LUND UNIVERSITY

Sparse Modeling of Grouped Line Spectra

Kronvall, Ted

2015

Sparse Modeling of Grouped Line Spectra

with Applications in Audio Processing

Ted Kronvall

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118
SE-221 00 Lund
Sweden

http://www.maths.lth.se/

Licentiate Theses in Mathematical Sciences 2015:3
ISSN 1404-028X

LUTFMS-2017-2015

Printed in Sweden by Media-Tryck, Lund 2015

Acknowledgements

This thesis marks the halfway point of my doctoral education in mathematical statistics at Lund University. In this section, I would like to express my deep gratitude towards the many people who have given me their support, both practical and emotional, in my academic work. First of all, I would like to say thank you to my supervisor Prof. Andreas Jakobsson, who goes out of his way to always be there and find time for me, to point me in the right direction, and to help me with matters big and small. Secondly, I would like to mention my co-authors, especially Dr. Stefan Ingi Adalbjörnsson, Johan Swärd, Simon Burgess, Maria Juhlin, and Filip Elvander. It is my strong belief that the best achievements are accomplished in collaboration, and I am grateful for the opportunity to work with such smart and diligent people. In particular, I am thankful to Stefan, whom, as my senior colleague, has given me much help and guidance. Furthermore, I am grateful to the technical and administrative staff, James Hakim, Lise-Lotte Mörner, Maria Lövgren, and Mona Forsler, for helping me with all those those everyday matters which in turn makes me able to do my part. Also, I would like to mention my colleagues at the department of mathematical statistics, for help in creating a pleasant and socially stimulating work environment. The banter, the fun gossip, and the occasional insights into new geeky hobbies offer much relaxation from the pressure of everyday workload. On a more personal level, I would like to say thank you to my mother and father, Karin and Andrzej, who have been there for me since the beginning. And last, but certainly not least, thank you Hanna, my lovely girlfriend, for sharing my ups and downs, and for being the centre of my life.

Lund, June 2015 *Ted Kronvall*

Abstract

This licentiate thesis focuses on clustered parametric models for estimation of line spectra, when the spectral content of a signal source is assumed to exhibit some form of grouping. Different from previous parametric approaches, which generally require explicit knowledge of the model orders, this thesis exploits sparse modeling, where the orders are implicitly chosen. For line spectra, the non-linear parametric model is approximated by a linear system, containing an overcomplete basis of candidate frequencies, called a dictionary, and a large set of linear response variables that selects and weights the components in the dictionary. Frequency estimates are obtained by solving a convex optimization program, where the sum of squared residuals is minimized. To discourage overfitting and to infer certain structure in the solution, different convex penalty functions are introduced into the optimization. The cost trade-off between fit and penalty is set by some user parameters, as to approximate the true number of spectral lines in the signal, which implies that the response variable will be sparse, i.e., have few non-zero elements. Thus, instead of explicit model orders, the orders are implicitly set by this trade-off. For grouped variables, the dictionary is customized, and appropriate convex penalties selected, so that the solution becomes group sparse, i.e., has few groups with non-zero variables. In an array of sensors, the specific time-delays and attenuations will depend on the source and sensor positions. By modeling this, one may estimate the location of a source. In this thesis, a novel joint location and grouped frequency estimator is proposed, which exploits sparse modeling for both spectral and spatial estimates, showing robustness against sources with overlapping frequency content. For audio signals, this thesis uses two different features for clustering. Pitch is a perceptual property of sound that may be described by the harmonic model, i.e., by a group of spectral lines at integer multiples of a fundamental frequency, which we estimate by exploiting a novel adaptive total variation penalty. The other feature, chroma, is a concept in musical theory, collecting pitches at powers of 2 from each other into groups. Using a chroma dictionary, together with appropriate group sparse penalties, we propose an automatic transcription of the chroma content of a signal.

Keywords

line spectra, parameter estimation, convex optimization, group-sparsity, block-sparsity, dictionary learning, ADMM, adaptive penalty, total variation, multi-pitch estimation, chroma, audio processing, TDOA, near-field localization, amplitude modulation

Contents

List of papers

This thesis is based on the following papers:

A Ted Kronvall, Filip Elvander, Stefan Ingi Adalbjörnsson, and Andreas Jakobsson, "An Adaptive Penalty Approach to Multi-pitch Estimation". *23rd European Signal Processing Conference*, Nice, France, August 31 - September 4, 2015.

B Stefan Ingi Adalbjörnsson, Ted Kronvall, Simon Burgess, Kalle Åström, and Andreas Jakobsson, "Sparse Localization of Harmonic Audio Sources". Submitted to *IEEE Transactions on Audio, Speech, and Language Processing*.

C Ted Kronvall, Stefan Ingi Adalbjörnsson, and Andreas Jakobsson, "Joint DOA and Multi-pitch Estimation via Block Sparse Dictionary Learning", *22nd European Signal Processing Conference*, Lisbon, Portugal, September 1-5, 2014.

D Ted Kronvall, Maria Juhlin, Johan Swärd, Stefan Ingi Adalbjörnsson, and Andreas Jakobsson, "Sparse Modeling of Chroma Features". To be submitted.

Additional papers not included in the thesis:

1. Maria Juhlin, Ted Kronvall, Johan Swärd, and Andreas Jakobsson, "Sparse Chroma Estimation for Harmonic Non-stationary Audio", *23rd European Signal Processing Conference*, Nice, France, August 31 - September 4, 2015.

2. Ted Kronvall, Maria Juhlin, Stefan Ingi Adalbjörnsson, and Andreas Jakobsson, "Sparse Chroma Estimation for Harmonic Audio", *40th IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brisbane, Australia, April 19-24, 2015.

3. Stefan Ingi Adalbjörnsson, Johan Swärd, Ted Kronvall, and Andreas Jakobsson, "A Sparse Approach for Estimation of Amplitude Modulated Sinusoids", *The Asilomar Conference on Signals, Systems, and Computers*, Asilomar, USA, November 2-5, 2014.

4. Ted Kronvall, Stefan Ingi Adalbjörnsson, and Andreas Jakobsson, "Joint DOA and Multi-pitch Estimation using Block Sparsity", *39th IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 4-9, 2014.

5. Ted Kronvall, Naveed R. Butt, and Andreas Jakobsson, "Computationally Efficient Robust Widely Linear Beamforming for Improper Non-stationary Signals", *21st European Signal Processing Conference*, Marrakech, Morocco, September 9-13, 2013.

6. Ted Kronvall, Johan Swärd, Andreas Jakobsson, "Non-Parametric Data-Dependent Estimation of Spectroscopic Echo-Train Signals", *38th IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 26-31, 2013.

Lists of notation

Typical notational conventions

$\mathbf{a}$, $\mathbf{b}$, ...	boldface lower case letters denote column-vectors
$\mathbf{A}$, $\mathbf{B}$, ...	boldface upper case letters denote matrices
A, a, Δ, α, ...	non-bold letters generally denote scalars
$(\cdot)^T$	vector or matrix transpose
$(\cdot)^H$	Hermitian (conjugate) transpose
$(\cdot)^\dagger$	Moore-Penrose pseudo-inverse
$\hat{(\cdot)}$	an estimated parameter
$(\cdot)_+$	positive threshold of real scalar, $(a)_+ = \max(0, a)$
$\{\cdot\}$	the set of elements or other entities
$\lvert\cdot\rvert$	magnitude of complex scalar
$\lvert\lvert\cdot\rvert\rvert$	the Euclidean norm of a vector, $\lvert\lvert a\rvert\rvert = \sqrt{\mathbf{a}^H\mathbf{a}}$
$\lvert\lvert\cdot\rvert\rvert_q$	the ℓ_q-norm of a vector, $\lvert\lvert a\rvert\rvert_q = \left(\sum_p \lvert a_p\rvert^q\right)^{1/q}$ but is not a proper norm for $p < 1$
$\lvert\lvert\cdot\rvert\rvert_0$	the ℓ_0-"norm" of a vector, $\lvert\lvert a\rvert\rvert_0 = \sum_p \lvert a_p\rvert^0$
$\lvert\lvert\cdot\rvert\rvert_\mathcal{F}$	the Frobenius norm of a matrix
abs$(\cdot)$	element-wise magnitude of (a vector or matrix)
arg$(\cdot)$	element-wise complex argument of
$\mathbb{R}^{n\times m}$	the real $n \times m$-dimensional space
$\mathbb{R}^n$	the real n-dimensional plane ($\mathbb{R}$ is used for $n = 1$)
$\mathbb{C}^{n\times m}$	the complex $n \times m$-dimensional space
$\mathbb{C}^n$	the complex n-dimensional space
$\mathbb{Q}$	the set of rational numbers
$\mathbb{Z}$	the set of integers
$\mathbb{N}$	the set of natural numbers
Im$(\cdot)$	the imaginary part of
Re$(\cdot)$	the real part of
i	the imaginary unit, $\sqrt{-1}$, unless otherwise specified

$\forall$	for all (members in the set)
$\triangleq$	defined as
$\approx$	approximately equal to
$\times$	multiplied by, or dimensionality
$\otimes$	Kronecker product by
∂	differential of
$\in$	belongs to (a set)
$\subseteq$	is a subset of (a set)
Prob$(\cdot)$	probability of event
$\mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$	the multivariate Normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{R}$
Cov$(\cdot)$	the covariance matrix
arg max$(\cdot)$	the argument that maximizes
arg min$(\cdot)$	the argument that minimizes
vec$(\cdot)$	column-wise vectorization of a matrix
diag$(\cdot)$	diagonal matrix with specified diagonal vector
1-D, 2-D, ...	one-dimensional, two-dimensional, ...

Specific notations

N	total number of time points
t, n	time, an index or in seconds
M	total number of sensors
m	sensor index
$y(t)$	sampled signal at time t
$\mathbf{y}$	$N \times 1$ sample signal vector
$\mathbf{Y}$	$N \times M$ sample signal array matrix
$x(t), \mathbf{x}, \mathbf{X}$	noise-free signal
$e(t), \mathbf{e}, \mathbf{E}$	sample noise
$\mathbf{I}$	the identity matrix (of unspecified dimension)
$\mathbf{I}_n$	the $n \times n$ identity matrix
P	the total number of dictionary atoms
p	index of dictionary atoms
L_p	the total number of components in the p:th atom
f_p	normalized frequency of the p:th atom
ℓ	the index of a component within an atom
$\mathbf{W}$	$N \times \sum_p L_p$ dictionary, if not otherwise defined
$\mathbf{W}_p$	$N \times L_p$ sub-dictionary for the p:th atom
$\mathbf{w}_p, \mathbf{w}_{p,\ell}$	single dictionary component for the p:th atom
$\mathbf{a}$	amplitude vector corresponding to the dictionary
$\mathbf{a}_p$	amplitude vector for the p:th sub-dictionary
a_p	amplitude for the p:th dictionary component
$L(\cdot)$	the Lagrangian function
λ	sparsity-promoting regularlization parameter
λ_j	regularization parameter for the j:th penalty function
$\tau, \boldsymbol{\tau}$	time-delay(s), TOA(s), or TDOA(s)
$\vartheta, \boldsymbol{\vartheta}$	DOA(s)
$\boldsymbol{\Psi}, \boldsymbol{\psi}, \ldots$	bold-face greek letters generally denote parameter sets

List of abbreviations

ANLS	Approximative Non-linear Least Squares
ADMM	Alternating Directions Method of Multipliers
BEAMS	Block sparse Estimation of Amplitude Modulated Signals
CCA	Cross-Correlation Analysis
CCD	Cyclic Coordinate Descent
CEAMS	Chroma Estimation of Amplitude Modulated Signals
CEBS	Chroma Estimation using Block Sparsity
CRLB	Cramér-Rao Lower Bound
DFT	Discrete Fourier Transform
DOA	Direction-Of-Arrival
HALO	Harmonic Audio LOcalization
IAPEBS	Iterative Adaptive Pitch Estimation using Block Sparsity
KKT	Karush-Kuhn-Tucker
LARS	Least Angle RegreSsion
LASSO	Least Absolute Shrinkage and Selection Operator
LS	Least Squares
NLS	Non-linear Least Squares
MC	Monte Carlo
MIR	Music Information Retreival
ML	Maximum Likelihood
PEBS	Pitch Estimation using Block Sparsity
PEBSI-Lite	PEBS - Improved and Lighter
PEBS-TV	PEBS - Total Variation
RMSE	Root Mean Square Error
SFL	Sparse Fused LASSO
SGL	Sparse Group LASSO
SNR	Signal-to-Noise Ratio
SOC	Second Order Cone
STFT	Short-Time Fourier Transform
TDOA	Time-Difference-Of-Arrival

TOA	Time-Of-Arrival
TR	Tikhonov Regularization
TV	Total Variation
ULA	Uniform Linear Array

Introduction

These lines introduce a licentiate thesis in the wide and exciting field of signal processing. It takes the perspective of *statistical* signal processing, especially that of Kay (1993) [1] and Scharf (1991) [2], whose good practices hopefully will shine through in the analysis, solution, and execution done here. In particular, the problems raised share a strong connection to spectral estimation, and many fundamental results are based upon the standard reference of Stocia and Moses (2005) [3]. In line with this heritage, this thesis attempts to judge performance from a statistical point of view, i.e., whether estimation procedures are good or bad in terms of, e.g., *efficiency*, *consistency*, and *bias*. The methods presented in this thesis are closely related mathematically. The basic underlying assumption is the parametric sinusoidal model, where signals are assumed to be well modeled as super-positioned complex sinusoids, having both linear and non-linear parameters, corrupted by some additive noise. Furthermore, the thesis is in particular concerned with sinusoids that experience some form of natural grouping, or structure, and where interest lies in specifying properties of this structure. Grouping of components often pose combinatorial issues, as the structural criteria may be implicitly defined, or as groups may have overlapping components, which the thesis will focus on solving using sparse modeling via convex optimization. Thereby, explicit model orders may be set implicitly, by adding regularization on the parameters, and so alleviate the need of model order estimation, which is a difficult problem necessary for parametric modeling. The main formulation and analysis for sparse modeling derives from the work of Tibshirani (1996) [4], herein extended with a variety of criterions which enforce a certain structure. Experience shows how grouped sinusoids can be used to describe the tonal part in acoustical signals, where the frequency components of an audio source often exhibit a predetermined relationship, from which a cluster may be formed. One such predetermined relationship, commonly used in audio applications, is *pitch*, which groups spectral content according to a model for human perception of sound. Another method of grouping the spectral content is according to *chroma*, which is a feature that is important in music information retrieval (MIR) applications. Furthermore, this thesis will touch upon the field of array processing, where sig-

nals are also attributed with some spatial information. In fact, many results in spectral analysis may be used in array processing, and vice versa, as these fields are highly related. To give some fundamental context for the papers of which this thesis consists, some preliminaries from spectral analysis, sparse estimation, audio analysis, and array processing will constitute the bulk of this introductory chapter. Lastly, an overview of the papers in this thesis is given.

1 Preliminaries on estimating line spectra

This section will introduce some preliminary results for parametric estimation of line spectra. For many applications, a periodic signal of interest may often be well described by the sinusoidal model

$$y(t) = x(t) + e(t), \quad x(t) = \sum_{k=1}^{K} a_k e^{i2\pi f_k t} \tag{1}$$

where $x(t)$ denotes the noise-free super-positioning of K sinusoidal components, that are sampled in some form of additive noise, $e(t)$, typically on a unit grid, $t = 0, \ldots, N-1$. For the k:th component, a_k and $f_k \in [0, 1)$ denote the complex-valued amplitude and the frequency, respectively. By forming the sample vector

$$\mathbf{y} = \begin{bmatrix} y(0) & \ldots & y(N-1) \end{bmatrix}^T \tag{2}$$

the sinusoidal model (1) may be equivalently formulated as

$$\mathbf{y} = \mathbf{x} + \mathbf{e}, \quad \mathbf{x} = \sum_{k=1}^{K} \mathbf{w}_k a_k = \mathbf{W}\mathbf{a} \tag{3}$$

where the noise-free signal vector, $\mathbf{x}$, and the noise vector, $\mathbf{e}$, are defined similarly to $\mathbf{y}$. Thus, some simple algebraic manipulations allows the signal vector to be compactly expressed as a matrix-vector multiplication, given that

$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_1 & \ldots & \mathbf{w}_K \end{bmatrix} \tag{4}$$

$$\mathbf{w}_k = \begin{bmatrix} e^{i2\pi f_k 1} & \ldots & e^{i2\pi f_k (N-1)} \end{bmatrix}^T \tag{5}$$

$$\mathbf{a} = \begin{bmatrix} a_1 & \ldots & a_K \end{bmatrix}^T \tag{6}$$

The noise-free signal vector may therefore also be seen as a linear combination of the columns in $\mathbf{W}$, which represents each sinusoidal component, using the complex weights in $\mathbf{a}$. If K is known *a priori*, it may be convenient to view (3) as a non-linear regression problem, where the spectral components at frequencies $\boldsymbol{\Psi} = \{f_k\}_{k=1}^K$ are multiplied by the linear amplitudes. This formulation allows the unknown parameters to be estimated using the well-known Least Squares (LS) criterion, given by

$$\left\{\hat{\boldsymbol{\Psi}}_{\mathrm{LS}}, \hat{\mathbf{a}}_{\mathrm{LS}}\right\} = \arg\min_{\boldsymbol{\Psi},\mathbf{a}} ||\mathbf{y} - \mathbf{W}\mathbf{a}||_2^2 \tag{7}$$

i.e, as the arguments minimizing the sum of squared model residuals. A closed form estimate of the amplitudes for a given selection of $\boldsymbol{\Psi}$ can be obtained by solving the normal equations, i.e.,

$$\hat{\mathbf{a}}_{\mathrm{LS}} = \mathbf{W}^{\dagger}\mathbf{y}, \quad \mathbf{W}^{\dagger} \triangleq \left(\mathbf{W}^H\mathbf{W}\right)^{-1}\mathbf{W}^H\mathbf{y} \tag{8}$$

which, if inserted into (7), gives the Non-Linear LS (NLS) criterion

$$\hat{\boldsymbol{\Psi}}_{\mathrm{LS}} = \arg\max_{\boldsymbol{\Psi}} \mathbf{y}^H\mathbf{W}\left(\mathbf{W}^H\mathbf{W}\right)^{-1}\mathbf{W}^H\mathbf{y}. \tag{9}$$

One may then, for instance, form the frequency estimates by maximizing the NLS criteria over a K-dimensional grid. Furthermore, as is shown in, e.g., [5], the NLS estimation errors of $\boldsymbol{\Psi}$ will have the asymptotic covariance matrix

$$\mathrm{Cov}(\hat{\boldsymbol{\Psi}}) = \frac{6\sigma^2}{N^3}\,\mathrm{diag}\left(\begin{bmatrix} \frac{1}{a_1^2} & \cdots & \frac{1}{a_K^2} \end{bmatrix}\right) \tag{10}$$

where diag($\mathbf{c}$) denotes a diagonal matrix the vector $\mathbf{c}$ along its diagonal. In the case of white Gaussian noise, i.e., $\mathbf{e} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, the covariances in (10) reach the Cramér-Rao Lower Bound (CRLB), as was shown in, e.g., [6], which gives the lower bound for the covariance matrix of any unbiased estimator of $\boldsymbol{\Psi}$. A similar analysis can be done for $\hat{\mathbf{a}}_{\mathrm{LS}}$, showing that the NLS method provides a statistically efficient estimate of the parametric line spectra. However, the NLS criterion works poorly in practice for this problem, and the reason is twofold. Firstly, (9) is highly multimodal and the global maximum is very sharp, and so, to obtain the correct estimates, the maximization needs to be well initialized, as well as evaluated over a sufficiently fine grid. Secondly, any two frequencies must be sufficiently separated in order for the estimator to work properly. To see this,

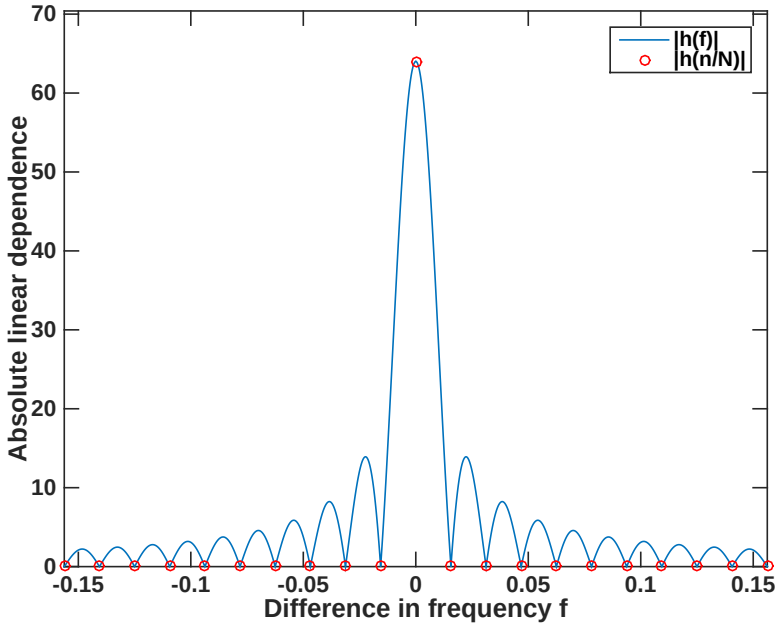


Figure 1: Absolute values of the complex function $h(\Delta f)$, measuring the amount of linear dependence between two Fourier vectors, spaced in frequency by Δf. The example illustrates the function for $N = 64$ samples, where orthogonality is found at every $n/N, n \neq 0$.

consider the square matrix $\mathbf{W}^H\mathbf{W}$ in the middle of the NLS criterion, which needs to be inverted. This matrix measures the linear dependence between the components in $\mathbf{W}$, and some simple calculations shows that its p,k element will only depend on the difference in frequency [3, p. 160], i.e.,

$$\left\{\mathbf{W}^H\mathbf{W}\right\}_{p,k} = \mathbf{w}_p^H\mathbf{w}_k \tag{11}$$

$$= \begin{cases} N & f_p = f_k \\ e^{i2\pi(f_k-f_p)}\frac{e^{i2\pi N(f_k-f_p)}-1}{e^{i2\pi(f_k-f_p)}-1} & f_p \neq f_k \end{cases} \tag{12}$$

$$\triangleq h(f_k - f_p) \tag{13}$$

where a special case is $h(n/N) = 0$, for $n = \{n \in \mathbb{Z} : n \neq 0\}$. An example of this function can be seen in Figure 1, which shows the absolute values of the function for $N = 64$. Thus, if two frequencies are too closely spaced, the columns of $\mathbf{W}$ become linearly dependent, making the inversion and the estimation problem ill-conditioned. In fact, under the quite restrictive assumption that all frequencies in $\mathbf{\Psi}$ are spaced some non-zero distance apart, say n/N, $\mathbf{W}^H\mathbf{W} = N\mathbf{I}$ and (9) reduces to

$$\hat{\mathbf{\Psi}} = \arg\max_{\mathbf{\Psi}} \left|\left|\mathbf{W}^H\mathbf{y}\right|\right|_2^2, \quad \hat{\mathbf{a}}_{\text{LS}} = \frac{1}{N}\mathbf{W}^H\mathbf{y} \tag{14}$$

which is the sum of periodogram estimates in $\mathbf{\Psi}$. Given this, some remarks the performance of the periodogram for estimation of line spectra may be noted.

