomposition (SVD) provides a means for factoring any matrix, $\boldsymbol{A} \in \mathbb{C}^{m \times n}$ as

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H,$$

where $\boldsymbol{U} \in \mathbb{C}^{m \times m}$ and $\boldsymbol{V} \in \mathbb{C}^{n \times n}$ are unitary and $\boldsymbol{\Sigma}$ is real non-negative diagonal with the elements arranged in non-increasing order [17, ch. 7].

$$\boldsymbol{\Sigma} = diag(\sigma_1, \ldots, \sigma_p), \qquad p = min(m, n).$$

The singular values are the diagonal elements of $\boldsymbol{\Sigma}$. These are real positive numbers ordered such that

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = 0.$$

The rank of $\boldsymbol{A}$ is equal to the number of non-zero eigenvalues: $rank(\boldsymbol{A}) = r$.

The matrix $\boldsymbol{A}$ can be seen as a linear operator that is capable of performing some transformation on an input vector

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}.$$

The column space of $\boldsymbol{A}$ is spanned by the independent columns of $\boldsymbol{A}$. The dimensionality of the column space is equal to the number of independent colums, i.e. $rank(\boldsymbol{A}) = r$. The column space corresponds to the range of the transformation $\boldsymbol{A}\boldsymbol{x}$, i.e. the set of all values the matrix product $\boldsymbol{A}\boldsymbol{x}$ can take. If we wish to find a solution to the equation $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}$ it is thus required that $\boldsymbol{b}$ resides in the column space of $\boldsymbol{A}$.

The nullspace of $\boldsymbol{A}$ corresponds to the vectors $\boldsymbol{x}$ which solve the equation $\boldsymbol{0} = \boldsymbol{A}\boldsymbol{x}$. The dimensionality of the nullspace is determined by the number of linear dependent rows in $\boldsymbol{A}$ since the linear combination of independent rows can only be $\boldsymbol{0}$ for the trivial case of $\boldsymbol{x} = \boldsymbol{0}$. The number of linear dependent rows is $m - r$.

The row space of $\boldsymbol{A}$ is spanned by the independent rows of $\boldsymbol{A}$. The dimensionality of the row space is equal to the number of independent rows, i.e. $rank(\boldsymbol{A}) = r$. The row space can also be decribed as the column space of the conjugate transpose of $\boldsymbol{A}$, i.e. the row space is spanned by $\mathcal{R}(\boldsymbol{A}^H)$. The row space is the orthogonal complement of the nullspace of $\boldsymbol{A}$.

The left nullspace of $\boldsymbol{A}$ corresponds to the vectors which cannot be written as the matrix product $\boldsymbol{A}\boldsymbol{x}$. The dimensionality of the left nullspace is determined by the number of linear dependent columns in $\boldsymbol{A}$. The left nullspace is the orthogonal complement of the range of $\boldsymbol{A}$.

The singular value decomposition of $\boldsymbol{A}$ provides orthonormal bases for all four fundamental subspaces of $\boldsymbol{A}$. A visual interpretation of the four fundamental subspaces is given in B.1. For a discussion of the computational aspects of the SVD see e.g. [16, sec. 8.6].

First	r	columns of $\boldsymbol{U}$:	Column space	$=$	$\mathcal{R}(\boldsymbol{A})$
Last	$m-r$	columns of $\boldsymbol{U}$:	Nullspace	$=$	$\mathcal{N}(\boldsymbol{A})$
First	r	columns of $\boldsymbol{V}$:	Row space	$=$	$\mathcal{R}(\boldsymbol{A}^H)$
Last	$n-r$	columns of $\boldsymbol{V}$:	Left nullspace	$=$	$\mathcal{N}(\boldsymbol{A}^H)$

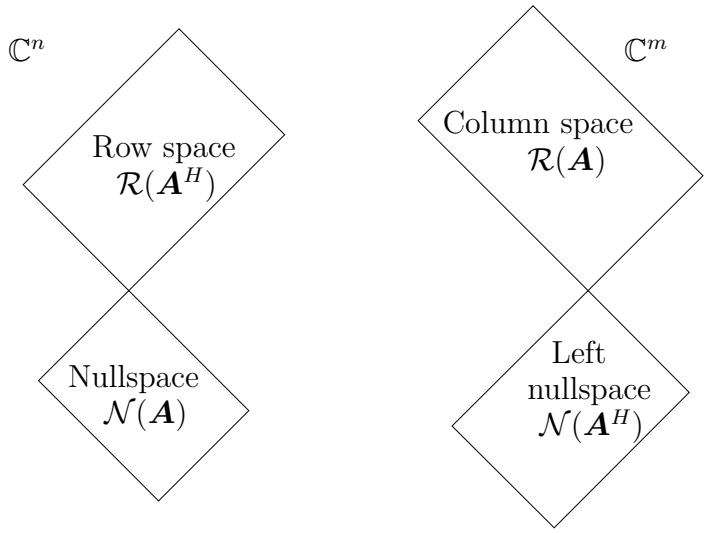

Figure B.1: *Visualization of the four fundamental matrix subspaces [36].*

Appendix C

Signal and noise subspaces

In subspace based signal analysis methods, the notion of signal and noise subspaces for a signal matrix is often used. In the following, we give a short introduction to the idea of signal and noise subspaces.

Consider a signal $x(k)$ consisting of one cosine with frequency ω and amplitude a

$$x(k) = a \cdot \cos(\omega k). \tag{C.1}$$

Using Eulers relation, this can be written as the sum of two complex exponentials

$$x(k) = \frac{a}{2}\left(e^{j\omega k} + e^{-j\omega k}\right). \tag{C.2}$$

We now construct a Hankel structured signal matrix

$$\boldsymbol{X}(k) = \begin{bmatrix} x(k) & x(k+1) & \cdots & x(k+M) \\ x(k+1) & x(k+2) & \cdots & x(k+M+1) \\ \vdots & \vdots & & \vdots \\ x(k+m) & x(k+m+1) & \cdots & x(k+N-1) \end{bmatrix}. \tag{C.3}$$

By the so-called Vandermonde decomposition, this matrix can also be written on the following form

$$\boldsymbol{X}(k) = \begin{bmatrix} 1 & 1 \\ e^{j\omega} & e^{-j\omega} \\ \vdots & \vdots \\ e^{j\omega m-1} & e^{-j\omega m-1} \end{bmatrix} \begin{bmatrix} \frac{a}{2} \cdot e^{j\omega k} & \\ & \frac{a}{2} \cdot e^{j\omega k} \end{bmatrix} \begin{bmatrix} 1 & e^{j\omega} & \cdots & e^{j\omega M} \\ 1 & e^{-j\omega} & \cdots & e^{-j\omega M} \end{bmatrix}. \tag{C.4}$$

By comparing (C.4) to (C.3) and (C.2) it can be verified that this is indeed true. We can identify (C.4) as a sum of two independent outer vector products. Since the rank of an outer vector product is one, we see that the matrix $\boldsymbol{X}(k)$ is of rank two. In general we may say that a signal matrix for a signal consisting of d complex exponentials will be of rank d. Thus, the signal resides in a d-dimensional subspace of $\mathbb{C}^m$. This subspace we denote the signal subspace. The singular value decomposition (see appendix B) can be used to determine this subspace.

Consider now the case where noise is added to the signal, $x(k)$

$$s(k) = x(k) + n(k), \tag{C.5}$$

where $n(k)$ is a stationary, zero-mean, white, gaussian noise signal. If we construct a Hankel structured signal matrix from $s(k)$ it will in general be full rank. Since the noise is uncorrelated, it will span the entire vector space $\mathbb{C}^m$. We may still say that the signal resides in a signal subspace (although it should more accurately now be denoted a signal-plus-noise subspace). The orthogonal complement of the signal subspace, we denote the noise subspace. If we know the dimensionality of the signal subspace, it can be estimated using the singular value decomposition. It will be spanned by the left singular vectors corresponding to the greatest singular values.

By setting the singular values which signify the noise subspace to zero, a rank d matrix approximating the signal matrix can be found. This is the best rank d approximant of the signal matrix in the Frobenius norm. This matrix only contains the signal-subspace part of the original signal matrix.