Remark 1: For a single sinusoid in white Gaussian noise, i.e., $K = 1$, the periodogram is the ML estimator, as $\mathbf{W}^H\mathbf{W} = \mathbf{w}^H\mathbf{w} = N$. To obtain the estimate in (14), one usually evaluates the periodogram on an oversampled DFT grid, i.e.,

$$\mathbf{\Psi} = \left\{\frac{p}{sN}\right\}_{p=0,\ldots,sN-1} \tag{15}$$

where s is the oversampling or super-resolution factor, and picking the largest peak of the corresponding magnitude estimate

$$|\hat{\mathbf{a}}| = \frac{1}{N}|\mathbf{W}^H\mathbf{y}| \tag{16}$$

An example is shown in Figure 2, where a single sinusoid in white Gaussian noise have been measured in $N = 64$ samples. A finely oversampled DFT estimate is compared to the true frequencies and to the LS estimate for the correct frequency (8). As can be seen from the figure, the peak of the DFT estimate coincides with the true frequency and the LS amplitude estimate.

Remark 2: For $K > 1$, one usually proceeds in the same manner as for one sinusoid. In the unlikely case that all frequencies are separated by at least $1/N$ and lie exactly on the standard DFT grid, where $s = 1$, the periodogram would be an efficient estimator, as $\mathbf{W}^H\mathbf{W} = N\mathbf{I}$ and so (9) and (14) are equal. Otherwise, when the frequencies lie off-grid, the periodogram is typically a reasonable, but not an efficient, estimator [3, p. 161].

Remark 3: The resolution of the periodogram is limited, so that two sinsuoids closely spaced in frequency are only likely to be resolved if that spacing is at least

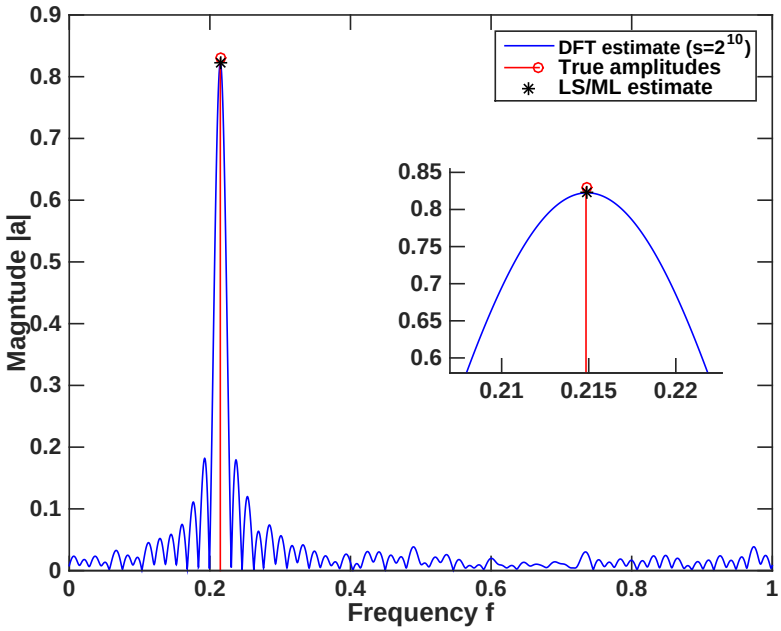


Figure 2: DFT amplitude estimates for $K = 1$ sinusoids in white Gaussian noise, as compared to the true amplitudes, and the ML estimate for the correct frequency.

$1/N$. If spaced finer, they will appear to coincide in the resulting spectral estimate. An example of this is shown in Figure 3, where $K = 5$ sinusoids are measured at $N = 64$ samples. Two of the sinusoids are spaced by $\frac{2}{5N}$, i.e., being two grid-points apart if having super-resolution $s = 5$, and are thus not resolved using the oversampled DFT periodogram. However, if given the correct frequencies, the LS estimates from (8) resolves the peaks very accurate. Thus, the problem resides in finding the non-linear frequency parameters. Other commonly used parametric methods for frequency estimation with good statistical accuracy include the HOYW, MUSIC, and ESPRIT methods [3, ch. 4].

Remark 4: Throughout the analysis in this section, the model order, K, is assumed to be known, which is also a requirement for most parametric estimation methods. However, in practice, the model order is typically unknown, which re-

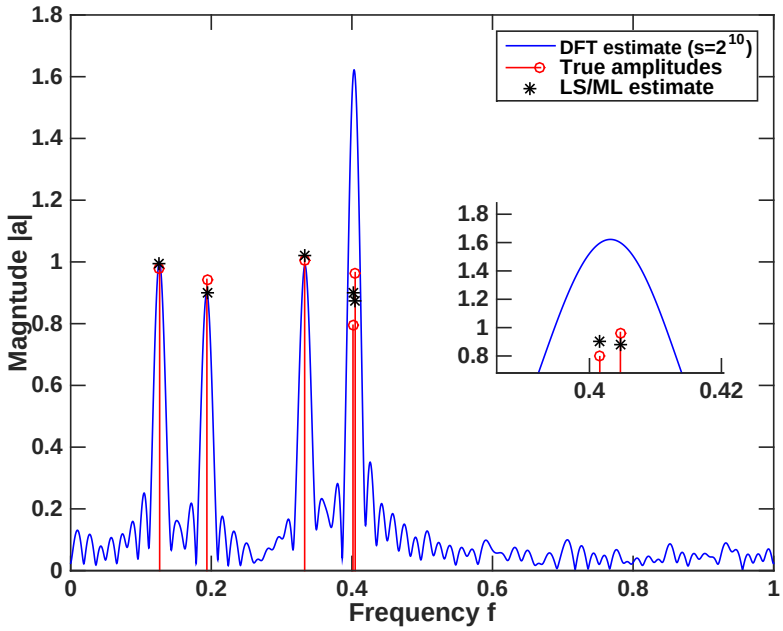


Figure 3: The DFT amplitude estimates for $K = 5$ sinusoids in white Gaussian noise, with two of them being spaced by $\frac{2}{5N}$, as compared to the true amplitudes, and the ML estimate for the correct frequencies.

quires an often difficult model order estimation procedure. In the next section, a methodology for line spectra with super-resolution capabilities is introduced, where explicit model orders are not required. This is the sparse modeling approach.

2 Sparse estimation of line spectra

In this section, a methodology for estimation of line spectra using sparse modeling is introduced. The main idea is quite simple, and may be put in a question: Given an over-complete set of candidate basis functions for a certain type of signal, which is the sparsest possible subset of them to model it with? The solution, which was first formalized in [4] using a statistical framework and convex ana-

lysis, is termed the Least Absolute Shrinkage and Selection Operator (LASSO). The LASSO solves an optimization problem where an LS cost is supplemented by a penalty function to avoid overfitting, i.e., to prevent having a surplus of active basis functions. The same idea goes under different acronyms, and is also referred to as the Basis Pursuit De-Noising (BPDN) method [7]. It has been the constant focal point of much research during the last decade and a half, and many prominent researchers have worked on the theoretical properties, solution algorithms, applications, and extensions of the method.

2.1 Promoting sparsity

A common application for the LASSO is the estimation of line spectra. In contrast with the NLS criteria in (9), which is non-linear, the LASSO formulation enables an entirely linear estimation problem. It represents the non-linear frequency parameters with an overcomplete set of candidate frequencies, each with a linear amplitude parameter, and then promotes a solution where only a small number of these amplitudes are non-zero. The LASSO solution is thus highly accurate, and also circumvents the requirement of explicitly defining the model order. To that end, let the K non-linear frequency parameters be represented by $P = sN \gg K$ sinusoidal basis functions, chosen such that some of them may well coincide with the true sinusoids. The problem thus reduces into finding the most sparse representation of the linear amplitude parameters. In this thesis, the approach is referred to as sparse modeling of sinusoids, as the sinusoidal signal is modeled by a sparse representation of an oversampled DFT basis. Re-using the same notation as in the previous section, let $\mathbf{\Psi}$ denote the candidate frequencies of the s-times oversampled DFT grid, as given in (15). The noise-free signal in (3) may thus be approximated by

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{w}_k a_k \approx \sum_{p=1}^{P} \mathbf{w}_p a_p = \mathbf{W}_{\mathbf{\Psi}} \mathbf{a} \tag{17}$$

where $\mathbf{W}_{\mathbf{\Psi}}$ denotes the over-complete sinusoidal base, and is also referred to as the signal dictionary, as it constitutes all possible representations of the signal. From the dictionary, the signal should be well described using few sinusoidal components, being commonly referred to as the atoms, i.e., the smallest building blocks of the dictionary. For notational convenience, the subscript of the dictionary is dropped. Thus, the LASSO for line spectra solves the optimization

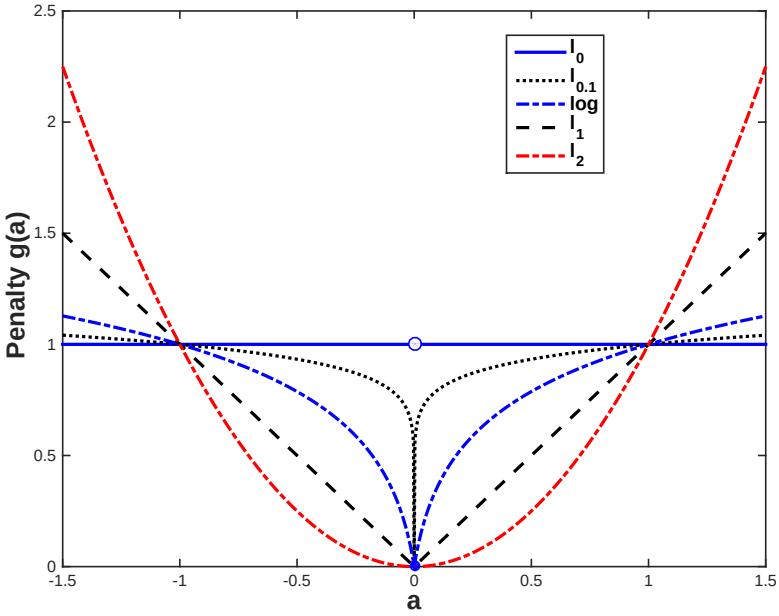


Figure 4: A comparison of different penalty functions for a scalar input x. The ℓ_0 penalty is the most sparsity-enforcing, as any deviation from zero adds cost. Only the ℓ_1 and ℓ_2 functions are convex, whereof only the former enforces sparsity.

problem

$$\underset{\mathbf{a}}{\text{minimize}}\ \frac{1}{2}\,||\mathbf{y}-\mathbf{W}\mathbf{a}||_2^2 + N\lambda||\mathbf{a}||_1 \tag{18}$$

where the sum of squared residuals cost function is extended by a sparsity-promoting penalty-function, i.e., the ℓ_1 norm of the amplitudes, which is multiplied by N for scaling and a positive constant λ. Thus, for every non-zero amplitude component a_p which helps to increase the fit in (18), a corresponding penalty $N\lambda|a_p|$ is added to the cost, and so λ trades off solution fit for sparsity. To illustrate why the ℓ_1 norm promotes sparsity, Figure 4 shows some suggestions of other possible penalty functions, namely

$$||\mathbf{a}||_0 = \sum_{p=0}^{P-1} 1\{a_p \neq 0\} \tag{ℓ_0}$$

$$||\mathbf{a}||_q = \left(\sum_{p=0}^{P-1} |a_p|^q \right)^{1/q} \tag{ℓ_q}$$

$$g(\mathbf{a}) = \frac{1}{1+c} \sum_{p=0}^{P-1} \ln\left(1 + c|a_p|\right) \tag{log}$$

for $q = \{0.1, 1, 2\}$, and where c in (log) is a positive constant, which increases the absolute slope close to zero. In the figure, c is set to 20. What is of interest, is whether and how fast any deviation from zero adds a cost. The most sparsity-promoting penalties are thus ℓ_0, ℓ_q, for $0 < q < 1$, and log. All of these are, however, non-convex, which strongly restrict their practical utility. For correctness, note that ℓ_q, for $q < 1$, is not a proper norm. As a stark contrast, the ℓ_2 penalty does not promote sparsity, but rather the opposite, as a small deviation from zero adds a relatively small penalty. Hence, being both convex and sparsity-promoting, the ℓ_1 is not an unintuitive choice of penalty. As a convex optimization problem, the LASSO enjoys the attractive property that if a local minima is found, it is also the global minima. From convex analysis, a necessary and sufficient condition for a solution to be optimal is that it fulfills the Karush-Kuhn-Tucker (KKT) conditions. In this thesis, solutions are found using numerical methods that ensures KKT; for simpler problems, they may also be solved analytically. In the next section, a closed form solution for the LASSO is derived using KKT, as it may give a qualitative understanding of the effect of penalizing the LS problem, as well as the effect of λ. The cost function in (18) is a real scalar function, say g, which takes complex arguments, i.e., $g : \mathbb{C}^{P\times 1} \to \mathbb{R}$. Instead of using complex analysis, the problem may be formulated by considering the real and imaginary parts of the arguments and dictionary separately, i.e., $g : \mathbb{R}^{2P\times 1} \to \mathbb{R}$. Next, the LASSO solution is derived for real-valued arguments, to which the omitted complex-valued case is a simple, but notationally tedious, extension.

2.2 LASSO solution via KKT

Consider the real-valued case where $\mathbf{a} \in \mathbb{R}^{P\times 1}$, and where $\mathbf{W} \in \mathbb{R}^{N\times P}$ is a dictionary with arbitrary atoms. The KKT conditions for the LASSO thus state

that [8]

$$\frac{\partial f}{\partial \mathbf{a}} = -\mathbf{W}^T\left(\mathbf{y} - \mathbf{W}\mathbf{a}\right) + N\lambda\mathbf{v} = \mathbf{0} \tag{19}$$

$$v_p = \begin{cases} \frac{a_p}{|a_p|} & \forall a_p \neq 0 \\ \in [-1, 1] & \forall a_p = 0 \end{cases} \tag{20}$$

where $\mathbf{v}$ is the sub-gradient of the penalty function $||\mathbf{a}||_1$. In order to solve (19) and (20) for $\mathbf{a}$, some additional notation is convenient. Define $\mathbf{W}_{-p}$ and $\mathbf{a}_{-p}$ as the dictionary and the response variable where the p:th component is left out. With some matrix algebra, (19) may be compactly expressed as

$$a_p = \breve{a}_p - \lambda v_p \tag{21}$$

for $p = 1, \ldots, P$, where

$$\breve{a}_p \triangleq \frac{1}{N}\mathbf{w}_p^T\left(\mathbf{y} - \mathbf{W}_{-p}\mathbf{a}_{-p}\right) \tag{22}$$

denotes the unconstrained solution of a_p, given $\mathbf{a}_{-p}$, and may be interpreted as the linear dependence between the p:th dictionary component and the signal residual, i.e., where the components in $\mathbf{a}_{-p}$ have been removed from the original signal. Furthermore, (21) can be used to show that

$$\begin{array}{lll} \lambda & \geq |\breve{a}_p| & \forall a_p = 0 \\ |a_p| & = |\breve{a}_p| - \lambda \geq 0 & \forall a_p \neq 0 \end{array} \tag{23}$$

which, together with some further manipulations, yields the closed form LASSO solution as

$$\hat{a}_p = \left(1 - \frac{\lambda}{|\breve{a}_p|}\right)_+ \breve{a}_p \tag{24}$$

for each component in $\mathbf{a}$, where $(c)_+$ returns c if positive, zero otherwise. This solution fulfills the KKT conditions and yields the global optimum of (18). From (24), the effect of λ becomes clearly visible. Any $\lambda > 0$ will induce a bias, which for $\lambda < |a_p|$ will be equal to $-\lambda$, and for $\lambda \geq |a_p|$, will be equal to $-\breve{a}_p$, as the LASSO solution will be zero. It also becomes clear why the ℓ_1-norm promotes sparsity, as it takes the non-sparse unconstrained solution and soft-thresholds it,

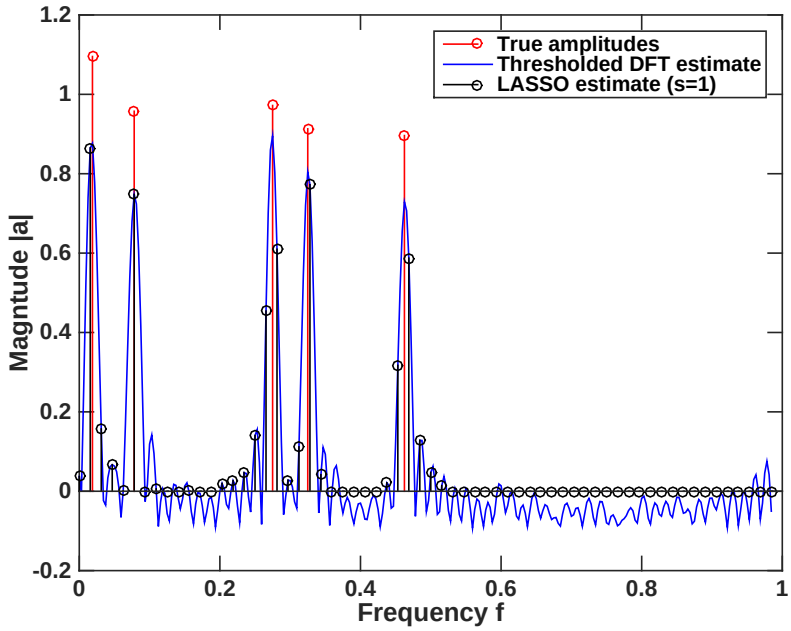


Figure 5: The LASSO amplitude estimates for $K = 5$ well separated sinusoids in white Gaussian noise. In comparison with a thresholded DFT estimate, $\hat{a}_{\text{DFT}} - \lambda$.

such that small components become zero. For the special case when the dictionary is orthogonal, i.e., $\mathbf{W}^T\mathbf{W} = N\mathbf{I}$, $\breve{a}_p$ does not depend on $\mathbf{a}_{-p}$, and will be simple to calculate for all components. For line spectra, the orthogonality implies that the LASSO solution is the equivalent of soft-thresholding the periodogram, as $\mathbf{W}$ becomes the DFT matrix with $s = 1$. An example of this may be seen in Figure 5, where five sinusoids, well-spaced by more than $1/N$ from each other, are estimated using the LASSO with such an orthogonal dictionary, for $N = 64$. For comparison, the DFT estimate thresholded with the bias, i.e., $\hat{a}_{\text{DFT}} - \lambda$, is shown, as to illustrate the soft-thresholding of the LASSO estimate.

2.3 Super-resolution and robust recovery

For non-orthogonal dictionaries, the LASSO solution is less easily interpreted, as the parameters will have some degree of mutual dependence. This can be

illustrated using (13), as

$$\breve{a}_p = \frac{1}{N}\left(\mathbf{w}_p^T\mathbf{y} - \sum_{q\neq p}\mathbf{w}_p^T\mathbf{w}_q a_q\right) \tag{25}$$

$$= \frac{1}{N}\mathbf{w}_p^T\mathbf{y} - \frac{1}{N}\sum_{q\neq p} h(f_p - f_q)a_q \tag{26}$$

Intuitively, consider two frequency components, f_p and $f_{p'}$, that are very closely spaced, such that $h(f_p - f_{p'}) \approx 1$. If an estimate of $a_{p'}$ reasonably well represents the true spectral energy in that component, then $\breve{a}_p$ becomes small, making $\hat{a}_p$ small. But the same also holds for $a_{p'}$ if an estimate a_p well represents the true spectral energy, and so the LASSO instead seemingly arbitrarily divides magnitude between highly coherent components. An example of this is seen in Figure 6, where the LASSO estimate for a dictionary with super-resolution $s = 20$ is plotted. The two closely spaced sinusoids are resolved, but their respective magnitude is divided between several dictionary elements. As a comparison, the ℓ_2-regularized estimate, i.e., the Tikhonov Regularization (TR) or ridge regression estimate, is also plotted. The TR is a smoothing estimate used when the solution of a linear system is not unique, and so solves

$$\underset{\mathbf{a}}{\text{minimize}}\ \frac{1}{2}||\mathbf{y} - \mathbf{W}\mathbf{a}||_2^2 + \varkappa||\mathbf{a}||_2^2 \tag{27}$$

which has the closed form solution

$$\hat{\mathbf{a}}_{\text{TR}} = (\mathbf{W}^H\mathbf{W} + \varkappa\mathbf{I})^{-1}\mathbf{W}^H\mathbf{y} \tag{28}$$

Note that the addition of $\varkappa\mathbf{I}$ makes the solution stable, and the larger $\varkappa$ becomes, the more $\hat{\mathbf{a}}_{\text{TR}}$ resembles a scaled version of the DFT estimate, which may also be seen from the figure. In spite of the LASSO's ambiguity in estimation of highly coherent dictionary atoms, it will generally have good super-resolution performance, and will typically cope with resolutions upwards of $5 \leq s \leq 10$ [9]. However, in terms of theoretical estimation guarantees, the results are quite pessimistic. There are several different methods of assessing the suitability of a dictionary for sparse estimation, which include the exact recovery coefficient (ERC) [10], the

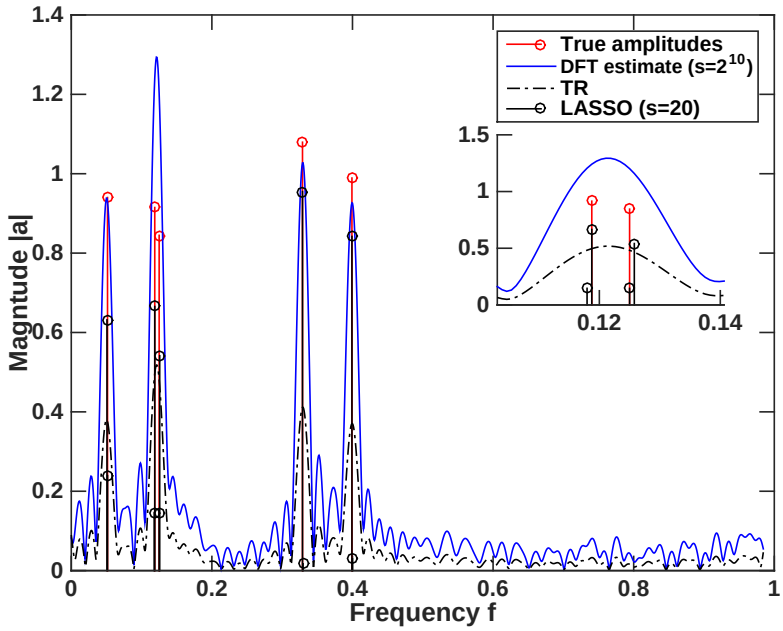


Figure 6: The LASSO, TR, and DFT amplitude estimates for $K = 5$ sinusoids in white Gaussian noise, where two of them are spaced by 2/5N, as compared to the true amplitudes.

spark [11], the two restricted isometry criteria (RIC)[1] [12], and the two coherence measures in [10] (cumulative coherence) and in [11] (mutual coherence). As noted in [13], only the latter two can be readily calculated for an arbitrary dictionary. Focusing on the mutual coherence, it is defined as the maximum linear dependence present in the dictionary, which for line spectra becomes

$$\mu(\mathbf{W}) \triangleq \max_{f_p \neq f_q} \frac{1}{N} \left| h(f_p, f_q) \right| \tag{29}$$

for $f_p, f_q \in \mathbf{\Psi}$. The theoretical implications for mutual coherence in line spectra was examined in [14], where it is claimed that a sufficient condition for for robust

[1] The two RICs are the well-known restricted isometry property (RIP) and the restricted orthogonality property (ROP), respectively.

recovery is that $\mu \leq \sqrt{2} - 1$. In contrast to practical observations this would thus correspond to a minimal grid point spacing of approximately $|f_p - f_q| \geq 2N/3$, and so robust recovery in this sense is not possible for super-resolution dictionaries, i.e., for $s > 1$. Except for super-resolution, another issue affecting performance is off-grid effects. Re-examining Figure 5, it is apparent that the LASSO does not robustly recover the correct number of frequency components, i.e., $||\hat{\mathbf{a}}||_0 \neq K$, even if an orthogonal dictionary is used. In spite of this, it has been found that if choosing the largest peaks of the LASSO estimate, rather than all non-zero parameters, sparse modeling works well for line spectra in practice [9]. In addition, the estimation may be further improved by using the LASSO estimates as an initial solution to the NLS method.

2.4 Choosing the level of regularization

The KKT conditions in section 2.2 give a qualitative understanding of the effect of λ. Thus, from (24), it is clear that any component p with an unconstrained estimate $|\breve{a}_p| \leq \lambda$ will have the LASSO solution $\hat{a}_p = 0$. The choice of λ can be therefore be seen as an implicit choice of the model order. In its essence, λ should be chosen such that any noise peak in the residual spectra should be zero, while any signal peak is still resolved. This is accomplished by setting λ larger than the largest noise peak, and lower than lowest signal peak. As a consequence, for signals with very low SNR, where some of the noise peaks may be larger than the signal peaks, the sparse modeling approach may either include noise peaks in, or exclude signal peaks from, the estimates. Choosing λ thus becomes very difficult, and some compromise must be done. In this section, some different approaches for setting the level of regularization will be presented, where the signal is assumed to have a spectral representation where the signal resides above the noise floor. For white Gaussian noise, the unconstrained LS estimate has the statistical properties

$$\hat{\mathbf{a}}_{\mathrm{LS}} \sim \mathcal{N}(\mathbf{a}, \sigma^2(\mathbf{W}^H\mathbf{W})^{-1}) \tag{30}$$

and for a noise-only signal, i.e., $K = 0$, implying that $\mathbf{a} = \mathbf{0}$, the LASSO solution should also be $\hat{\mathbf{a}} = \mathbf{0}$. If σ^2 is known, this may be ensured with probability $1 - \alpha$, if λ is chosen such that

$$\lambda = \left\{ \lambda : \mathrm{Prob}\left(\max_p |a_{LS,p}| \leq \lambda \right) = (1 - \alpha) \right\} \tag{31}$$

i.e., such that the magnitude of the largest LS parameter estimate is smaller than λ with probability $1-\alpha$. For an orthogonal dictionary, i.e. $P = N$ for line spectra, the LS estimates become uncoupled, and (31) becomes equivalent to the maximum of N independent and identically distributed variables, i.e.,

$$\lambda = \left\{ \lambda : \text{Prob}\left(\xi \leq \frac{\lambda^2 N}{\sigma^2} \right)^N = (1-\alpha) \right\} \tag{32}$$

where $\xi = \breve{a}_p^2/(\sigma^2/N) \sim \chi^2(1)$, as the square of a standard Gaussian random variable is χ^2 distributed with one degree of freedom. Solving for λ, one obtains

$$\lambda = \sigma \sqrt{\frac{1}{N} F_\xi^{-1}\left((1-\alpha)^{1/P} \right)} \tag{33}$$

where F_ξ^{-1} is the inverse of the χ^2 distribution function. In the examples of this section, this approach was used, with $\alpha = 0.001$. One may also think of the solution in terms of dynamic range, especially if the noise level is very low. Thus, one may decide upon a dynamic range of δ (dB), such that if the signal is normalized by its largest spectral line, the regularization becomes

$$\lambda = \sqrt{10^{-\delta/10}} \tag{34}$$

which implies that the maximal dynamic range, i.e., difference in signal power between two components, is $|\delta|$ dB. For example, for $\delta = 20$ dB, this yields $\lambda = 0.1$. Another approach, suggested in [15], is to evaluate the solution for different levels of λ, i.e., $\hat{\mathbf{a}}(\lambda)$, and thereafter evaluate the solutions by some model order criteria. Of course, this requires solving the optimization problem several times, increasing the computational burden. A way of circumventing this was proposed in [16], which claims to (for real-valued signals) solve the entire solution path of $\hat{\mathbf{a}}(\lambda)$ with the same computational complexity as if solving for a single λ. This is an interesting idea, although not further discussed herein.

3 Sparse estimation of grouped line spectra

A natural extension of the LASSO methodology is to account for a grouping or clustering behavior of the dictionary atoms. Thus, a subset of dictionary atoms may be assigned to a cluster which, if active, assumes that all atoms in the cluster

are active, i.e., has non-zero parameters. For spectral estimation, this may be the case if a certain signal source contains multiple sinusoids, which share a predetermined relationship. Thus, the estimation procedure may be simplified into only searching for a single frequency, e.g., the fundamental frequency of a pitch signal, from which the frequencies of the other components may directly be obtained. This has the statistical benefits of reducing the degrees of freedom, and thus increasing the precision of the estimate. Another advantage of using sparse modeling is that it alleviates the burden of precise model order knowledge. This may be especially beneficial for grouped line spectra, where several groups are superimposed in a signal. For such signals, a combinatorial issue of determining which components belong to which group arise, to which sparse modeling is a good remedy, as it only claims that the number of groups should be as few as possible.

3.1 The Group LASSO

The principal methodology for grouping of dictionary atoms, introduced in [17], is called the Group LASSO, and is a simple extension of the LASSO to where clusters are pre-defined; to show this, a slight change of notation is required. Therefore, let

$$\mathbf{\Psi} = \left\{ \left\{ \psi(f_p, \ell) \right\}_{\ell=1,\ldots,L_p} \right\}_{p=1,\ldots,P} \tag{35}$$

be the set of candidate frequencies represented in the dictionary, divided into P groups of L_p atoms in each, and where $\psi(f, \ell)$ is a known function of the fundamental frequency, f, and the ℓ:th component. In the same manner, let the dictionary $\mathbf{W}$ be structured into P sub-dictionaries, each consisting of L_p atoms, such that

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & \ldots & \mathbf{W}_P \end{bmatrix} \tag{36}$$

$$\mathbf{W}_p = \begin{bmatrix} \mathbf{w}_{p,1} & \ldots & \mathbf{w}_{p,L_p} \end{bmatrix} \tag{37}$$

$$\mathbf{w}_{p,\ell} = \begin{bmatrix} e^{i2\pi\psi(f_p,\ell)1} & \ldots & e^{i2\pi\psi(f_p,\ell)(N-1)} \end{bmatrix}^T \tag{38}$$

where $\mathbf{W}_p$ is the sub-dictionary corresponding to the candidate fundamental frequency f_p. Furthermore, let $\mathbf{a}$ be structured similarly, i.e.,

$$\mathbf{a} = \begin{bmatrix} \mathbf{a}_1 & \ldots & \mathbf{a}_P \end{bmatrix}^T \tag{39}$$

$$\mathbf{a}_p = \begin{bmatrix} {a_{p,1}}^T & \ldots & {a_{p,L_p}}^T \end{bmatrix}^T. \tag{40}$$

The Group LASSO may thus be defined as the solution to the convex optimization problem

$$\underset{\mathbf{a}}{\text{minimize}}\ \frac{1}{2}||\mathbf{y}-\mathbf{W}\mathbf{a}||_2^2+N\lambda\sum_{p=1}^{P}\sqrt{L_p}||\mathbf{a}_p||_2 \tag{41}$$

where instead of using the ℓ_1-norm for the entire variable vector, the group sparse penalty is the equivalent of taking the ℓ_1-norm on a vector consisting of ℓ_2-norms for each group. This implies that sparsity is enforced for the entire group, whereas no sparsity is imposed for the components within a group. The Group LASSO is also a convex optimization problem, and to obtain a qualitative understanding of the effect of the penalty and of λ, a closed form solution is derived using the KKT conditions, which state (for the real-valued case) that

$$\frac{\partial f}{\partial \mathbf{a}_p}=-\mathbf{W}_p^T\left(\mathbf{y}-\mathbf{W}\mathbf{a}\right)+N\lambda\sqrt{L_p}\mathbf{u}_p=\mathbf{0},\quad\forall\mathbf{a}_p \tag{42}$$

$$\mathbf{u}_p=\begin{cases}\frac{\mathbf{a}_p}{||\mathbf{a}_p||_2} & \forall\mathbf{a}_p\neq\mathbf{0}\\ \in\{[\mathbf{u}:||\mathbf{u}||_2\leq 1\} & \forall\mathbf{a}_p=\mathbf{0}\end{cases} \tag{43}$$

where $\mathbf{u}_p$ is the sub-gradient of the group sparse penalty. Similar to the LASSO, let $\mathbf{W}_{-p},\mathbf{a}_{-p}$ denote the dictionary and the variable vector where the p:th group is left out. With some linear algebra, (42) is equal to

$$(\mathbf{W}_p^T\mathbf{W}_p)\mathbf{a}_p=(\mathbf{W}_p^T\mathbf{W}_p)\breve{\mathbf{a}}_p-\lambda\sqrt{L_p}N\mathbf{u}_p \tag{44}$$

where, similarly to (22), $\breve{\mathbf{a}}_p$ denotes the unconstrained solution;

$$\breve{\mathbf{a}}_p\triangleq(\mathbf{W}_p^T\mathbf{W}_p)^{-1}\mathbf{W}_p^T\left(\mathbf{y}-\mathbf{W}_{-p}\mathbf{a}_{-p}\right) \tag{45}$$

Using some matrix algebra, one obtains

$$\mathbf{a}_p=\left(\mathbf{W}_p^T\mathbf{W}_p+\frac{\lambda\sqrt{L_p}N}{||\mathbf{a}_p||_2}\mathbf{I}\right)^{-1}(\mathbf{W}_p^T\mathbf{W}_p)\breve{\mathbf{a}}_p \tag{46}$$

In this form, a solution of the Group LASSO is quite difficult to obtain, as $\mathbf{a}_p$ also occurs in the right hand side of the equation. However, if assuming that the sub-dictionary for the p:th group is orthogonal, i.e., $\mathbf{W}_p^T\mathbf{W}_p=N\mathbf{I}$, one may show

that

$$\lambda\sqrt{L_p} \geq ||\breve{\mathbf{a}}_p||_2 \qquad \forall \mathbf{a}_p = \mathbf{0} \tag{47}$$

$$||\mathbf{a}_p||_2 = ||\breve{\mathbf{a}}_p||_2 - \lambda\sqrt{L_p} \geq 0 \qquad \forall \mathbf{a}_p \neq \mathbf{0} \tag{48}$$

which, if inserted into (46), yields the closed form solution for the Group LASSO as

$$\hat{\mathbf{a}}_p = \left(1 - \frac{\lambda\sqrt{L_p}}{||\breve{\mathbf{a}}_p||_2}\right)_+ \breve{\mathbf{a}}_p \tag{49}$$

and therefore also fulfills the KKT conditions. Thus, it becomes apparent that the Group LASSO does not promote sparsity within groups, for each component. It instead makes the solution group-wise sparse, by soft-thresholding all components in proportion to the ℓ_2 norm of the group, putting the entire group to zero if $||\breve{\mathbf{a}}_p||_2 \leq \lambda\sqrt{L_p}$. Lastly, for the Group LASSO, some remarks may be noted:
Remark 1: The purpose of $\sqrt{L_p}$ is to make λ for the Group LASSO comparable in dimension to that of the LASSO, as, e.g., if $\breve{a}_{p,1} = \cdots = \breve{a}_{p,L_p} = a$, then $||\mathbf{a}_p||_2 = \sqrt{L_p}|a|$ and the entire group is set to zero if $|a| \leq \lambda$. It is also worth noting that, for $L_1 = \cdots = L_P = 1$, the regular LASSO is obtained.
Remark 2: For the Group LASSO, recovery guarantees can be formulated in a manner similar to the LASSO, as in, e.g., [18, 19], but shows a similar degree pessimism. In its essence, robust recovery, in the sense of only allowing non-zero groups for the true spectral lines, is only possible if $\mathbf{W}^T\mathbf{W} = N\mathbf{I}$. However, in practice, the Group LASSO is quite robust to coherency both in, and between, groups.
Remark 3: The closed-form solution of the LASSO with complex-valued variables, i.e., $\mathbf{a} \in \mathbb{C}^{P\times 1}$, is obtained using the Group LASSO. By modeling the real and imaginary parts of the variables separately in a real-valued response variable, a group sparse penalty must be used to ensure the correct sparsity structure, where $\{\mathrm{Re}(a_p), \mathrm{Im}(a_p)\}$ defines the p:th group.

3.2 Other variations on the LASSO

One may customize the penalty functions in order to address a specific sparse structure, to which there exist a multitude of different approaches, some of which

will be described here. In general, all of them solve a convex optimization problem on the form

$$\underset{\mathbf{a}}{\text{minimize}}\ \frac{1}{2}\,||\mathbf{y}-\mathbf{W}\mathbf{a}||_2^2 + N\sum_{j\in\mathcal{J}} g_j(\mathbf{a},\lambda_j) \tag{50}$$

where g_j denotes the j:th penalty function, enforcing a certain sparsity structure, with λ_j denoting its corresponding regularization parameter, which weights the importance between the sparsity promoted by g_j and the model fit. What separates different methods, is this which functions that are included in the set $\mathcal{J}$. In [20], Simon et al. introduce the Sparse Group LASSO (SGL), in which the representation of the signal, $\mathbf{a}$, may have sparsity also within groups. Such a sparsity pattern is enforced by combining (18) and (41), thereby regularizing the solution with two penalty functions, i.e.,

$$g = \left\{\ \lambda_1||\mathbf{a}||_1,\quad \lambda_2\textstyle\sum_{p=1}^{P}\sqrt{L_p}||\mathbf{a}_p||_2\ \right\} \tag{51}$$

Another variation is the Generalized LASSO, introduced in [21], which use a penalty function on the form

$$g = \left\{\ \lambda||\mathbf{D}\mathbf{a}||_1\ \right\} \tag{52}$$

where $\mathbf{D}$ is a linear transformation matrix, such that the ℓ_1-norm is imposed on a linear combination of the components in $\mathbf{a}$. A popular choice of $\mathbf{D}$ is the first-order difference matrix, defined as

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 & -1 \end{bmatrix} \tag{53}$$

which yields the difference between every two adjacent parameters. Commonly, this function is used in combination with the standard ℓ_1-norm, i.e.,

$$g = \left\{\ \lambda_1||\mathbf{a}||_1,\quad \lambda_2\textstyle\sum_{p=2}^{P}|a_p - a_{p-1}|\ \right\} \tag{54}$$

where $\mathbf{a}$ is indexed for the normal LASSO, i.e., as in (17)-(18). This method is called the Sparse Fused LASSO (SFL), introduced in [22]. The SFL has a

grouping effect, where if adjacent dictionary components have similar energy, they are fused into groups without a pre-defined structure. Simultaneously, if components are too weak, they are soft-thresholded to zero. Thus, SFL enforces both grouping and sparsity.

3.3 Choosing the levels of regularization

As detailed in section 2.4, the level of sparsity is implicitly selected by setting the λ-parameter. Similarly to the LASSO, the solution of the Group LASSO for the p:th group is set to zero if $||\breve{\mathbf{a}}_p|| \leq \lambda\sqrt{L_p}$. An expression similar to (33) may thus be derived for the Group LASSO. If the power of the noise is unknown, as it most often is, one may also examine the unconstrained solution $\breve{\mathbf{a}}_p$, or an approximation of it, and thereafter set λ larger than the perceived grouped noise floor. Alternatively, the idea of dynamic range, described earlier for the LASSO, can also be applied for the group LASSO. Calculating the solution path $\hat{\mathbf{a}}(\lambda)$ over an interval of different λ:s, and thereafter choosing the best according to some model order criteria, is also a viable approach [17]. However, as the number of penalty functions increase, so does the complexity in choosing the levels of regularization. Not only does the total penalty cost need to balance the model fit, but each of the penalties also needs to be weighed against each other, as to find the sought sparsity pattern. A standard method for tuning the sparse penalty parameters is to use cross-validation, but for most purposes, this is impractical, as the computational complexity would be very high. However, there are some simple heuristics which makes the tuning process quite manageable. One approach which usually works well is to tune λ_j for the j:th penalty independently of the other penalties, using some of the methods previously described. However, as the penalties are added together, the total level of regularization will be too high, which might remove components in the signal, if the amount of headroom above the noise level is small. Another approach is to tune one penalty at a time, in a prioritized order. For instance, with the SGL, one may tune the group penalty first, as to find the appropriate number of blocks. Then, one may tune the ℓ_1 penalty to find the appropriate sparsity pattern within the active groups. In practice, when the signal is much stronger than the noise, the solution is often quite insensitive to the choice of λ.

4 Solving convex programs

A major reason why the optimization problems described in this thesis are cast as convex, besides from the favorable theoretical properties described above, is that there exists a solid framework for finding numerical solutions to convex problems. Using the methodology of disciplined convex programming described in [23], the corresponding software package, CVX [24], makes implementation very approachable. Thus, new prototype methods, with novel penalty functions that infer certain structure on the solution, may be experimented with in a straightforward manner. CVX uses commonly available interior point methods such as SeDuMi [25] and SDPT3 [26], for instance, to find solutions which approximately fulfill the KKT conditions of the corresponding convex problem. However, CVX is in general too computationally burdensome for practical estimation of the optimization problems considered in this thesis, to which there exist some alternative implementations. For the LASSO-type of problems, two such approaches are the ADMM framework used in [27], and the Least Angle RegreSsion (LARS) [16], where the latter uses a Cyclic Coordinate Descent (CCD) approach. The methods described in this paper are efficiently implemented using the ADMM, which the next subsection describes in more detail.