Example C.1. *Consider a signal consisting of three real sinusoids in additive white gaussian noise. From a signal block of length* 100*, we construct a Hankel signal matrix of dimensions* 87×16*. By taking the SVD of the signal matrix, we get an estimate of the signal and noise subspaces. In figure C.1, the singular value spectrum is shown. Since the model order is known, we know that the first six singular values belong to the signal subspace and the rest of the singular values is associated with the noise subspace. Thus, we know that the signal subspace is spanned by the first six left singular vectors, and the noise subspace is spanned by the remaining left singular vectors.*

□

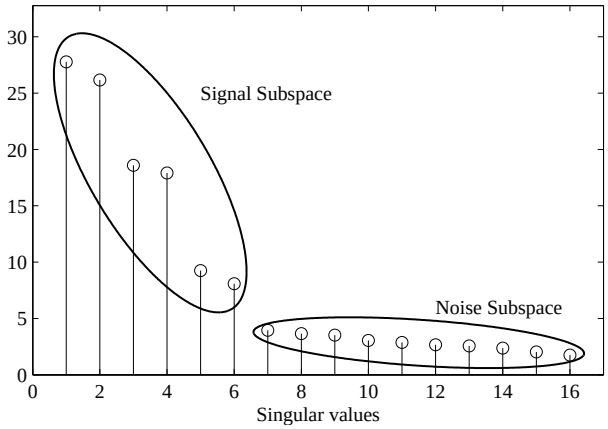

Figure C.1: *Singular value spectrum for a signal consisting of three sinusoids in white gaussian noise. This illustrates the subspace identification problem. The three sinusoids in the signal will span a six-dimensional subspace, whereas the noise spans the whole space.*

Appendix D

Sound files

The files used for experiments have primarily been found in the sound quality assestment material (SQAM) database [44], which is a collection of audio files chosen by the European Broadcasting Union (EBU), for sound quality assestment. The files have been chosen for their characteristic sounds, i.e. tonality, transient behaviour, and noise-like sections. In addition to the files from the SQAM, two other files suplied from our supervisors have also been included.

Each file is sampled at 44.1 kHz and is recorded with audio-CD quality, i.e. each sample is represented in 16 bits resolution. Each file has been reduced to include only a few seconds.

bass47_1_short.wav	A wav-file from the SQAM, of male bass singer.
gspi35_2_short.wav	A wav-file from the SQAM, of a glockenspiel.
quar48_1_short.wav	A wav-file from the SQAM, of a quartet of male and female singers.
spfe49_1_short.wav	A wav-file from the SQAM, of a female speaker reading a english text.
spme50_1_short.wav	A wav-file from the SQAM, of a male speaker reading a english text.
trpt21_2_short.wav	A wav-file from the SQAM, of a trumpet.
sicas3_orig.wav	A wav-file from the supervisors, ABBA - Head over heels, from the album: The Visitors.
sicas13_orig.wav	A wav-file form the supervisors, symphony orchestra and a choir.

Both the shortened, and full version of each file can be found on the CD-ROM in the wav-files section.

Appendix E

Project proposal

The following is the original project proposal written by Jesper Jensen and Søren Holdt Jensen.

Perceptual Unitary ESPRIT Algorithm

Sinusoidal models, which aim at representing signal segments as sums of sinusoidal functions, have proven to provide accurate and flexible representations of a large class of acoustic signals including audio and speech signals. For speech and audio processing, sinusoidal models have been applied in areas such as speech coding, speech enhancement, speech signal transformations, music synthesis, and more recently low bit-rate audio coding.

The problem of robust and accurate estimation of perceptually relevant model parameters based on an observed signal segment is of critical importance in any sinusoidal model based system, especially for coding purposes where a limited set of model parameters is necessary. A potential class of algorithms for solving this estimation problem are the so-called subspace based algorithms. However, a well-known problem with this algorithm class is the difficulty in representing only the perceptually relevant time/frequency regions of the signal in question by exploiting the masking properties of the human auditory system. Only recent research [1] has shown how to achieve this. While the algorithm described in [1] certainly outperforms non-perceptual parameter estimation algorithms, it has two drawbacks: i) it uses exponentially damped sinusoids as basis functions (and thus requires as many as four parameters per sinusoid: amplitude, damping, frequency, and phase), and ii) it is computationally complex. This project attemts to eliminate the problems i) and ii) related to the algorithm in [1] through use of the so-called Unitary ESPRIT algorithm [13], which is a subspace-based algorithm for estimating sinusoids. The goal is to design a variant of the Unitary ESPRIT algorithm which takes perceptual relevance into account in the estimation process. Although theoretically advanced and highly demanding, the computational complexity of the Unitary ESPRIT is lower than the algorithm in [1]. Furthermore, Unitary ESPRIT uses undamped sinusoids as basis functions and thus reduces the parameter set of [1] by eliminating the damping factors. A successful outcome of this project is of current and immediate interest for parametric, low-rate audio coding purposes.