4.1 Outline of the ADMM

In general, ADMM solves problems of the form

$$\underset{\mathbf{z}}{\text{minimize}} \quad f_1(\mathbf{z}) + f_2(\mathbf{G}\mathbf{z}) \tag{55}$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are closed, proper, and convex functions, and $\mathbf{G}$ is a known matrix. By introducing the new variable $\mathbf{u} = \mathbf{G}\mathbf{z}$, and adding this condition to the optimization problem, the ADMM approach is to iterate between solving for $\mathbf{z}$, while keeping $\mathbf{u}$ constant, and vice versa. The problem (55) may thus be equivalently expressed as [27]

$$\begin{aligned} &\underset{\mathbf{z}}{\text{minimize}} \, f_1(\mathbf{z}) + f_2(\mathbf{u}) + \frac{\mu}{2} ||\mathbf{G}\mathbf{z} - \mathbf{u}||_2^2 \\ &\text{subject to } \mathbf{G}\mathbf{z} - \mathbf{u} = \mathbf{0} \end{aligned} \tag{56}$$

for any smoothing parameter μ, as the penalty term disappears when the constraint is fulfilled. To solve this convex program, the augmented Lagrangian for

the scaled form ADMM [27, p. 15] is formed as

$$L_\mu(\mathbf{z}, \mathbf{u}, \mathbf{d}) = f_1(\mathbf{z}) + f_2(\mathbf{u}) + \frac{\rho}{2}||\mathbf{G}\mathbf{z} - \mathbf{u} + \mathbf{d}||_2^2 \tag{57}$$

where $\mathbf{d}$ denotes the scaled dual variable. At iteration $(k+1)$, the parameters are obtained by solving

$$\mathbf{z}^{(k+1)} = \arg\min_{\mathbf{z}} L_\mu(\mathbf{z}, \mathbf{u}^{(k)}, \mathbf{d}^{(k)}) \tag{58}$$

$$\mathbf{u}^{(k+1)} = \arg\min_{\mathbf{u}} L_\mu(\mathbf{z}^{(k+1)}, \mathbf{u}, \mathbf{d}^{(k)}) \tag{59}$$

and then updating the scaled version dual variable as

$$\mathbf{d}^{(k+1)} = \mathbf{d}^{(k+1)} - \mu(\mathbf{G}\mathbf{z}^{(k+1)} - \mathbf{u}^{(k+1)}) \tag{60}$$

Clearly, using the ADMM optimization scheme is worthwhile when (58)-(59) are such that they may be carried out much easier than the original problem in (55). For the LASSO, this is precisely the case, as will be shown in the next section.

4.2 LASSO via ADMM

To solve the LASSO using an ADMM approach, consider an augmented optimization problem equivalent to the one in (18), i.e.,

$$\begin{aligned} &\underset{\mathbf{a},\mathbf{b}}{\text{minimize}}\ \frac{1}{2}||\mathbf{y} - \mathbf{W}\mathbf{a}||_2^2 + N\lambda||\mathbf{b}||_1 + \frac{\mu}{2}||\mathbf{a} - \mathbf{b}||_2^2 \\ &\text{subject to } \mathbf{a} - \mathbf{b} = \mathbf{0} \end{aligned} \tag{61}$$

to which the augmented Lagrangian for the scaled form ADMM may be expressed as

$$L_\mu(\mathbf{a}, \mathbf{b}, \mathbf{d}) = \frac{1}{2}||\mathbf{y} - \mathbf{W}\mathbf{a}||_2^2 + N\lambda||\mathbf{b}||_1 + \frac{\mu}{2}||\mathbf{a} - \mathbf{b} + \mathbf{d}||_2^2 \tag{62}$$

such that $\mathbf{d}$ denotes the scaled dual variable. To find the expressions which minimize (62) with respect to $\mathbf{a}$ and $\mathbf{b}$, similar to (58) and (59), one must differantiate the Langrangian, set the derivative to zero, and solve for the current variable at iteration $k+1$. For $\mathbf{a}$, this yields an expression similar to the TR estimate in (28), i.e.,

$$\mathbf{a}^{(k+1)} = \left(\mathbf{W}^H\mathbf{W} + \mu\mathbf{I}\right)^{-1}\left(\mathbf{W}^H\mathbf{y} + \mu\left(\mathbf{b}^{(k)} - \mathbf{d}^{(k)}\right)\right) \tag{63}$$

For $\mathbf{b}$, the Lagrangian is non-differentiable, due to the ℓ_1 penalty. However, notice that that the two terms which depend on $\mathbf{b}$ resembles a simplified version of the LASSO, where the parameters in $b_p,\ p = 1, \ldots, P$ are uncoupled, and may thus be obtained using the closed-form expression

$$b_p^{(k+1)} = \left(1 - \frac{N\lambda}{\mu|\breve{b}_p^{(k+1)}|}\right)_+ \frac{\breve{b}_p^{(k+1)})}{|\breve{b}_p^{(k+1)}|} \tag{64}$$

where

$$\breve{b}_p^{(k+1)} = a_p^{(k+1)} + d_p^{(k)} \tag{65}$$

is formed from the p:th elements of the vectors $\mathbf{a}^{(k+1)}$ and $\mathbf{d}^{(k)}$, respectively. Finally the dual variable is updated as in (60), with $\mathbf{G} = \mathbf{I}, \mathbf{z} = \mathbf{a}$, and $\mathbf{u} = \mathbf{b}$.

5 Preliminaries for selected applications

This thesis deals with two specific applications for sparse modeling, namely audio and array processing. In this section, some preliminary assumptions from both of these topics are presented, which are then treated in greater detail in the later part of the thesis.

5.1 Parametric Modeling of Audio

In modern audio processing, one primarily deals with the digital representation of sound waves, i.e., longitudinal waves where a medium[2] is compressed and decompressed. The bulk of research over the last decades has been focused on speech processing, as to fill the emerging need of solutions for digital communication (see, e.g., [28], and the references therein) However, in more recent years, much research in audio processing has also been devoted to musical signals, quite unsurprisingly, given the large role of digital media in everyday life (for an overview, see, e.g., [29]). Combined, the two fields of speech and music processing are formidably vast, and they cannot possibly be given any form of justice in merely a few pages. Instead, some brief excerpts are given, as to give some context to the methods of which this thesis consist. For both fields, given the nature of sound,

[2] Sounds in air are typically recorded using microphones, but sounds in water are also often considered, which are then recorded by hydrophones.

signals are periodic and, for our purposes, their spectral representations are highly relevant. Many audio signals are well described as narrowband, i.e., the spectral energy is largely limited to a few very limited intervals on the frequency axis. As a result, parametric estimation using the sinusoidal model is often a good approach to quantifying the properties of speech and music. A common model used for voiced speech and tonal music is the harmonic model, or pitch model, which is of the form [30]

$$y(t) = x(t) + e(t),\ x(t) = \sum_{\ell=1}^{L} a_\ell e^{i2\pi f \ell t} \tag{66}$$

for a single pitch. Measured in some form of additive noise, $e(t)$, the pitch signal, $x(t)$, consists of a group of complex-valued[3] sinusoids, whose relation are described, using the notation from section 3, as

$$\psi(f, \ell) = f\ell, \quad \ell \in \mathcal{L} \tag{67}$$

where the frequency components are thus equally spaced, by f, on the frequency axis, for all indices of ℓ in an index set $\mathcal{L}$. Typically, a pitch is defined by its fundamental frequency, i.e., $\psi(f, 1) = f$, and the individual sinusoids are referred to as its harmonics. A common misconception is that the fundamental is always the lowest frequency in the pitch, which is obviously only true if $1 \in \mathcal{L}$. This is, however, not always the case, as some harmonics may be missing, including the fundamental. Instead, it is in most cases better to view the fundamental frequency as the smallest distance between two adjacent harmonics in a pitch group. Thus, if a certain pitch f has the following set of harmonics,

$$\mathcal{L} = \{2, 4, 6, 8, \ldots, 2L\} \tag{68}$$

it may preferably be seen as a pitch with fundamental frequency $f' = 2f$, and corrsponding set $\mathcal{L}' = \{1, 2, 3, 4, \ldots, L\}$ of harmonics. As there might be ambiguities as how to chose f and $\mathcal{L}$, such as, e.g., the example given above, the basic assumption, which is extensively used in the thesis, is that the spectral envelope of the pitch should be smooth [31], i.e., that adjacent harmonics should be of

[3] Naturally, audio signals are not complex-valued. However, by using the analytic representation of the real-valued signals, both analysis and estimation may be greatly simplified. This is mainly because real-valued signals contain two spectral lines for every frequency f present in the signal, located at $\pm f$, where the negative component is removed in the analytic signal.

comparable magnitude. This is obviously not the case for the pitch described in (68), as its uneven harmonics have zero magnitude. Promoting such smoothness to avoid ambiguity is one of the objectives of paper A. Another common property for harmonic audio signals, in particular for some musical instruments, is a slight, but systematic, deviation from even distances between harmonics. This is referred to as inharmonicity, which for stringed instruments may be well described as

$$\psi(f,\ell) = f\ell\sqrt{1+\ell^2 B}, \quad \ell \in \mathcal{L} \tag{69}$$

where B is called the inharmonicity coefficient, specific to each string; typically $B \in [10^{-5}, 10^{-3}]$ [32]. Another feature of audio, highly related to pitch and especially used in musical contexts, is chroma. Chroma is the representation of the fundamental frequency frequency on a cyclical scale. To that end, consider the chroma parameter $c \in [0,1)$, to which the corresponding fundamental frequencies may be expressed as

$$f = f_{\text{base}} 2^{c+m}, \quad \forall m \in \mathbb{Z} \tag{70}$$

where m is referred to the octave, and where f_b is a tuning or offset frequency, defining the specific location of a chroma in frequency. This implies that the linear frequency scale collapses into a cyclic chroma scale, as all fundamental frequencies which fulfill (70), for some integer o, belong to the same chroma, i.e., if $f \in c$, then

$$f' \in c \Rightarrow f' \in \left\{ \ldots, \frac{f}{8}, \frac{f}{4}, \frac{1}{2}, f, 2f, 4f, 8f, 16f, \ldots \right\} \tag{71}$$

and all fundamentals in a chroma are thus related by some power of 2. One benefit of the chroma representation is that it groups together pitches that have largely overlapping frequency content, which makes chroma estimation much less ambiguous than pitch estimation. In music, the chroma representation is a common grouping criterion, as all pitches in a chroma are perceived as similar by human hearing [29]. In the Western musicological system, for instance, the chroma interval is discretized into twelve semitones, uniformly spaced on $[0,1)$, i.e.,

$$c \in \left\{ 0, \frac{1}{12}, \frac{2}{12}, \ldots, \frac{11}{12} \right\} \tag{72}$$

In paper D, the chroma model for Western music is used with sparse modeling to form estimates, cruder than pitch but more robust, of the spectral components of an audio signal.

5.2 Spatial modeling of sinusoids

In the field of array processing, a common objective is to locate signal emitting sources by measuring their emissions over an array of sensors. The emitted energy may be of various types, e.g., acoustic or electromagnetic, to which different corresponding types of sensors are used. In this section, some basic results for source localization is given as to facilitate a bit of background for the methods presented in paper B and C. In general, the objective may be put as finding the distribution of energy in the spatial domain. If assuming that all sensors have the same gain, the signal model for the impinging source signal at the m:th sensor may be expressed as

$$y_m(t) = x(t - \tau_m) + e_m(t) \tag{73}$$

where τ_m is the source-sensor time-delay with respect to some reference point, such that the source signal $x(t)$ is at each sensor delayed with respect to the specific geometry of the array. Consider that $x(t)$ follows the sinusoidal signal model in (1). As such a signal is formed by a sum of narrowband components, the time-delay in (73) may typically be well modeled as a phase offset in each component, exponentially proportional to its frequency, i.e.,

$$y_m(t) = \sum_{k=1}^{K} a_k e^{i2\pi f_k(t-\tau_m)} + e_m(t) \tag{74}$$

which for the sample vector is equivalent to

$$\mathbf{y}_m = \sum_{k=1}^{K} \mathbf{w}_k a_k e^{-i2\pi f_k \tau_m} + \mathbf{e}_m, \tag{75}$$

using the same notation as in Section 1, but where $(\cdot)_m$ denotes the m:th sensor. By column-wise stacking the sample vectors for all sensors, i.e.,

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 & \dots & \mathbf{y}_M \end{bmatrix} \tag{76}$$

the signal model for the entire array may be expressed as

$$\mathbf{Y} = \sum_{k=1}^{K} \mathbf{w}_k a_k \mathbf{z}^T + \mathbf{E} = \mathbf{W}\,\mathrm{diag}(\mathbf{a})\mathbf{Z}^T + \mathbf{E} \tag{77}$$

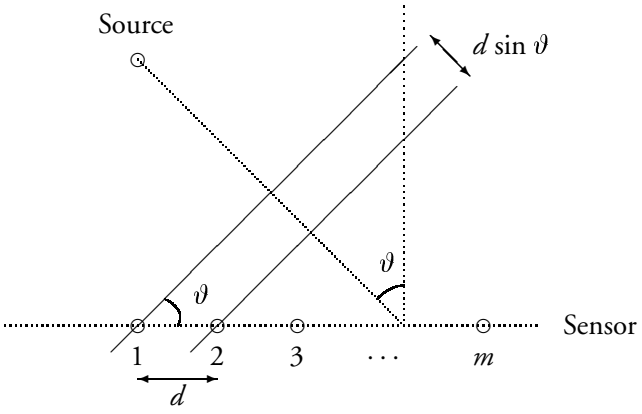


Figure 7: Principle sketch of a far-field point-source, which from ϑ emit planar wavefronts, that are impinging on a ULA, with equidistant sensor spacing d.

where $\mathbf{W}$ and $\mathbf{a}$ are defined as in Section 1, where $\mathbf{E}$ is the noise matrix defined similarly to (76), and where

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \dots & \mathbf{z}_K \end{bmatrix} \tag{78}$$

$$\mathbf{z}_k = \begin{bmatrix} e^{-i2\pi f_k \tau_1} & \dots & e^{-i2\pi f_k \tau_M} \end{bmatrix}^T \tag{79}$$

denote the phase offset for each sinusoidal component in each sensor, which depend on both the frequency and the time-delay. The time-delays are inherently related to both the source position and the geometry of the array, whose relation may be modeled by imposing some assumptions on the source, and the array, respectively. Two assumptions, which are very common for localization in array processing, are

- The source is a point source in the far-field, i.e., the source is at an infinite distance from the sensor array. This implies that the impinging signal wavefronts are essentially planar, so that a source's location solely depends on its Direction-Of-Arrival (DOA).
- The sensors are positioned as a Uniform Linear Array (ULA), meaning that they are equidistantly placed on a line. This implies that the positions will be defined to a 2-D space of locations, described by DOA and distance.

Figure 7 illustrates these two assumptions, where the DOA is denoted as the 1-D angular deviation from the array's normal, denoted $\vartheta \in [-\pi, \pi]$. Note that the ULA will not discriminate between a source impinging from the front or from the back of the array. From these assumptions, time-delays may thus be expressed as a function of DOA, i.e.,

$$\tau_m = \frac{d \sin(\vartheta)}{c}(m-1) \tag{80}$$

where d and c is the sensor distance, and the wave propagation speed, respectively. Therefore, (79) may be equivalently expressed as

$$\mathbf{z}_k = \left[\begin{array}{cccc} 1 & e^{-i2\pi f_k \frac{d \sin(\vartheta)}{c}} & \dots & e^{-i2\pi f_k \frac{d \sin(\vartheta)}{c}(M-1)} \end{array} \right]^T \tag{81}$$

where

$$\left| f_k \frac{d \sin(\vartheta)}{c} \right| \leq \frac{1}{2} \Rightarrow d \leq \frac{c}{2 f_k} \tag{82}$$

should be fulfilled as to guarantee that aliasing effects are avoided. For the far-field source and ULA case, $\mathbf{z}_k$ may thus be seen as a uniformly sampled spatial DFT vector. These preliminaries are utilized in paper C, where (74) is used to form a joint frequency and DOA estimator for the pitch model (66), when multiple pitches are impinging on an array of sensors, possibly from different locations. In paper B, the preliminaries presented herein are extended, and a joint multi-pitch and location estimator is proposed, for sources which are near-field rather than far-field, and when the array's geometry is arbitrary rather than uniformly linear.

6 Outline of the papers in this thesis

This section briefly summarizes the papers of which this thesis consist, together with information of where they have been published or submitted.

Paper A: An Adaptive Penalty Approach to Multi-pitch Estimation

In paper A, we propose a novel adaptive penalty approach to estimate the parameters in the multi-pitch model with the use of sparse modeling. We examine the PEBS/PEBS-TV methods introduced in [33], which formulate the problem

as a SGL, in which difficulties arise for pitch candidates at half of the true fundamental frequency (halflings). In PEBS-TV, an additional penalty function, based on the total variation cost, is introduced, which is shown to migitate such issues. However, this method requires tuning three regularization parameters, which we circumvent in this paper by using the adaptive approach, where a total variation penalty is efficiently utilized, which enables us to drop the group-LASSO penalty altogether. The method may thus be seen as solving a series of convex problems, where each is a SFL, having two tuning parameters. The strength of using total variation penalty compared to block-sparsity is that total variation promotes solutions which have smooth parameter envelopes, which discourages halflings, as they will have every other amplitude equal to zero. The method is shown to work well for highly coherent dictionaries, and even outperforms the method in [33].

The work in paper A has been published in part as

> Ted Kronvall, Filip Elvander, Stefan Ingi Adalbjörnsson, and Andreas Jakobsson, "An Adaptive Penalty Approach to Multi-pitch Estimation". *23rd European Signal Processing Conference*, Nice, France, August 31 - September 4, 2015.

Paper B: Sparse Localization of Harmonic Audio Sources

In paper B, we use a two-step procedure to form joint estimates of pitches and near- or far-field locations from measurements on an arbitrary, but calibrated, sensor array. In the first step, a SGL generalized for array signals is used to find the active pitches. Then, for estimated pitch, another variation on the SGL is used on the estimated parameters, which contain information of both TDOA and signal attenuation. This information is consequently exploited to form location estimates, which may be more than one for each pitch. The implications of using the sparse modeling approach is interesting, as it facilitates an opportunity to position sources despite of reverberation effects, which usually are detrimental to localization. The performance of the proposed method is validated using both synthetic and real recorded signals, showing promising results.

The work in paper B has been published/submitted in part as

> Stefan Ingi Adalbjörnsson, Ted Kronvall, Simon Burgess, Kalle Åström, and Andreas Jakobsson, "Sparse Localization of Harmonic Audio Sources". Submitted to *IEEE Transactions on Audio, Speech, and Language Processing*.

Ted Kronvall, Stefan Ingi Adalbjörnsson, and Andreas Jakobsson, "Joint DOA and Multi-pitch Estimation using Block Sparsity", *39th IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 4-9, 2014.

Paper C: Joint DOA and Multi-pitch Estimation via Block Sparse Dictionary Learning

In paper C, we introduce a dictionary learning approach of estimating the joint pitch and DOA estimates for an unknown number of pitch signals impinging on a ULA. The method builds on the spatial pitch model in paper B, but where each pitch source may only originate from a single direction. As the two parameters, frequency and DOA, are non-linear and intertwined in the signal model, the linearization with the Group-LASSO is not straight-forward. Instead, the DOA may be seen as phase offset, different for each sensor according to the specific array geometry, that may be learned using the dictionary learning framework, reminiscent to [34]. Thus, the method alternates between estimating the pitches present in the signal, using an extension of the SGL for array signals, and between learning the phase offset parameters governed by the DOA. The estimating procedures solves a series of convex problems, where each iteration improves the joint pitch and DOA estimate, until convergence.

The work in paper C has been published in part as

Ted Kronvall, Stefan Ingi Adalbjörnsson, and Andreas Jakobsson, "Joint DOA and Multi-pitch Estimation via Block Sparse Dictionary Learning", *22nd European Signal Processing Conference*, Lisbon, Portugal, September 1-5, 2014.

Paper D: Sparse Chroma Estimation for Harmonic Audio

In paper D, we take a different approach to the pitch estimation problem. Instead of focusing on estimating the parameter group of a specific pitch, we form groups of all pitches that belong to the same chroma group, which is all pitches that are at some power of 2 from each other. The chroma is a concept from musical theory, and transcribing a piece of audio with respect to its chroma content is a pre-processing step that is done for a variety of different MIR applications. In this paper, we propose a solution where we use a combination of group-sparsity

and total variation, such that the group-sparsity promotes solutions where few chroma blocks are active, and where total variation discourages misclassifications due to musical harmony, as the chroma groups have some partly overlapping frequency components. The method is numerically evaluated for a synthetic violin signal, which is known to be well modeled with grouped sinusoids, and indicates a preferred performance for transcription purposes. In this paper, we also allow the amplitude of each component to vary over time, which is modeled using a spline basis, where the number of spline knots are chosen according to preference. As this approach increases the number of parameters proportional to the number of knots, the method is especially suitable for longer sequences of data, where, for audio signals longer than 40 ms, the signal exhibits a large degree of non-stationarity. The approach may also be beneficial for sounds that are very transient, or for capturing the onset of a signal. We show that for a recorded violin signal, the proposed method estimates the signal envelope more accurately than for constant amplitudes, although at a higher computational cost.

The work in paper D has been published in part as

> Ted Kronvall, Maria Juhlin, Stefan Ingi Adalbjörnsson, and Andreas Jakobsson, "Sparse Chroma Estimation for Harmonic Audio", *40th International Conference on Acoustics, Speech, and Signal Processing*, Brisbane, Australia, April 19-24, 2015.
>
> Maria Juhlin, Ted Kronvall, Johan Swärd, and Andreas Jakobsson, "Sparse Chroma Estimation for Harmonic Non-stationary Audio", *23rd European Signal Processing Conference*, Nice, France, August 31 - September 4, 2015.
>
> Stefan Ingi Adalbjörnsson, Johan Swärd, Ted Kronvall, and Andreas Jakobsson, "A Sparse Approach for Estimation of Amplitude Modulated Sinusoids", *The Asilomar Conference on Signals, Systems, and Computers*, Asilomar, USA, November 2-5, 2014.

References

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*, Prentice-Hall, Englewood Cliffs, N.J., 1993.

[2] L. L. Scharf, *Statistical Signal Processing: Detection Estimation, and Time Series Analysis*, Addison-Wesley, New York, 1991.

[3] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Prentice Hall, Upper Saddle River, N.J., 2005.

[4] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society B*, vol. 58, no. 1, pp. 267–288, 1996.

[5] P. Stoica and A. Nehorai, "Statistical Analysis of Two Nonlinear Least-Squares Estimators of Sine-Wave Parameters in the Colored-Noise Case," *Circuits, Systems, and Signal Processing*, vol. 8, no. 1, pp. 3–15, 1989.

[6] P. Stoica, R. Moses, B. Friedlander, and T. Söderström, "Maximum Likelihood Estimation of the Parameters of Multiple Sinusoids from Noisy Measurements," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 378–392, March 1989.

[7] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From Sparse Solutions of Systems of Equations to Sparse Modelling of Signals and Images," *SIAM Review*, vol. 51, 2009.

[8] R.J. Tibshirani, "The Lasso Problem and Uniqueness," *Electronic Journal of Statistics*, vol. 7, no. 0, pp. 1456–1490, 2013.

[9] P. Stoica and P. Babu, "Sparse Estimation of Spectral Lines: Grid Selection Problems and Their Solutions," *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 962–967, Feb. 2012.

[10] J.A. Tropp, "Just Relax: Convex Programming Methods for Identifying Sparse Signals in Noise," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1030–1051, March 2006.

[11] D.L. Donoho, M. Elad, and V.N. Temlyakov, "Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 6–18, Jan 2006.

[12] E. J. Candes, J. Romberg, and T. Tao, "Robust Uncertainty Principles: Exact Signal Reconstruction From Highly Incomplete Frequency Information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[13] Z. Ben-Haim, Y.C. Eldar, and M. Elad, "Coherence-Based Performance Guarantees for Estimating a Sparse Vector Under Random Noise," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5030–5043, Oct 2010.

[14] J. Karlsson and L. Ning, "On Robustness of l1-Regularization Methods for Spectral Estimation," in *IEEE 53rd Annual Conference on Decision and Control*, Dec 2014, pp. 1767–1773.

[15] C. D. Austin, R. L. Moses, J. N. Ash, and E. Ertin, "On the Relation Between Sparse Reconstruction and Parameter Estimation With Model Order Selection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 560–570, 2010.

[16] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, April 2004.

[17] M. Yuan and Y. Lin, "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[18] Y. V. Eldar, P. Kuppinger, and H. Bolcskei, "Block-Sparse Signals: Uncertainty Relations and Efficient Recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.

[19] X. Lv, G. Bi, and C. Wan, "The Group Lasso for Stable Recovery of Block-Sparse Signal Representations," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1371–1382, 2011.

[20] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A Sparse-Group Lasso," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.

[21] R.J. Tibshirani and J. Taylor, "The Solution Path of the Generalized Lasso," *The Annals of Statistics*, vol. 39, no. 3, pp. 1335–1371, June 2011.

[22] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society B*, vol. 67, no. 1, pp. 91–108, January 2005.

[23] M. Grant, *Disciplined Convex Programming*, Ph.D. thesis, Information Systems Laboratory, Department of Electrical Engineering, Stanford University, 2004.

[24] Inc. CVX Research, "CVX: Matlab Software for Disciplined Convex Programming, version 2.0 beta," `http://cvxr.com/cvx`, Sept. 2012.

[25] J. F. Sturm, "Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11-12, pp. 625–653, August 1999.

[26] R. H. Tutuncu, K. C. Toh, and M. J. Todd, "Solving semidefinite-quadratic-linear programs using SDPT3," *Mathematical Programming Ser. B*, vol. 95, pp. 189–217, 2003.

[27] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[28] J. Benesty, M. Sondhi, M. Mohan, and Y. Huang, *Springer handbook of speech processing*, Springer, 2008.

[29] M. Müller, D. P. W. Ellis, A. Klapuri, and G. Richard, "Signal Processing for Music Analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, 2011.

[30] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Signal Processing*, vol. 88, no. 4, pp. 972–983, April 2008.

[31] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 11, no. 6, pp. 804–816, 2003.

[32] H. Fletcher, "Normal vibration frequencies of stiff piano string," *Journal of the Acoustical Society of America*, vol. 36, no. 1, 1962.

[33] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Multi-Pitch Estimation Exploiting Block Sparsity," *Elsevier Signal Processing*, vol. 109, pp. 236–247, April 2015.

[34] C. D. Austin, J. N. Ash, and R. L. Moses, "Dynamic Dictionary Algorithms for Model Order and Parameter Estimation," *IEEE Transactions on Signal Processing*, vol. 61, no. 20, pp. 5117–5130, October 2013.


A

Paper A

An Adaptive Penalty Approach to Multi-Pitch Estimation

Ted Kronvall, Filip Elvander, Stefan Ingi Adalbjörnsson, and Andreas Jakobsson

Centre for Mathematical Sciences, Lund University, Lund, Sweden

Abstract

This work treats multi-pitch estimation, and in particular the common misclassification issue wherein the pitch at half of the true fundamental frequency, here referred to as a sub-octave, is chosen instead of the true pitch. Extending on current methods which use an extension of the Group LASSO for pitch estimation, this work introduces an adaptive total variation penalty, which both enforce group- and block sparsity, and deal with errors due to sub-octaves. The method is shown to outperform current state-of-the-art sparse methods, where the model orders are unknown, while also requiring fewer tuning parameters than these. The method is also shown to outperform several conventional pitch estimation methods, even when these are virtued with oracle model orders.

Key words: multi-pitch estimation, block sparsity, adaptive sparse penalty, total variation, ADMM

1 Introduction

Pitch estimation, i.e., estimating the fundamental frequency of a group of harmonically related sinusoids, is a problem arising in a variety of fields, not least in audio processing. For example, correctly determining the pitches present in a signal is a fundamental building block in many music information retrieval applications, such as automatic music transcription and genre classification [1]. However, pitch estimation for multi-pitch signals is a difficult problem, and although notable efforts have been made to find reliable multi-pitch estimators, (see e.g. [2]), most of the currently available methods which use the harmonic structure depend on *a priori* model order information, i.e., knowing the number of pitches present, as well as the number of harmonic overtones for each pitch. Such information is in general notoriously difficult to obtain. Our approach is instead to solve the problem in a group sparse modeling framework, which allows us to avoid making explicit assumptions on the number of pitches, nor the number of harmonics. Instead, the number of components in the signal is chosen implicitly, by the setting of some tuning parameters. These tuning parameters determine how appropriate a given pitch candidate is to be present in the signal and may be set using some simple heuristics, or by using cross-validation.The sparse modeling approach has earlier been used for audio (see, e.g., [3]), and specifically for sinusoidal components in [4]. We extend on these works by exploiting the harmonic structure of the signals in a block sparse framework, where each block represents a candidate pitch. A similar method was introduced in [5], where block sparsity was enforced using block-norms, penalizing the number of active pitches. As the block-norm penalty, under some circumstances, cannot distinguish a true pitch from its sub-octave, i.e., the pitch with half of the true fundamental frequency, the method is also complemented by a total variation penalty, which is shown to solve such issues. Total variation penalties are often applied in image analysis to obtain block-wise smooth image reconstructions (see, e.g., [6]). For audio data, one can similarly assume that signals often are block-wise smooth, as the harmonics of a pitch are expected to be of comparable magnitude [7]. Enforcing this feature will specifically deal with octave errors (due to present sub-octaves), as, in the noise free case, only every other harmonic of the sub-octave will have non-zero power. In this paper, we show that a total variation penalty, in itself, is enough to enforce a block sparse solution, if utilized efficiently. More specifically, by making the penalty function adaptive, we may improve upon the convex approximation used in [5], allowing us to drop the block-norm penalty altogether, and so reduce the

number of tuning parameters. In some estimation scenarios, e.g., when estimating chroma using the approach in [8], this would simplify the tuning procedure significantly. Furthermore, we show that the proposed method performs comparably to that of [5], albeit with the notable improvement of requiring fewer tuning parameters. The method operates by solving a series of convex optimization problems, and as this class of problems generally are computationally cumbersome to solve, we present an efficient algorithm based on the alternating directions method of multipliers (ADMM); we refer to e.g. [9] for further details of the ADMM.

2 Multi-pitch signal model

Consider a complex-valued[1] signal consisting of K pitches, where the kth pitch is constituted by a set of L_k harmonically related sinusoids, defined by the component having the lowest frequency ω_k, such that

$$x(t) = \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} a_{k,\ell} e^{i\omega_k \ell t} \tag{1}$$

for $t = 1, \ldots, N$, where $\omega_k \ell$ is the frequency of the ℓth harmonic in the kth pitch, and with $a_{k,l}$ denoting its magnitude and phase. The occurrence of such harmonic signals is often in combination with non-sinusoidal components, such as for instance, colored broadband noise or non-stationary impulses. In the scope of this work, we only treat the narrowband components of the signal, although noting that audio signals often also contain other features of notable perceptual importance such as the signal's timbre. In general, selecting model orders in (8) is a daunting task, with both the number of sources, K, and the number of harmonics in each of these sources, L_k, being unknown, as well as often being structured such that different sources may have spectrally overlapping overtones. In order to remedy this, we propose a relaxation of the model onto a predefined grid of $P \gg K$ candidate fundamentals, each having $L_{\max} \geq \max_k L_k$, harmonics. Here, we chose the candidates so numerous and so finely spaced that the approximation

$$x(t) \approx \sum_{p=1}^{P} \sum_{\ell=1}^{L_{\max}} a_{p,\ell} e^{i\omega_p \ell t} \tag{2}$$

[1] For notational simplicity and computational efficiency, we here use the discrete-time analytical signal formed from the measured (real-valued) signal.

holds sufficiently well. We are only interested in such approximations where few, ideally K, of the fundamentals will have non-zero power, and so steps must be taken to ensure this sparse behavior of the to be estimated amplitudes $a_{p,l}$. This approach may be seen as a sparse linear regression problem reminiscent of [4] and has been thoroughly examined in the context of pitch estimation in, e.g., [5, 10, 11]. For notational convenience, we define the set of all amplitude parameters to be estimated as

$$\mathbf{\Psi} = \{\mathbf{\Psi}_{\omega_1}, \ldots, \mathbf{\Psi}_{\omega_P}\} \tag{3}$$

$$\mathbf{\Psi}_{\omega_k} = \{a_{k,1}, \ldots, a_{k,L_{\max}}\} \tag{4}$$

where, as described above, most $a_{k,\ell}$ in $\mathbf{\Psi}$ will be zero. It should be noted that the sparse pattern of $\mathbf{\Psi}$ will be group-wise, so that if a pitch with fundamental frequency ω_p is not present, then neither will any of its harmonics, i.e., $\mathbf{\Psi}_{\omega_p} = \mathbf{0}$. Furthermore, when a pitch is present, we may expect that not all $L_{\max}$ harmonics will be non-zero, but only the actual L_k ones. For candidate pitches at fractions of the present pitch, there will be a partial fit of its harmonics, which may render misclassification, which is a cause for errors, which occurs when a present pitch at ω_k may be perfectly modeled by a pitch at $\omega_k/2$ if $L_{\max} \geq 2L_k$, where then every other harmonic, i.e., $\ell = 2, 4, 6, \ldots, 2L_k$, are non-zero and the others equal to zero. To take these attributes into account and to avoid misclassifications, we propose the iterative approach detailed in the next section.

3 Block-sparse estimation using the total variation penalty

Considering a measured time-frame of the sought signal, we expect it to be corrupted by noise and perhaps other non-sinusoidal structure, i.e., $y(t) = x(t) + e(t)$, where $e(t)$ is such an additive broadband noise. In order to estimate the parameter set $\mathbf{\Psi}$, one often strives to minimize the squared residual cost function

$$g_1(\mathbf{\Psi}) = \frac{1}{2}\sum_{t=1}^{N}\left|y(t) - \sum_{p=1}^{P}\sum_{\ell=1}^{L_{\max}} a_{p,\ell}e^{i\omega_p\ell t}\right|^2 \tag{5}$$

where $|\cdot|$ denotes the absolute value. However, this function will not enforce said sparsity. As requiring exactly sparse solutions leads to combinatorially infeasible optimization problems, we herein adopt a convex modeling approach using

a number of convex cost functions. To discourage spurious harmonics, we introduce a constraint on the ℓ_1-norm of $\mathbf{\Psi}$ by

$$g_2(\mathbf{\Psi}) = \sum_{p=1}^{P} \sum_{\ell=1}^{L_{\max}} |a_{p,\ell}| \tag{6}$$

which is a convex approximation of the ℓ_0 penalty. Parameter estimation using a weighted sum of g_1 and g_2 is widely used in the literature, being referred to as the *lasso* [12]. Taking the block-wise sparse behavior described above into account, we further introduce

$$g_3(\mathbf{\Psi}) = \sum_{p=1}^{P} \sqrt{\sum_{\ell=1}^{L_{\max}} a_{p,\ell}^2} \tag{7}$$

which also is a convex function. The inner sum corresponds to the ℓ_2-norm, and does not enforce sparsity within each pitch, whereas instead the outer sum, corresponding to the ℓ_1-norm, enforces sparsity between pitches. Thereby, adding the $g_3(\mathbf{\Psi})$ constraint will penalize the number of non-zero pitches. However, if we for some p have $2L_p \leq L_{\max}$, the above penalties have no way of discriminating between the correct pitch candidate ω_p and the spurious sub-octave candidate $\omega_p/2$. However, as the sub-octave will only contribute to the harmonic signal at every other frequency in its block, one may reduce the risk of such a misclassification by further adding the penalty

$$\breve{g}_4(\mathbf{\Psi}) = \sum_{q=1}^{PL_{\max}-1} \Big| |a_{q+1}| - |a_q| \Big| \tag{8}$$

where the reparametrization is $q = (p-1)L_{\max} + \ell$, which would add a cost to blocks where there are notable magnitude variations between neighboring harmonics. Regrettably, (8) is not convex, but a simple convex approximation would be $\tilde{g}_4$, detailed as

$$\tilde{g}_4(\mathbf{\Psi}) = \sum_{q=1}^{PL_{\max}-1} |a_{q+1} - a_q| \tag{9}$$

which would be a good approximation of (8) if all the harmonics had the same phase. Clearly, this may not be the case, resulting in that the penalty in (9) would also penalize the correct candidate. An illustration of this is found by considering the worst-case scenario, when all the adjacent harmonics are completely out of phase and have the same magnitudes, i.e., $a_{p,\ell+1} = a_{p,\ell}e^{i\pi}$ with magnitude $|a_{p,\ell}| = r$, for $\ell = 1, \ldots, L_p - 1$. Then, the penalty in (9) will yield a cost of $\tilde{g}_4(\mathbf{\Psi}_{\omega_p}) = 2rL_p$ rather than the desired $\breve{g}_4(\mathbf{\Psi}_{\omega_p}) = 2r$. The cost may also be compared with that of (6), which is $g_2(\mathbf{\Psi}_{\omega_p}) = rL_p$, suggesting that this would add a relatively large penalty. More interestingly, for the sub-octave candidate, the cost will be just as large, i.e. if $\omega_{p'} = \omega_p/2$, then $\tilde{g}_4(\mathbf{\Psi}_{\omega_{p'}}) = 2rL_p$ provided that $L_{\max} \geq 2L_p$, thereby offering no possibility of discriminating between the true pitch and its sub-octave. Obviously, such a worst case scenario is just as unlikely as having all harmonics same-phased, if assuming that the phases are evenly distributed on $[0, 2\pi)$. Instead, the $\tilde{g}_4$ penalty of the true pitch will be slightly smaller than its sub-octave, on average, and together with (7), the scale tips in favour of the true pitch, as shown in [5]. We may thus conclude that the combination of g_3 and $\tilde{g}_4$ provides a block sparse solution where sub-octaves are usually discouraged. However, it should be noted that such a solution requires the tuning of two functions to control the block sparsity. In this work, we propose to simplify the described algorithm by improving the approximation in (9), by using an adaptive penalty approach. In order to do so, let $\varphi_{k,\ell}$ denote the phase of the component with frequency $\omega_{k,\ell}$ and collect these phases in the parameter set

$$\mathbf{\Phi} = \{\mathbf{\Phi}_{\omega_1}, \ldots, \mathbf{\Phi}_{\omega_P}\} \tag{10}$$

$$\mathbf{\Phi}_{\omega_k} = \{\varphi_{k,1}, \ldots, \varphi_{k,L_{\max}}\} \tag{11}$$

The penalty function in (9) may then be modified to

$$g_4(\mathbf{\Psi}, \mathbf{\Phi}) = \sum_{q=1}^{PL_{\max}} |a_{q+1}e^{-\varphi_{q+1}} - a_q e^{-\varphi_q}| \tag{12}$$

thus penalizing only differences in magnitude. In order to do so, the phases $\varphi_{k,\ell}$ need to be estimated as the arguments of the latest available amplitude estimates $a_{k,\ell}$. As a result, (12) yields an improved approximation of (8), avoiding the issues of (9) described above, and also promotes a block sparse solution. And so, the block-norm penalty function g_3 may be omitted, which simplifies the algorithm

noticeably. Thus, we form the parameter estimates by solving

$$\hat{\mathbf{\Psi}} = \arg\min_{\mathbf{\Psi}} \sum_{j=1,2} \lambda_j g_j(\mathbf{\Psi}) + \lambda_4 g_4(\mathbf{\Psi}, \mathbf{\Phi}) \tag{13}$$

where $\lambda_1 = 1$, and where λ_i, for $i = 2, 4$, are user-defined regularization parameters that weigh the importance of each penalty function and the residual cost. To form the convex criteria and to facilitate the implementation, consider the signal expressed in matrix notation as

$$\mathbf{y} = \begin{bmatrix} y(1) & \dots & y(N) \end{bmatrix}^T \tag{14}$$

$$= \sum_{p=0}^{P} \mathbf{W}_p \, \mathbf{a}_p + \mathbf{e} \triangleq \mathbf{W}\mathbf{a} + \mathbf{e} \tag{15}$$

where

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & \dots & \mathbf{W}_P \end{bmatrix} \tag{16}$$

$$\mathbf{W}_p = \begin{bmatrix} \mathbf{z}^1 & \dots & \mathbf{z}^{L_{\max}} \end{bmatrix} \tag{17}$$

$$\mathbf{z}_p = \begin{bmatrix} e^{i\omega_p 1} & \dots & e^{i\omega_p N} \end{bmatrix}^T \tag{18}$$

$$\mathbf{a} = \begin{bmatrix} \mathbf{a}_1^T & \dots & \mathbf{a}_P^T \end{bmatrix}^T \tag{19}$$

$$\mathbf{a}_p = \begin{bmatrix} a_{p,1} & \dots & a_{p,L_{\max}} \end{bmatrix}^T \tag{20}$$

The dictionary matrix $\mathbf{W}$ is constructed of P horizontally stacked blocks, or dictionary atoms $\mathbf{W}_p$, where each is a matrix with $L_{\max}$ columns and N rows. In order to obtain an acceptable approximation of (8), the problem must be solved iteratively, where the last solution is used to improve the next. To pursue an even sparser solution, a re-weighting procedure is simultaneously used for g_2, similar to that in [13]. The solution is thus found at the k-th iteration by solving

$$\hat{\mathbf{a}}^{(k)} = \arg\min_{\mathbf{a}} \sum_{j=1,2,4} g_j(\mathbf{H}_j^{(k)}\mathbf{a}, \lambda_j) \tag{21}$$

where $\mathbf{H}_1^{(k)} = \mathbf{W}$, $\mathbf{H}_2^{(k)} = \operatorname{diag}\big(1/(\|*\| \, \hat{\mathbf{a}}_1^{(k-1)} + \varepsilon)\big)$,
$\mathbf{H}_4^{(k)} = \mathbf{F} \operatorname{diag}(\arg\big(\hat{\mathbf{a}}^{(k-1)}\big))^{-1}$, and with

$$g_1(\mathbf{H}_1^{(k)}\mathbf{a}, 1) = \frac{1}{2} ||\mathbf{y} - \mathbf{W}\mathbf{a}||_2^2 \tag{22}$$