Bibliography

[1] Jesper Jensen, Richard Heusdens, and Søren Holdt Jensen, "A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids," *Accepted for publication in IEEE Transactions on Speech and Audio Processing*, March 2003.

[2] Jesper Jensen, Søren Holdt Jensen, and Egon Hansen, "Exponential sinusoidal modeling of transitional speech segments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 1999, vol. 1, pp. 473–476.

[3] Kris Hermus, Werner Verhelst, and Patrick Wambacq, "Psycho-acoustic modeling of audio with exponentially damped sinusoids," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, vol. 2, pp. 1821–1824.

[4] Robert J. McAulay and Thomas F. Quatieri, "Speech analysis/synthesis based on a sinusiodal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, August 1986.

[5] Tony S. Verma, *A perceptually based audio signal model with application to scalable audio compression*, Ph.D. thesis, Stanford University, 1999.

[6] Julius O. Smith and Xavier Serra, "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusiodal representation," in *Proceedings of the International Computer Music Conference*, 1987.

[7] Jesper Jensen and John H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 731–740, 2001.

[8] Tony S. Verma and Teresa H. Y. Meng, "A 6kbps to 85kbps scalable audio coder," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, June 2000, vol. 2, pp. 877–880.

[9] Alle-Jan Van Der Veen, Ed F. Deprettere, and A. Lee Swindlehurst, "Subspace-based signal analysis using singular value decomposition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1277–1308, 1993.

[10] Sabine van Huffel, Hua Chen, Caroline Decanniere, and Paul Van Hecke, "Algorithm for time-domain NMR data fitting based on total least squares," *Journal of Magnetic Resonance*, vol. 110, no. 2, pp. 228–237, 1994.

[11] Steven van de Par, Armin Kohlrausch, Ghassan Charestan, and Richard Heusdens, "A new psychoacoustical masking model for audio coding applications," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2002, vol. 2, pp. 1805–1808.

[12] Philippe Lemmerling, Ioannis Dologlou, and Sabine Van Huffel, "Speech compression based on exact modeling and structured total least norm optimization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1998, vol. 1, pp. 353–356.

[13] Martin Haardt and Josef A. Nossek, "Unitary ESPRIT: How to obtain increased estimation accuracy with a reduced computational burden," *IEEE Transactions on Signal Processing*, vol. 43, no. 5, pp. 1232–1242, May 1995.

[14] R. V. Cox, *Speech Coding Standards*, chapter 2, pp. 49–78, Elsevier Science B. V., 1995.

[15] Ted Painter and Andreas Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, April 2000.

[16] Gene H. Golub and Charles F. Van Loan, *Matrix computations*, Johns Hopkins, 3rd edition, 1996.

[17] Todd K. Moon and Wynn C. Stirling, *Mathematical methods and algorithms for signal processing*, Prentice Hall, 2000.

[18] Michael Mark Goodwin, *Adaptive Signal Models: Theory, Algorithms, and Audio Applications*, Ph.D. thesis, University of California, Berkeley, 1997.

[19] Alan V. Oppenheim and Ronald W. Schafer, *Discrete-Time Signal Processing*, Prentice-Hall International, 1999.

[20] Mats Viberg and Björn Ottersten, "Sensor array processing based on subspace fitting," *IEEE Transactions on Signal Processing*, vol. 39, no. 5, pp. 1110–1121, 1991.

[21] Petre Stoica and Magnus Jansson, "On forward-backward MODE for array signal processing," *Digital Signal Processing*, vol. 7, no. 4, pp. 239–252, 1997.

[22] Alex B. Gersham and Petre Stoica, "On unitary and forward-backward MODE," *Digital Signal Processing*, vol. 9, no. 2, pp. 67–75, 1999.