Algorithm 1 The proposed PEBSI-Lite algorithm

initialize $k := 0$, $\mathbf{H}_4^{(0)} = \mathbf{F}$, and
  $\mathbf{a}^{(0)} = \mathbf{z}_{\text{save}} = \mathbf{d}_{\text{save}} = \mathbf{0}^{PL_{\max} \times 1}$
repeat {adaptive penalty scheme}
  initialize $\ell := 0$, $\mathbf{u}^{(2)}(0) = \mathbf{a}^{(k)}$,
    $\mathbf{z}(0) = \mathbf{z}_{\text{save}}$, and $\mathbf{d}(0) = \mathbf{d}_{\text{save}}$
  repeat {ADMM scheme}
    $\mathbf{z}(\ell) = \left(\mathbf{G}^{(k)H}\mathbf{G}^{(k)}\right)^{-1}\mathbf{G}^{(k)H}\left(\mathbf{u}(\ell) + \mathbf{d}(\ell)\right)$
    $\mathbf{u}^{(1)}(\ell+1)$
      $= \frac{\mathbf{y} - \mu\left(\mathbf{H}_1\mathbf{z}(\ell+1) - \mathbf{d}^{(1)}(\ell)\right)}{1+\mu}$
    $\mathbf{u}^{(2)}(\ell+1)$
      $= \mathbf{T}\left(\mathbf{H}_2\mathbf{z}(\ell+1) - \mathbf{d}^{(2)}(\ell), \frac{\lambda_2}{\mu}\right)$
    $\mathbf{u}^{(3)}(\ell+1)$
      $= \mathbf{T}\left(\mathbf{H}_4^{(k)}\mathbf{z}(\ell+1) - \mathbf{d}^{(3)}(\ell), \frac{\lambda_4}{\mu}\right)$
    $\mathbf{d}(\ell+1)$
      $= \mathbf{d}(\ell) - \left(\mathbf{G}^{(k)}\mathbf{z}(\ell+1) - \mathbf{u}(\ell+1)\right)$
    $\ell \leftarrow \ell + 1$
  until convergence
  store $\mathbf{a}^{(k)} = \mathbf{u}^{(2)}(\text{end})$, $\mathbf{z}_{\text{save}} = \mathbf{z}(\text{end})$, and
    $\mathbf{d}_{\text{save}} = \mathbf{d}(\text{end})$
  update $\mathbf{H}_4^{(k+1)} = \mathbf{F}\text{diag}\left(\arg\left(\mathbf{a}^{(k)}\right)\right)^{-1}$
  $k \leftarrow k + 1$
until convergence

$$g_2(\mathbf{H}_2^{(k)}\mathbf{a}, \lambda_2) = \lambda_2 \left|\left|\mathbf{H}_2^{(k)}\mathbf{a}\right|\right|_1 \tag{23}$$

$$g_4(\mathbf{H}_4^{(k)}\mathbf{a}, \lambda_4) = \lambda_4 \left|\left|\mathbf{H}_4^{(k)}\mathbf{a}\right|\right|_1 \tag{24}$$

where diag$(\cdot)$ denotes a diagonal matrix, arg$(\cdot)$ is the element-wise complex argument, and $\varepsilon \ll 1$. Also, $\mathbf{I}$ denotes the identity matrix, and $\mathbf{F}$ is a first order difference matrix, having elements $\mathbf{F}\{n,n\} = 1$, $\mathbf{F}\{n,n+1\} = -1$, for $n = 1,\ldots,PL_{\max} - 1$, and zeros everywhere else. As intended, the minimization in (21) is convex, and may be solved using one of many convex solvers publicly available, such as, for instance, the interior point methods SeDuMi [14]

or SDPT3 [9]. These are, however, quite computationally burdensome and will scale poorly with increased data length and larger grid. Instead, we here propose an efficient implementation using ADMM. In brief, ADMM is a method where the original problem is split into two or more subproblems, using a number of auxiliary variables, which are solved independently in an iterative fashion. The problem in (21) may be implemented in a similar manner as was done [6], thus requiring only two tuning parameters, λ_2 and λ_4. The proposed method compares to PEBS and PEBS-TV introduced in [5] as improving upon the former, and requiring less tuning than the latter. We therefore term the proposed method PEBSI-Lite. An outline of its implementation is given in Algorithm 1 where $\mathbf{z}, \mathbf{u}, \mathbf{d}$ are the introduced auxiliary variables, μ is an inner convergence variable, and

$$\mathbf{G}^{(k)} = \left[\begin{array}{ccc} \mathbf{H}_1^T, & \mathbf{H}_2^{(k)T}, & \mathbf{H}_4^{(k)T} \end{array} \right]^T \tag{25}$$

$$\mathbf{u} = \left[\begin{array}{ccc} \mathbf{u}^{(1)T}, & \mathbf{u}^{(2)T}, & \mathbf{u}^{(3)T} \end{array} \right]^T \tag{26}$$

$$\mathbf{d} = \left[\begin{array}{ccc} \mathbf{d}^{(1)T}, & \mathbf{d}^{(2)T}, & \mathbf{d}^{(3)T} \end{array} \right]^T \tag{27}$$

$$\mathbf{T}\left(\mathbf{x}, \xi\right) = \frac{\max(|\mathbf{x}| - \xi, 0)}{\max(|\mathbf{x}| - \xi, 0) + \xi} \odot \mathbf{x} \tag{28}$$

such that the solution is given as $\hat{\mathbf{a}} = \mathbf{z}(\ell_{\text{end}})$ at iteration k_{end}.

4 Numerical results

In order to examine the performance of the proposed estimator, we evaluate it using a simulated dual-pitch signal, measured in white Gaussian noise at different Signal-to-Noise Ratios (SNR), ranging from -5 dB to 20 dB in steps of 5 dB. At each level of SNR, 200 Monte Carlo simulations are performed, each simulation generating a signal with fundamental frequencies $[600, 730]$ Hz. To reflect the performance in presence of off-grid effects, the fundamental frequencies are randomly chosen at each simulation uniformly $\pm d/2$ from the chosen frequencies, where d is the grid point spacing. The phases of the harmonics in each pitch are chosen uniformly on $[0, 2\pi)$, whereas all have unit magnitude. The signal is sampled at $f_s = 48$ kHz on a time frame of 10 ms, yielding $N = 480$ samples per frame. As a result, the pitches are spaced by just over f_s/N, which is the resolution limit of the periodogram. This is also seen in Figure 1, illustrating the resolution

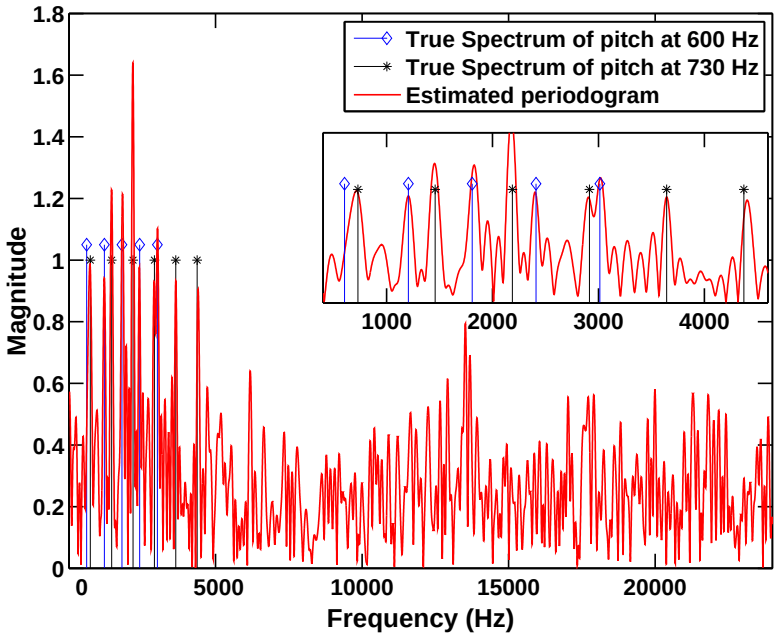


Figure 1: The periodogram estimate and the true signal amplitudes for one of the realizations studied in Figure 2, for SNR $= -5$ dB

of the periodogram as well as the frequencies of the harmonics, at SNR $= -5$ dB. From the figure, it may be concluded that the signal contains more than one harmonic source, as the observed peaks are not harmonically related. Furthermore, it is clear that the fundamental frequencies are not separated by the periodogram, indicating that any pitch estimation algorithm based on the periodogram would suffer notable difficulties. In order to form our estimates, we begin by using a coarse dictionary with candidate pitches uniformly distributed on the interval $[280, 1500]$ Hz, thus also including $\omega_p/2$ and $2\omega_p$ for both pitches. The coarse resolution is $d = 10$ Hz, i.e., still a super-resolution of $1/10N$. After estimation on this grid, a zooming step is taken where a new grid with spacing $d/10$ is laid $\pm 2d$ around each pitch having non-zero power. This zooming approach is taken for the proposed method, as well as for PEBS and PEBS-TV. Comparisons are also made with the ANLS, ORTH, and the harmonic Capon estimators, which have

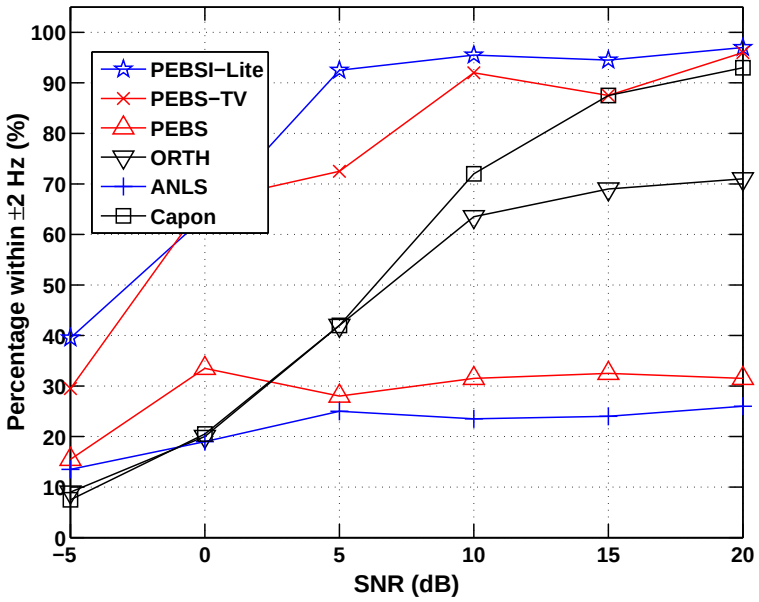


Figure 2: Percentage of estimated pitches where both fundamental frequencies lie at most 2 Hz, or $d/5 = 1/50N$, from the ground truth, plotted as a function of SNR. Here, the pitches have $[5, 6]$ harmonics, respectively, and $L_{\max} = 10$.

been given the oracle model orders (see [15] for more details on these methods). The simulation and estimation procedure is performed for two cases; one where the number of harmonics L_k are set to $[5, 6]$ and one where L_k are set to $[10, 11]$. In the former case, we set $L_{\max} = 10$ and in the latter we set $L_{\max} = 20$, i.e. well above the true number of harmonics. Figures 2 and 3 show the percentage of pitch estimates where both lie within ± 2 Hz from the true values for the six compared methods, for the case of $[5, 6]$ and $[10, 11]$ harmonics, respectively. As is clear from the figures, the proposed method performs as well, or better, than the PEBS-TV algorithm, although requiring fewer tuning parameters. In this setting, PEBS performs poorly, as the generous choices of $L_{\max}$ allows it to ambiguously pick the sub-octave, as predicted.

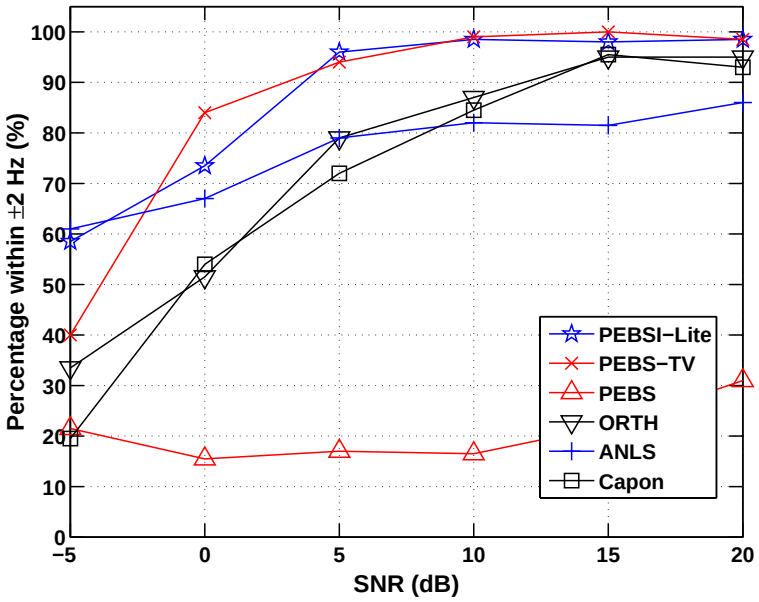


Figure 3: Percentage of estimated pitches where both fundamental frequencies lie at most 2 Hz, or $d/5 = 1/50N$, from the ground truth, plotted as a function of SNR. Here, the pitches have $[10, 11]$ harmonics, respectively, and $L_{\max} = 20$.

References

[1] M. Müller, D. P. W. Ellis, A. Klapuri, and G. Richard, "Signal Processing for Music Analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, 2011.

[2] M. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, Morgan & Claypool, 2009.

[3] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE Transactions on Signal Processing*, vol. 51, no. 1, pp. 101–111, jan. 2003.

[4] J. J. Fuchs, "On the Use of Sparse Representations in the Identification of Line Spectra," in *17th World Congress IFAC*, Seoul, jul 2008, pp. 10225–10229.

[5] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Multi-Pitch Estimation Exploiting Block Sparsity," *Elsevier Signal Processing*, vol. 109, pp. 236–247, April 2015.

[6] M. A. T. Figueiredo and J. M. Bioucas-Dias, "Algorithms for imaging inverse problems under sparsity regularization," in *Proc. 3rd Int. Workshop on Cognitive Information Processing*, May 2012, pp. 1–6.

[7] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 11, no. 6, pp. 804–816, 2003.

[8] T. Kronvall, M. Juhlin, S. I. Adalbjörnsson, and A. Jakobsson, "Sparse Chroma Estimation for Harmonic Audio," in *40th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Brisbane, Apr. 19-24 2015.

[9] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of

Multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[10] T. Kronvall, S. I. Adalbjörnsson, and A. Jakobsson, "Joint DOA and Multi-Pitch Estimation Using Block Sparsity," in *39th IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Florence, May 4-9 2014.

[11] T. Kronvall, S. I. Adalbjörnsson, and A. Jakobsson, "Joint DOA and Multi-pitch estimation via Block Sparse Dictionary Learning," in *22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, Sept. 1-5 2014.

[12] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society B*, vol. 58, no. 1, pp. 267–288, 1996.

[13] E. J. Candes, M. B. Wakin, and S. Boyd, "Enhancing Sparsity by Re-weighted l_1 Minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, Dec. 2008.

[14] R. H. Tutuncu, K. C. Toh, and M. J. Todd, "Solving semidefinite-quadratic-linear programs using SDPT3," *Mathematical Programming Ser. B*, vol. 95, pp. 189–217, 2003.

[15] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Signal Processing*, vol. 88, no. 4, pp. 972–983, April 2008.

B

Paper B

Sparse Localization of Harmonic Audio Sources

Stefan Ingi Adalbjörnsson, Ted Kronvall, Simon Burgess, Kalle Åström, and Andreas Jakobsson

Centre for Mathematical Sciences, Lund University, Lund, Sweden

Abstract

In this paper, we propose a novel method for estimating the locations of near- and/or far-field harmonic audio sources impinging on an arbitrary, but calibrated, sensor array. Using a joint pitch and location estimation formed in two steps, we first estimate the fundamental frequencies and complex amplitudes under a sinusoidal model assumption, whereafter the location of each source is found by utilizing both the difference in phase and the relative attenuation of the magnitude estimates. As audio recordings often consist of multi-pitch signals exhibiting some degree of reverberation, where both the number of pitches and the source locations are unknown, we propose to use sparse heuristics to avoid the necessity of detailed a priori assumptions on the spectral and spatial model orders. The method's performance is evaluated using both simulated and measured audio data, with the former showing that the proposed method achieves near-optimal performance, whereas the latter confirms the method's feasibility when used with real recordings.

Key words: Multi-pitch estimation, near-field and far-field localization, TDOA, block sparsity, convex optimization, ADMM, non-convex sparsity

1 Introduction

Sound localization has been a topic of interest in a wide range of applications for centuries, and is well known to be a difficult problem, especially in a reverberating room environment (see, e.g., [1–7], and the references therein). Typically, a source is located in relation to an array of sensors by exploiting the time delay between sensors for when they receive its emitted signal. In the literature, this is referred to as either time of arrival (TOA) estimation, if the time of signal emission is known, or otherwise time difference of arrival (TDOA) estimation, where only the relative time delays are used. Common techniques for delay estimation include different variations on cross-correlation or canonical correlation analysis (CCA), which then allows the sources to be located in a second step using tri- och multi-lateration (see, e.g., [8]). Such estimates may also be further improved by matching the relative received signal gains to a model for signal attenuation. If the source is far from the sensor array, i.e., in the far-field, its range may not be determined due to the lack of curvature of the impinging sound pressure wavefront, which is then approximately planar, making the range estimation problem ill-posed. The scope is then restricted to determining the direction of arrival (DOA) of the source relative to the sensor array for the 2-D case, or determining azimuth and elevation angles for a 3-D scenario. Historically, such methods are not restricted to sound, but are commonly used, in e.g., military applications, with electromagnetic signals (see, e.g., [9–11]). Perhaps, partly due to differences in application for near-field and far-field techniques, these problems are often treated separately. In this work, and for our purposes with audio signals, the two problems may indifferently be treated together. A common issue with correlation-based techniques is that of reverberation. Although often described in a temporal sense as a filter for each sensor through which the signal is convoluted [12], it may also be analyzed using a spatial formulation. In principle, reverberation occurs when the original source signal is received together with a number of reflections of it, which are both time delayed and dislocated in space with respect to the original. Localization in reverberant environments is still very much an open topic, although several correlation-based approaches exist which shows some degree of robustness (see, e.g., [2]). By assuming a temporal and spectral parametric structure on the received signals, localization may be improved by jointly forming estimates of location together with the parameters of such structures. This is quite common for audio signals such as voiced speech [12], and many forms of harmonic audio sources, such as stringed, wind, and pitched percussion in-

struments [13], which typically have lots of structure. At a glance, the spectral distribution of energy for such signals is typically broadband, but further analysis shows that it is in fact dominantly multi-narrowband, and may be well described using the harmonic model, i.e., as a sum of harmonically related sinusoids [14]. Under this assumption, a source's difference in delay and attenuation when received at the different sensors translates into phase shifted and magnitude scaled versions of the original signal. Exploiting this, joint estimation of the DOA and the pitch frequency has been addressed, such as in [15–17], wherein the authors consider the estimation of the DOA of a single harmonic sound source using a uniform linear array (ULA) of receiver sensor, typically assuming oracle knowledge of the number of harmonic signals in the sound source. Here, we extend on these works, albeit with some generalizations. We are allowing for an unknown number of near- or far-field harmonic sources, each having an unknown number of harmonics, to impinge on an arbitrary, but calibrated, sensor array, in the presence of some degree of reverberation. This feat is attempted through the use of a sparse recovery framework, which avoids making explicit assumptions on the number of harmonic signals, i.e., the number of pitches, as well as for the number of source locations for each pitch. Instead, only an implicit constraint which controls a lower threshold for acceptable source power is needed, which may typically be set using some simple heuristics. Sparse recovery frameworks have in earlier works been found to allow high quality estimates for sinusoidal signals; typical examples include [18–21], wherein the sparse signal reconstruction from noisy observations were accomplished with the by now well-known sparse least squares (LS) technique. More recently, the technique has been extended to the case of harmonically related audio signals [22, 23]. Using the techniques introduced there, we propose a two-step procedure, first creating a dictionary of candidate pitches to model the harmonic components of the sources, without taking the locations of the sources into account, and then, in a second step, a dictionary of possible locations, including simultaneously near- and far-field locations, to model the observed phase differences, as well as the relative attenuations, of the magnitudes of each sinusoidal component. In terms of computational complexity, the estimation problem in each of the two steps is convex, which thus guarantees convergence, and may be solved using a second order cone (SOC) program. As this is typically quite costly, we introduce a computationally efficient implementation based on the alternating direction method of multipliers (ADMM), which makes the proposed method very managable in an off-line estimation procedure. The remainder

of this paper is organized as follows: in the next section, we present the assumed signal model and discuss the imposed restrictions on the sensor array. Then, in section 3, we present the proposed pitch and localization estimator. Section 4 accounts for the ADMM-based implementation, followed in section 5 with an evaluation of the presented technique using both simulated and measured audio signals. Finally, we conclude on our work in section 6.

2 Spatial pitch signal model

In this work, we restrict our attention to the localization of complex-valued[1] harmonically related audio signals, consisting of $\tilde{K}$ distinct sources, $x_k(t)$, for $k = 1, \ldots, \tilde{K}$. Each source is thus assume to consist of L_k harmonically related sinusoids, such that it may be detailed as (see also [14])

$$x_k(t) = \sum_{\ell=1}^{L_k} a_{k,\ell} e^{j\omega_k \ell t} \tag{1}$$

where $\omega_k = 2\pi f_k / f_s$ is the normalized fundamental frequency, with sampling frequency f_s, and with $a_{k,l}$ denoting the complex amplitude of the ℓ:th harmonic.

2.1 Multi-sensor characteristics in near-field environments

When a source signal impinges on a sensor array, it is both delayed and attenuated, such that at sensor m it may be expressed as

$$x_{k,m}(t) \triangleq \frac{d_{k,1}}{d_{k,m}} x_k(t - \tau_{k,m}) \tag{2}$$

where $d_{k,m}$ denotes the sensor-source distance, i.e.,

$$d_{k,m} = \|\mathbf{s}_k - \mathbf{r}_m\|_2 \tag{3}$$

with $\mathbf{s}_k$ and $\mathbf{r}_m$ denoting the location coordinates of the k:th source and the m:th sensor, respectively, and $\|\cdot\|_2$ the Euclidean norm. Thus, (2) accounts for the approximative attenuation of the signal when propogating in space, according to the free-space path loss model. Furthermore, $\tau_{k,m}$ denotes the propagation delay,

[1] Clearly, the measured audio sources will be real-valued, but to simplify notation and in order to reduce complexity, we will here initially compute the discrete-time analytic signal versions of the measured signals, whereafter all processing is done on these signals (see also [14, 24]).

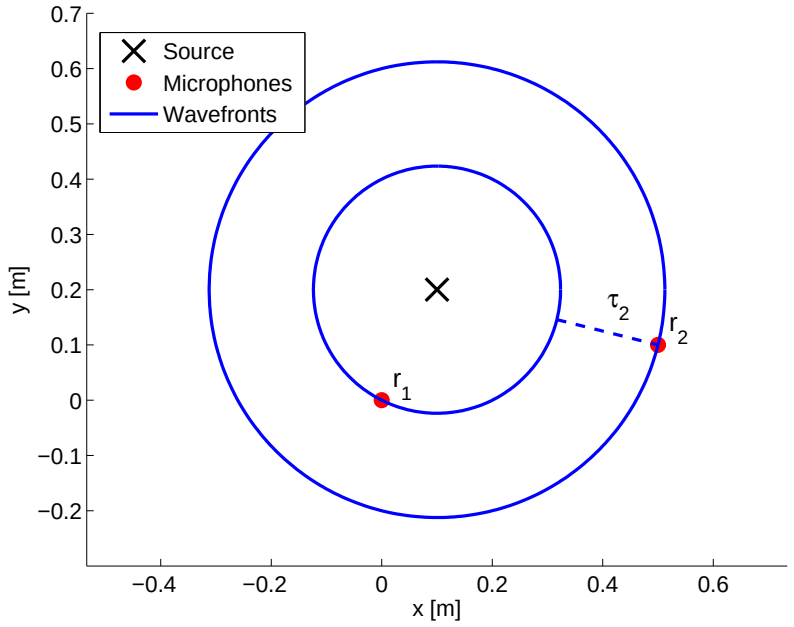


Figure 1: Illustration of a two sensor scenario, with spherical wavefronts propagating from the source. The dashed line shows the scaled TDOA of the second sensor with respect to the first sensor, i.e., τ_2.

i.e., the TDOA, relative to a selected reference sensor, say $m = 1$, so that

$$\tau_{k,m} = c^{-1}\left(d_{k,m} - d_{k,1}\right) \tag{4}$$

for $m = 1, \ldots, M$, where $\tau_{k,1} \triangleq 0$, with c denoting the propagation velocity. An illustration of this is shown in Figure 1, for the case of a single source and two sensors. When recording audio, we often obtain multi-pitch signals of the type

$$x(t) = \sum_{k=1}^{\tilde{K}} x_k(t) \tag{5}$$

which may be either a single source in the physical environment emitting multiple pitch signals, such as an instrument playing a chord, or multiple sources in the physical environment each emitting a single pitch, such as multiple speakers

talking at the same time from different locations. We may also receive a combination of these two types. Without loss of generality, we will hereafter term a source as a spatio-temporal object which has a unique combination of fundamental frequency and location. Two sources may thus have the same fundamental frequency or the same location in space, although not both. This has rather large implications when considering reverberation, where we, apart from the original source, also receive a large number of reflections of it, each reflection having highly similar spectral content, albeit differently attenuated and delayed, i.e., having different magnitudes and phases. All reflections will thus be modeled as separate sources, which implies that under such a model assumption $\tilde{K}$ generally becomes very large. If not seen as separate sources, however, the localization of the original source will become biased by the interference caused from its reflections. To see this, consider for example a sinusoid with frequency ω, magnitude a_1, and phase φ_1, measured in superimposition with its $S-1$ reverberating reflections, having magnitudes $a_2, \ldots, a_S$, and phases $\varphi_2, \ldots, \varphi_S$. For the mth sensor, the measured (noise-free) signal becomes

$$x_m(t) = \sum_{s=1}^{S} a_s e^{-j(\omega t+\varphi_s)} \triangleq b e^{-j(\omega_0 t+\psi)} \tag{6}$$

i.e., a single sinusoid with magnitude $b \in \mathbb{R}_+$ and phase $\psi \in [-\pi, \pi)$, generally being different from the original source. Thus, if trying to estimate the TDOA using phase estimates without taking all reflections into account, for instance by using a correlation-based measure, then only the biased phase, ψ, would be obtained. However, separation of all reflections for all fundamental frequencies is a quite difficult problem, and in this work, we propose to split the estimation procedure into two subproblems. In the first, we find the present fundamental frequencies, and then for each of these we separate the original source(s) from its reflections. To that end, consider $K \leq \tilde{K}$ as the number of unique fundamentals. The noisy signal measured at sensor m may thus be expressed as

$$y_m(t) = \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} b_{k,\ell,m} e^{j\omega_k \ell t} + e_m(t) \tag{7}$$

where the TDOA and attenuation of all S_k reflections of the k:th pitch, for overtone ℓ and sensor m, is gathered in the complex amplitude of the signal, $b_{k,\ell,m}$

using (2) in the same manner as in (6), i.e.,

$$b_{k,\ell,m} = \sum_{s=1}^{S_k} a_{k,\ell,s} \frac{d_{k,1,s}}{d_{k,m,s}} e^{-j\omega_k \ell \tau_{k,m,s}} \tag{8}$$

where $a_{k,\ell,s}$, $d_{k,m,s}$, and $\tau_{k,m,s}$ denote the amplitude, the distance to the mth sensor, and the TDOA for the sth reflection, respectively. Thus, as $\tilde{K} = \sum_{k=1}^{K} S_k$, the estimation procedure first finds the K active fundamentals, whereafter for each one, the original source is separated from its reflections. This approach offers great simplification in contrast to decoupling all $\tilde{K}$ sources simultaneously. To simplify presentation, and without loss of generality, we will here restrict our attention to the case when all sources and signals are restricted to a 2-D plane, i.e., $\mathbf{s} \in \mathbb{R}^2$ and $\mathbf{r} \in \mathbb{R}^2$.

2.2 Avoiding spatial aliasing in arbitrary array geometries

In the literature, keeping below half wavelength sensor spacing is generally preferred to avoid spatial aliasing, although some methods of circumventing this have been published, see e.g. [25]. In this work, we assume a calibrated, although arbitrary, sensor array, without requiring it to satisfy the pairwise half wavelength spacing. We will therefore briefly examine the spatial aliasing effect in the near-field environment, which is the phase difference ambiguity between sensors, resulting when the solution may map to several feasible source locations. To that end, consider a reverberation-free, delayed, and attenuated complex amplitude from a single sinusoidal signal, b. Naturally,

$$b_m = \frac{d_1}{d_m} a e^{-j\omega\tau_m} = \frac{d_1}{d_m} a e^{-j(\omega\tau_m + k2\pi)} \tag{9}$$

and thus the mapping between phase and TDOA is ambiguous for any $k \in \mathbb{Z}$. Considering a given TDOA, and by combining (3) and (4), one will note that any source $\mathbf{s}$ located on a half-space of an hyperbolic curve, i.e.,

$$\tau_m c = \|\mathbf{s} - \mathbf{r}_m\|_2 - \|\mathbf{s} - \mathbf{r}_1\|_2 \tag{10}$$

is a feasible location. To obtain a unique solution, we add additional sensors, and we may thus form new sensor pairs yielding new hyperbolas, where the feasible solution set will be restricted by the intersection of these curves. Ambiguity may

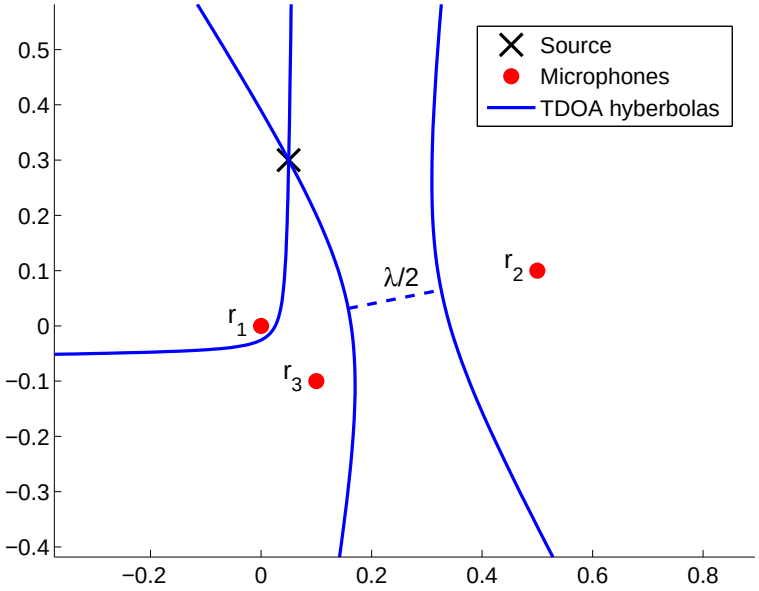


Figure 2: TDOA hyperbolas representing all feasible locations of a single source received by three sensors. As $||\mathbf{r}_2 - \mathbf{r}_1|| > \lambda/2$, spatial aliasing yields another hyperbola of feasible locations. And yet, in this case, there exists only one intersection between the hyperbolas and so the estimate may still be obtained unambiguously.

arise when, for each sensor pair, there exist another TDOA (and thus another k) which fulfills (9), giving rise to an additional hyperbolic curve of feasible points, also intersecting the hyperbolas for other sensor pairs. To identify such ambiguous cases, we first show that a feasible TDOA is restricted to an interval. Using the triangle inequality,

$$|\tau_m c| = \left| \|\mathbf{s} - \mathbf{r}_m\|_2 - \|\mathbf{s} - \mathbf{r}_1\|_2 \right| \leq \|\mathbf{r}_m - \mathbf{r}_1\|_2 \tag{11}$$

it is directly implied that the TDOA must satisfy

$$\tau_m c \in \left[-\|\mathbf{r}_m - \mathbf{r}_1\|_2 , \|\mathbf{r}_m - \mathbf{r}_1\|_2 \right] \tag{12}$$

i.e., is restricted by the sensor-sensor distance. And so, using (9), an estimate of $\arg b \in [-\pi, \pi]$ will map to any TDOA

$$\tau_m c = \frac{\lambda \arg b}{2\pi} + \lambda k \in \left[-\left\| \mathbf{r}_m - \mathbf{r}_1 \right\|_2, \left\| \mathbf{r}_m - \mathbf{r}_1 \right\|_2 \right] \tag{13}$$

where $k \in \mathbb{Z}$, and $\lambda = 2\pi c/\omega$ is the wavelength of the signal. Therefore, if the sensors are spaced by less than $\lambda/2$, the feasible τ_m is unique, and there is no ambiguity in the resulting estimates. If instead some sensors are spaced further apart than $\lambda/2$, then, for all such sensor pairs, there will be more than one feasible TDOA, thereby yielding as many hyperbolas indicating feasible source locations, with a minimum distance of $\lambda/2$ apart. Our main argument to relax the halv wavelength spacing limit is that, when using sufficiently many sensors, the feasible source locations are restricted to the intersection of many hyperbolas, which will, with a high probability, yield a unique solution. Consider an example illustrated in Figure 2, where a single source emits a 1000 Hz signal, which is recorded by three sensors. As shown in the figure, between sensors one and three, which are less than $\lambda/2$ apart, the source gives a single TDOA and a corresponding hyperbola, where the source may be located. Between sensors one and two, which are spaced by more than $\lambda/2$ apart, a second TDOA is feasible, λ/c apart from the true one, also fulfilling (13). However, as shown in the figure, the combined hyperbolas coincide in only a single feasible location, thus still allowing for an unambiguous estimate of the source location. Furthermore, for pitch signals, each overtone will yield a separate set of hyperbolas, which all must intersect to the same location, which further helps to avoid ambiguity. Modeling the attenuation between sensors also helps to avoid ambiguity. Examining the magnitude of the the complex amplitude in (9), we find that

$$|b_m| = \frac{d_1}{d_m}|a| \tag{14}$$

for each pair, consisting of the first and the m:th microphone, which limits $\mathbf{s}$ to lie on a circle. Using the same arguments as above, a feasible source location in terms of attenuation is thus the intersection of circles for all microphone pairs, and will further contribute to avoid spatial aliasing. Even if, despite of intersecting the feasible solutions for all harmonics in terms of both delay and attenuation, ambiguities still remain, then as more sensors are added to the array the set of possible locations quickly becomes small, and a unique solution generally exists,

even if not guaranteed. We thus deem that the imposed restriction on the array's geometry is mild.

3 Joint estimation of pitch and location

We proceed to detail the proposed two-step procedure to form reliable estimates of both the pitches and locations of the sources impinging on the array, without assuming detailed model knowledge of either the number of sources, K, the number of overtones for each source, L_k, the number of reflections experienced due to a possibly reverberant environment, S_k, or requiring knowledge about if sources are far- or near-field. In the first step, the magnitudes, phases, fundamental frequencies, and model orders of the present pitches are estimated, whereas, in the second step, the phase estimates are used to find the locations of these sources. Let

$$\mathbf{\Phi} = \left\{ \left\{ b_{k,\ell,m} \right\}_{\substack{\ell=1,\ldots,L_k \\ m=1,\ldots,M}}, \omega_k, L_k \right\}_{k=1,\ldots,K} \tag{15}$$

denote the set of unknown parameters to be determined in the first step. Minimizing the squared model residual in (7), an estimate of $\mathbf{\Phi}$ may thus be formed as

$$\hat{\mathbf{\Phi}} = \arg\min_{\mathbf{\Phi}} \sum_{t=1}^{N} \sum_{m=1}^{M} \left| y_m(t) - \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} b_{k,\ell,m} e^{j\omega_k \ell t} \right|^2 \tag{16}$$

Clearly, given the dimensionality of the problem, and the required model order estimation steps in order to determine K and L_k, this is a non-trivial problem, and needs to be modified to allow for an efficient solution, as is detailed below. Moving over to the second step, where the found magnitude and phase estimates, $\hat{b}_{k,l,m}$, are exploited to form estimates of the source locations, let

$$\mathbf{\Psi}_k = \left\{ \left\{ a_{k,\ell,s} \right\}_{\ell=1,\ldots,L_k}, \mathbf{s}_s \right\}_{s=1,\ldots,S_k} \tag{17}$$

be the amplitudes and coordinates for a present fundamental frequency k. The locations may be determined by minimizing the squared model residual in (8), i.e.,

$$\hat{\mathbf{\Psi}}_k = \arg\min_{\mathbf{\Psi}_k} \sum_{\ell=1}^{\hat{L}_k} \sum_{m=1}^{M} \left| \hat{b}_{k,\ell,m} - \sum_{s=1}^{S_k} a_{k,\ell,s} d_{k,m,s}^{-1} e^{-j\omega_k \ell \tau_{k,m,s}} \right|^2 \tag{18}$$

where $\tau_{k,m,s}$ and $d_{k,m,s}$ are functions of the location $\mathbf{s}_s$, as defined in (3) and (4). As before, this minimization is also non-trivial, requiring an estimate of S_k, and also needs to be modified to allow for a reasonably efficient solution. In the following, we will elaborate on the proposed modifications of the above minimizations. In order to do so, we first extend the sparse pitch estimation algorithm presented in [22, 23] to allow for multiple measurement vectors. In the second minimization, we then introduce a similar sparsity pattern to solve the localization problem. We begin by examining the extended pitch estimation algorithm.

3.1 Step 1: Sparse pitch estimation

Define the measurement matrix

$$\mathbf{Y} = \left[\begin{array}{ccc} \mathbf{y}(1) & \dots & \mathbf{y}(N) \end{array} \right]^T \tag{19}$$

where

$$\mathbf{y}(t) = \left[\begin{array}{ccc} y_0(t) & \dots & y_{M-1}(t) \end{array} \right]^T \tag{20}$$

denotes a sensor snapshot for each time point $t = 1, \dots, N$, with $(\cdot)^T$ being the transpose. The measurements may then be concisely expressed as

$$\mathbf{Y} = \sum_{k=1}^{K} \mathbf{W}_k \mathbf{B}_k + \mathbf{E} \tag{21}$$

where $\mathbf{E}$ denotes the combined noise term constructed similar to $\mathbf{Y}$, and

$$\mathbf{W}_k = \left[\begin{array}{ccc} \mathbf{w}_k^1 & \dots & \mathbf{w}_k^{L_k} \end{array} \right] \tag{22}$$

$$\mathbf{w}_k = \left[\begin{array}{ccc} e^{j\omega_k} & \dots & e^{j\omega_k N} \end{array} \right]^T \tag{23}$$

$$\mathbf{B}_k = \left[\begin{array}{ccc} \mathbf{b}_{k,1} & \dots & \mathbf{b}_{k,L_k} \end{array} \right]^T \tag{24}$$

$$\mathbf{b}_{k,\ell} = \left[\begin{array}{ccc} b_{k,\ell,1} & \dots & b_{k,\ell,M} \end{array} \right]^T \tag{25}$$

Reminiscent to the sparse estimation framework proposed in [18], we form an extended dictionary of feasible fundamental frequencies, $\omega_1, \dots, \omega_P$, where $P \gg K$, being chosen so large that K of these will reasonably well coincide with the true pitches in the signal. In the same manner, the number of harmonics of each

pitch is extended to an arbitrary upper level, say $L_{\max}$, for all dictionary elements. The signal model may thus be expressed as

$$\mathbf{Y} = \sum_{p=1}^{P} \mathbf{W}_p \mathbf{B}_k + \mathbf{E} = \mathcal{W}\mathcal{B} + \mathbf{E} \tag{26}$$

where the block dictionary matrices are formed by stacking the matrices such that

$$\mathcal{W} = \begin{bmatrix} \mathbf{W}_1 & \dots & \mathbf{W}_P \end{bmatrix} \tag{27}$$

$$\mathcal{B} = \begin{bmatrix} \mathbf{B}_1^T & \dots & \mathbf{B}_P^T \end{bmatrix}^T \tag{28}$$

Note from (11) that if the element (ℓ, r) of the matrix $\mathbf{B}_k$ is non-zero, the frequency $\ell\omega_k$ is present in the signal at sensor r. Furthermore, since we assume all sensors to receive essentially the same signal, although time-delayed, one may assume that for a harmonic signal, the rows off a non-zero $\mathbf{B}_k$ will either be non-zero, implying that the harmonic ℓ is present in the pitch, or zero, if the harmonic is missing. An appropriate criterion, that promotes a combination of model to data fit and the sparsity pattern just described, may thus be formed as

$$\underset{\mathcal{B}}{\text{minimize}} \left\{ \frac{1}{2} \left\| \mathbf{Y} - \mathcal{W}\mathcal{B} \right\|_{\mathcal{F}}^2 + \lambda \sum_{p=1}^{P} \sum_{\ell=1}^{L_p} \left\| \mathbf{b}_{p,\ell} \right\|_2 + \sum_{p=1}^{P} \gamma_p \left\| \mathbf{B}_p \right\|_{\mathcal{F}} \right\} \tag{29}$$

where two different kinds of group sparsities are imposed, and with $\|\cdot\|_{\mathcal{F}}$ denoting the Frobenius norm. This can be seen to be a generalization of the sparse group lasso to the multiple measurement case (see also [23, 26]). Here, the double sum of 2-norms in the second entry of the minimization should enforce sparsity in the solution in the rows of $\mathcal{B}$, and ideally only have as many non-zero rows as there are sinusoids in the signal. The third entry makes the solution (matrix) block sparse over the candidate pitches, penalizing the number of pitches with non-zero magnitude in the signal, ideally making them as many as there are pitches in the signal, i.e., K. Given an optimal point, $\hat{\mathcal{B}}$, the number of pitches is thus estimated as the number of non-zero matrices $\hat{\mathbf{B}}_k$, and, for each pitch, the number of harmonics, L_k, is estimated as the number of non-zero rows. The user parameters

λ, $\gamma_p \in \mathbb{R}_+$ weighs the fit of the solution to its vector and matrix sparsity, respectively. It is well known (see, e.g., [27]) that the amplitudes in the sparse estimate will be increasingly biased towards zero as sparse regularizers are increased. As we here intend to use both the estimated phases and the magnitudes, we propose to refine the amplitude estimates using a reweighting scheme similar to the one presented in [28]. This is accomplished by iteratively solving (29), such that at iteration $j+1$, one updates

$$\gamma_p^{(j+1)} = \frac{\gamma_p^{(0)}}{\left|\left|\hat{\mathbf{B}}_p^{(j)}\right|\right|_{\mathcal{F}} + \varepsilon} \tag{30}$$

where $\hat{\mathbf{B}}_p^{(j)}$ is block p of the optimal point for iteration j, and all $\gamma_p^{(0)}$ are set to be equal in the first iteration. As a result, the block matrices, $\hat{\mathbf{B}}_p^{(j)}$, which have a small Frobenius norm at iteration j will be penalized harder in the next step, whereas the ones that have a larger Frobenius norm will be penalized less, and as a result reducing the bias. The resulting algorithm can be seen as a sequence of iterative convex programs to approximate the concave $\log(\sum_{p=1}^{P} \gamma_p^{(0)} \left|\left|\mathbf{B}_p\right|\right|_{\mathcal{F}} + \varepsilon)$ penalty function [29], where ε is chosen as a small number to avoid numerical difficulties. The introduction of the reweighting yields sparser estimates due to the introduction of the log penalty [28, 30], and the resulting technique may be viewed as an alternative to using an information criterion (as was done in [23], to avoid spurious peaks caused by the signal model and data miss-match).