[23] A. Lee Swindlehurst, Björn Ottersten, Richard Roy, and Thomas Kailath, "Multiple invariance ESPRIT," *IEEE Transactions on Signal Processing*, vol. 40, no. 4, pp. 876–881, 1992.

[24] Richard Roy, Björn Ottersten, A. Lee Swindlehurst, and Thomas Kailath, "Multiple invariance ESPRIT," in *Twenty-Second Asilomar Conference on Signals, Systems and Computers*, 1989, vol. 2, pp. 583–587.

[25] Richard Roy and Thomas Kailath, "ESPRIT — Estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, July 1989.

[26] Richard Roy, *ESPRIT: Estimation of signal parameters via rotational invariance techniques*, Ph.D. thesis, Stanford University, 1987.

[27] Richard Roy, Arogyaswami Paulraj, and Thomas Kailath, "ESPRIT — A subspace rotation approach to estimation of parameters of cisoids in noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 5, pp. 1340–1342, October 1986.

[28] Arogyaswami Paulraj, Richard Roy, and Thomas Kailath, "Estimation of signal parameters via rotational invariance techniques — ESPRIT," in *Nineteeth Asilomar Conference on Circuits, Systems and Computers*, November 1985, pp. 83–89.

[29] Arogyaswami Paulraj, Richard Roy, and Thomas Kailath, "A subspace rotation approach to signal parameter estimation," *Proceedings of the IEEE*, vol. 74, no. 7, pp. 1044–1045, 1986.

[30] Sun-Yuan Kung, K. S. Arun, R. J. Gal-ezer, and D. V. Bhaskar Rao, "State-space and singular-value decomposition-based approximation methods for the harmonic retrieval problem," *Journal of the Optical Society of America*, vol. 73, no. 12, pp. 1799–1811, 1983.

[31] K. S. Arun and Bhaskar D. Rao, "An improved toeplitz approximation method," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1988, vol. 4, pp. 2352–2355.

[32] D. W. Tufts and R. Kumaresan, "Estimation of frequencies of multiple sinusoids: making linear prediction perform like maximum likelihood," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 975–989, 1982.

[33] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[34] D. V. Bhaskar Rao and K. V. S. Hari, "Performance analysis of esprit and tam in determining the direction of arrival of plane waves in noise," *Acoustics, Speech and Signal Processing*, vol. 37, no. 12, pp. 1990–1995, 1989.

[35] Sabine Van Huffel, *The total least squares problem : Computational aspects and analysis*, vol. 9 of *Frontiers in Applied Mathematics*, Society for Industrial and Applied Mathematics, 1991.

[36] Gilbert Strang, *Introduction to Linear Algebra*, Wellesley-Cambridge Press, 1998.

[37] Anna Lee, "Centrohermitian and skew-centrohermitian matrices," *Linear Algebra and its Applications*, vol. 29, pp. 205–210, 1980.

[38] Brian C.J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, 3rd edition, 1989.

[39] Danish Standards Association, *DS/EN ISO/IEC 11172-3:1995 Information technology — Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s — Part 3: Audio*, DS, 1995.

[40] Richard Heusdens, Renat Vafin, and W. Bastiaan Kleijn, "Sinusoidal modeling of audio and speech using psychoacoustic-adaptive matching pursuits," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, vol. 5, pp. 3281–3284.

[41] John G. Proakis and Dimitris G. Manolakis, *Digital Signal Processing. Principles, Algorithms, and Applications*, Prentice Hall, 3rd edition, 1996.

[42] Fengduo Hu, T. K. Sarkar, and Yingbo Hua, "Utilization of bandpass filtering for the matrix pencil method," *IEEE Transactions on Signal Processing*, vol. 41, no. 1, pp. 442–446, 1993.

[43] Hua Chen, Sabine Van Huffel, and Joos Vandewalle, "Bandpass prefiltering for exponential data fitting with known frequency region of interest," *Signal Processing*, vol. 48, no. 2, pp. 135–154, 1996.

[44] G. Spikofski and H. Jakubowski, "SQAM — the EBU compact disc for subjective assessments of audio systems," *EBU Review, Technical*, pp. 2–6, 1988.