It is worth noting that as we are here focusing on localization, we have selected to use a somewhat simplistic audio model that ignores several important features in harmonic audio signals, such as issues of inharmonicities, pitch halvings and doublings, and of the commonly occurring forms of amplitude modulation exhibited by most audio sources (see also [14]). Clearly, the used model could be refined reminiscent to models such as the one used in [23,31], introducing a total variation penalty to each column of $\mathcal{B}$, and/or using an uncertainty volume to allow for inharmonicity. However, for localization purposes, these issues are of less concern, as halvings/doublings and/or amplitude modulations will not affect the below localization procedure more than marginally. Inharmonicity is more pressing, but we have in our numerical studies found that given the size of the calibration errors, the inharmonicity is not affecting the solution significantly, and in the interest of reducing the complexity, we have here opted to exclude this aspect from the estimator.

As for the selection of the tuning parameters, one may use, for example, cross validation techniques, although it may be noted that, in high SNR cases, one can often get good results by simply inspecting the periodogram and by then setting the tuning parameters appropriately (see also [23] for a further discussion on this issue). Furthermore, we note that in the case of different noise variances at each sensor in the array, the Frobenius norm in the first entry of the minimization criterion may be replaced with a weighed Frobenius norm. Finally, we note that non-Gaussian noise distributions can also be used as long as the negative log-likelihood is convex.

3.2 Step 2: Sparse localization

According to the signal model (7), $\hat{\mathcal{B}}$ will inherently contain the TDOA and attenuation for all reflections of any fundamental frequency present in the signal, which enables a range of post-processing steps to, for instance, estimate position, track, and/or calibrate the sensors. Here, we limit our attention to estimating the source positions. Let $\hat{\mathcal{B}}$ denote the solution obtained from minimizing (29), and consider a scenario where the sources are well separated in their pitch frequencies, and, initially, suffering from negligible reverberation, implying that $S_1 = \ldots = S_P = 1$. Then, the minimization in (18) may be seen as a generalization of the time-varying amplitude modulation problem examined in [32] (see also [11]) to the case of several realizations of the same signal, sampled at irregular time points, and with a different initial phase for each realization. Reminiscent to the solution presented in [11, p. 186], one may thus find the source locations, for far-field signals, for every pitch p with non-zero amplitudes in $\mathbf{B}_p$, as

$$\hat{\mathbf{s}}_p = \arg\max_{\mathbf{s}_p} \sum_{\ell=1}^{L_p} \left| \sum_{m=1}^{M} \hat{b}^2_{p,\ell,m} e^{-j2\omega_p \ell \tau_{p,\ell,m}} \right|^2 \tag{31}$$

where the TDOAs $\tau_{p,\ell,m}$ are found as a function of the source location $\mathbf{s}_p$, using (4). This minimization may be well approximated by 1-D searches over range and DOA (or over range, azimuth, and elevation in the 3-D case). Considering also reverberating room environments, wherein each of the pitches may appear as originating from many different locations, the minimization needs to be extended to allow for varying number of reflections, S_k. To allow for such reflections, we proceed to model every non-zero amplitude block from the pitch estimation step

as

$$\mathbf{B}_k = \sum_{s=1}^{S_k} \operatorname{diag}\left(\mathbf{a}_{k,s}\right) \mathbf{U}_{k,s} + \boldsymbol{\mathcal{E}}_k \tag{32}$$

with diag($\mathbf{x}$) denoting a diagonal matrix with the vector $\mathbf{x}$ along its diagonal, $\boldsymbol{\mathcal{E}}_k$ the combined noise term constructed in the same manner as $\mathbf{B}_k$, and

$$\mathbf{U}_{k,s} = \begin{bmatrix} \mathbf{u}_{k,s}^1 & \dots & \mathbf{u}_{k,s}^{\hat{L}_k} \end{bmatrix} \tag{33}$$

$$\mathbf{u}_{k,s} = \begin{bmatrix} \frac{e^{j\omega_k \tau_{k,1,s}}}{1} & \dots & \frac{e^{j\omega_k \tau_{k,M,s}}}{d_{k,M,s}/d_{k,m,s}} \end{bmatrix}^T \tag{34}$$

$$\mathbf{a}_{k,s} = \begin{bmatrix} \mathbf{a}_{k,1,s} & \dots & \mathbf{a}_{k,\hat{L}_k,s} \end{bmatrix}^T \tag{35}$$

where $\tau_{k,m,s}$ and $d_{k,m,s}$ are related to the source location as given by (3) and (4), respecively. Analogously to the above procedure for the pitch estimation, we then extend the dictionary of feasible source locations for the kth source, $\mathbf{s}_1, \dots, \mathbf{s}_{S_k}$, onto a grid of $Q \gg S_k$ candidate locations $\mathbf{s}_q$, for $q = 1, \dots, Q$, with Q chosen large enough to allow some of the introduced dictionary elements to coincide, or closely so, with the true source locations in the signal. Clearly, this may force Q to be very large. Striving to keep the size of the dictionary as small as possible, we consider grid points in polar coordinates, with equal resolution for all considered DOAs, and linearly spaced grid points over the distance in each DOA. Thus, we get a denser grid in the close proximity to the sensor array, where the resolution capacity is highest, and then a less and less dense grid for sources further away from the array. Finally, to also allow for far-field sources, one may include one dictionary element for each direction at an infinite range, for which, naturally, the attenuation effect may be disregarded, i.e., $d_{k,m,s} \triangleq 1$ for all sensors. Thus, we may estimate the source locations for the k:th pitch using a sparse modelling framework as

$$\begin{aligned} \underset{\mathbf{a}_{k,1},\dots,\mathbf{a}_{k,Q}}{\text{minimize}} \Bigg\{ & \frac{1}{2} \left\| \mathbf{B}_k - \sum_{q=1}^{Q} \operatorname{diag} \mathbf{a}_{k,q} \mathbf{U}_{k,q} \right\|_{\mathcal{F}}^2 \\ & + \sum_{q=1}^{Q} \varkappa_q \left\| \mathbf{a}_{k,q} \right\|_2 + \rho \sum_{q=1}^{Q} \left\| \mathbf{a}_{k,q} \right\|_1 \Bigg\} \end{aligned} \tag{36}$$

where, again, two types of sparsity is imposed on the solution. The 2-norm penalty term imposes sparsity to the blocks $\mathbf{a}_{k,q}$, i.e., penalizing the number of source locations present in the signal. Furthermore, the 1-norm term penalizes the number of harmonics, to allow for cases when some sources may have missing harmonics. Thus, here the number of sources is estimated as the number of nonzero blocks in an optimal point and any zero elements within a block corresponding to a missing harmonic. Here, $\varkappa_q, \rho \in \mathbb{R}_+$ are tuning parameters, controlling the amount of sparsity and the weight between sparsity in pitches and in harmonics, respectively, whereas the factor ρ is only used if two sources share the same fundamental frequency but differ in which harmonics are present. Finally, $\varkappa_q$ may be updated in the same manner as described in section III.A. As shown in the following section, the optimization problem in (29) and (36) are equivalent, so these tuning parameters may be set in a similar fashion.

4 Efficient implementation

It is worth noting that both the minimization in (29) and (36) are convex, as the tuning parameters are non-negative and all the functions are convex. Their solutions may thus be found using standard convex minimization techniques, e.g., using CVX [33, 34], SeDuMi [35], or SDPT3 [36]. Regrettably, such solvers will scale poorly both with increasing data length, the use of a finer grid for the fundamental frequencies, and with the number of sensors. Furthermore, such implementations are unable to utilize the full structure of the minimization, and may, as a result, be computationally cumbersome in practical situations. To alleviate this, we proceed to formulate a novel ADMM re-formulation of the minimizations, offering efficient and fast implementations of both minimizations. For completeness and to introduce our notation, we briefly review the main steps involved in an ADMM (we refer the reader to [37, 38] for further details on the ADMM). Considering the convex optimization problem

$$\underset{\mathbf{z}}{\text{minimize}}\, f(\mathbf{z}) + g(\mathbf{z}) \tag{37}$$

where $\mathbf{z} \in \mathbb{R}^p$ is the optimization variable, with $f(\cdot)$ and $g(\cdot)$ being convex functions. Introducing the auxiliary variable, $\mathbf{u}$ (37) may be equivalently be expressed as

$$\underset{\mathbf{z},\mathbf{u}}{\text{minimize}}\, f(\mathbf{z}) + g(\mathbf{u}) \quad \text{subject to } \mathbf{z} - \mathbf{u} = \mathbf{0} \tag{38}$$

Algorithm 1 The ADMM algorithm

1: Initiate $\mathbf{z} = \mathbf{z}_0$, $\mathbf{u} = \mathbf{u}_0$, and $k = 0$
2: **repeat**
3: $\quad \mathbf{z}_{k+1} = \underset{\mathbf{z}}{\operatorname{argmin}} f(\mathbf{z}) + \frac{\mu}{2}||\mathbf{z} - \mathbf{u}_k - \mathbf{d}_k||_2^2$
4: $\quad \mathbf{u}_{k+1} = \underset{\mathbf{u}}{\operatorname{argmin}}\, g(\mathbf{u}) + \frac{\mu}{2}||\mathbf{z}_{\mathbf{k+1}} - \mathbf{u} - \mathbf{d}_k||_2^2$
5: $\quad \mathbf{d}_{k+1} = \mathbf{d}_k - (\mathbf{z}_{\mathbf{k+1}} - \mathbf{u}_{\mathbf{k+1}})$
6: $\quad k \leftarrow k + 1$
7: **until** convergence

since at any feasible point $\mathbf{z} = \mathbf{u}$. Under the assumption that there is no duality gap, which is true for the here considered minimizations, one may solve the optimization problem via the dual function defined as the infimum of the augmented Lagrangian, with respect to $\mathbf{x}$ and $\mathbf{z}$, i.e., (see also [37])

$$L_\mu(\mathbf{z}, \mathbf{u}, \mathbf{d}) = f(\mathbf{z}) + g(\mathbf{u}) + \mathbf{d}^T(\mathbf{z} - \mathbf{u}) + \frac{\mu}{2}||\mathbf{z} - \mathbf{u}||_2^2$$

The ADMM does this by iteratively maximizing the dual function such that at step $k + 1$, one minimizes the Lagrangian for one of the variables, while holding the other fixed at its most recent value, i.e.,

$$\mathbf{z}_{k+1} = \arg\min_{\mathbf{z}} L_\mu(\mathbf{z}, \mathbf{u}_k, \mathbf{d}_k) \tag{39}$$

$$\mathbf{u}_{k+1} = \arg\min_{\mathbf{u}} L_\mu\left(\mathbf{z}_{k+1}, \mathbf{u}_k, \mathbf{d}_k\right) \tag{40}$$

Finally, one updates the dual variable by taking a gradient ascent step to maximize the dual function, resulting in

$$\tilde{\mathbf{d}}_{k+1} = \tilde{\mathbf{d}}_k - \mu\left(\mathbf{z}_{k+1} - \tilde{\mathbf{d}}_{k+1}\right) \tag{41}$$

where μ is the dual variable step size. The general ADMM steps are summarized in Algorithm 1, using the scaled version of the dual variable $\mathbf{d}_k = \tilde{\mathbf{d}}/\mu$, which is more convenient for implementation. Thus, in cases when steps 3 and 4 of Algorithm 1 may be carried out more efficiently than for the original problem, the ADMM may be useful to form an efficient implementation of the considered minimization.

It may be noted that the minimizations in (29) and (36) are rather similar, both containing an affine function in a Frobenius norm, as well as a sum of the norm of different subset of the variable. In fact, by using the vec operation, i.e., vectorization, both minimizations may be shown to be equivalent with the problem

$$\underset{\mathbf{z}}{\text{minimize}}\Bigg\{ \frac{1}{2}\left\|\mathbf{y}-\mathbf{A}\mathbf{z}\right\|_2^2 + \gamma\sum_{k=1}^{P}\left\|\mathbf{z}_k\right\|_2 + \delta\sum_{k=1}^{P}\sum_{g=1}^{G_k}\left\|\mathbf{z}_{k,g}\right\|_2 \Bigg\} \tag{42}$$

where the complex variable $\mathbf{z}$ is given as

$$\mathbf{z} = \left[\begin{array}{ccc} \mathbf{z}_1^T & \dots & \mathbf{z}_P^T \end{array}\right]^T \tag{43}$$

$$\mathbf{z}_k = \left[\begin{array}{ccc} \mathbf{z}_{k,1}^T & \dots & \mathbf{z}_{k,G_k}^T \end{array}\right]^T \tag{44}$$

where each $\mathbf{z}_k$ and $\mathbf{z}_{k,g}$ denote complex vectors with G_k and O elements, respectively. For the minimization in (29), this implies that

$$\mathbf{y} = \text{vec}(\mathbf{Y}) \tag{45}$$

$$\mathbf{z} = \text{vec}(\mathcal{B}) \tag{46}$$

$$\mathbf{A} = \mathbf{I} \otimes \mathcal{W} \tag{47}$$

where $\otimes$ and $\mathbf{I}$ denote the Kronecker product and an M-dimensional identity matrix, respectively, with G_k being equal to the number of harmonics, L_k, and O equals the number of sensors, M. Similarly, for the minimization in (36),

$$\mathbf{y} = \text{vec}(\mathbf{B}_p) \tag{48}$$

$$\mathbf{z} = \mathbf{a}_k \tag{49}$$

$$\mathbf{A} = \tilde{\mathbf{V}}_k \tag{50}$$

where

$$\mathbf{a}_k = \left[\begin{array}{ccc} \mathbf{a}_{k,1}^T & \dots & \mathbf{a}_{k,Q}^T \end{array}\right]^T \tag{51}$$

$$\tilde{\mathbf{V}}_k = \left[\begin{array}{ccc} \tilde{\mathbf{V}}_{k,1} & \dots & \tilde{\mathbf{V}}_{k,Q} \end{array}\right] \tag{52}$$

and $\mathbf{V}_{k,q} = \mathbf{U}_{k,q} \otimes \mathbf{I}$, with $\tilde{\mathbf{V}}_{k,q}$ being formed by removing all columns from $\mathbf{V}_{k,q}$ that correspond to zeros in the vector vec(diag($\mathbf{a}_{k,q}$)), and G_k being equal to L_k and O equals 1. Thus, we can formulate an ADMM solution for (42) that solves both problem (29) and (36). To that end, defining

$$f(\mathbf{z}) = \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2^2 \tag{53}$$

$$g(\mathbf{u}) = \gamma \sum_{k=1}^{P} \|\mathbf{u}_k\|_2 + \delta \sum_{k=1}^{P} \sum_{g=1}^{Q_k} \left\|\mathbf{u}_{k,g}\right\|_2 \tag{54}$$

yields a quadratic problem in step 3 in Algorithm 1, with a closed form solution given by

$$\mathbf{z}_{k+1} = \left(\mu I + \mathbf{A}^H \mathbf{A}\right)^{-1} \left(\mu \left(\mathbf{u}_k - \mathbf{d}_k\right) + \mathbf{A}^H \mathbf{y}\right)$$

with $(\cdot)^H$ denoting the Hermitian transpose, whereas in step 4, by solving the sub-differential equations (see [23] for further details), one obtains

$$\mathbf{u}_{k+1} = \mathcal{S}^o \left(\mathcal{S}^i \left(\mathbf{z}_k - \mathbf{d}_k, \varkappa/\mu\right), \delta/\mu\right) \tag{55}$$

where the shrinkage operators $\mathcal{S}^o$ and $\mathcal{S}^i$ are defined using the vector shrinkage operator $\mathcal{S}$, defined for any vector $\mathbf{v}$ and positive scalar ξ such that

$$\mathcal{S}(\mathbf{v}, \xi) = \mathbf{v} \left(1 - \xi/||\mathbf{v}||_2\right)^+ \tag{56}$$

where $(\cdot)^+$ is the positive part of the scalar, and

$$\mathcal{S}(\mathbf{z}, \xi)^o = \left[\begin{array}{ccc} \mathcal{S}^T(\mathbf{z}_1, \xi) & \ldots & \mathcal{S}^T(\mathbf{z}_P, \xi) \end{array} \right]^T \tag{57}$$

$$\mathcal{S}(\mathbf{z}, \xi)^i = \left[\begin{array}{ccccc} \mathcal{S}^T(\mathbf{z}_{1,1}, \xi) & \ldots & \mathcal{S}^T(\mathbf{z}_{1,G_1}, \xi) & \ldots & \\ \mathcal{S}^T(\mathbf{z}_{P,1}, \xi) & \ldots & \mathcal{S}^T(\mathbf{z}_{P,G_P}, \xi) & & \end{array} \right]^T \tag{58}$$

The resulting algorithm is here termed the Harmonic Audio LOcalization using block sparsity (HALO) estimator.

5 Numerical comparisons

We proceed to examine the performance of the proposed estimator using both synthetic and measured audio signals, initially examining the performance using

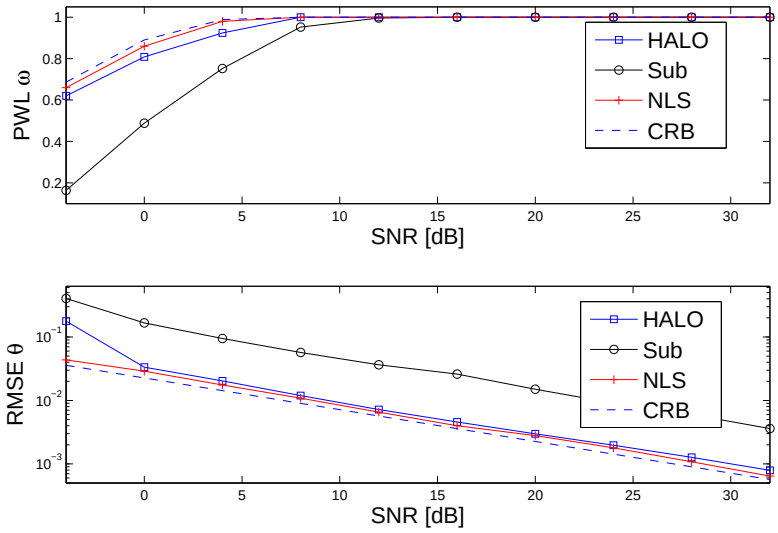


Figure 3: The PWL and RMSE for a single-pitch signal as compared with the optimal performance of an estimator reaching the CRB.

simulated audio signals. In the first examples, we limit ourselves to the case of letting a far-field signal impinge on a ULA. Figure 3 shows the percentage within limits (PWL), defined as the ratio of pitch estimates within a limit of ± 0.1 Hz from the true pitch, and the root mean square error (RMSE) of the DOA, defined as

$$\mathrm{RMSE}_{\vartheta} = \sqrt{\frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\left(\hat{\vartheta}_{k,i} - \vartheta_k\right)^2} \tag{59}$$

where n denotes the number of Monte Carlo (MC) simulation estimates, and K the number of pitches in the signal, for the resulting estimates. For comparison, we use the Cramér-Rao lower bound (CRB), the NLS estimator, and the Sub approach (see [15] for further details on these methods and for the corresponding CRB). These results have been obtained using $n = 250$ MC simulations of a single pitch signal, with $\omega_1 = 220$ Hz and $L_1 = 7$ harmonics, impinging from $\vartheta_1 = -30^\circ$, where both the NLS and the Sub estimators have been al-

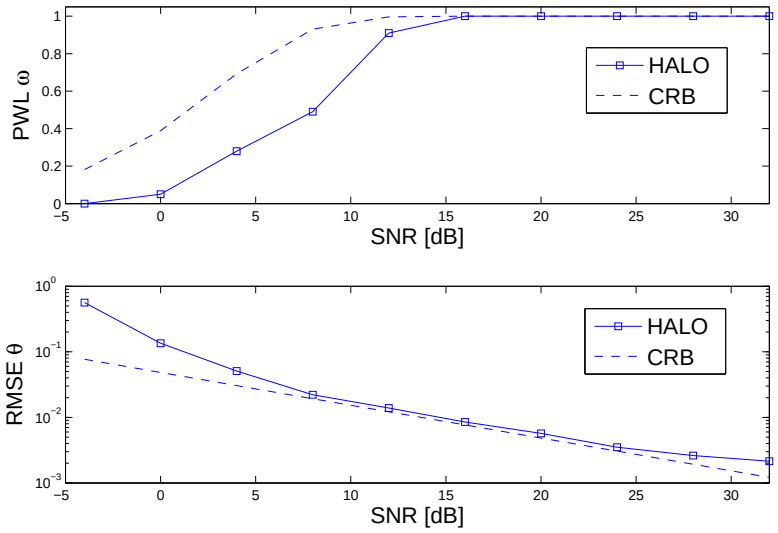


Figure 4: The PWL and RMSE for a multi-pitch signal with two pitches, as compared to the corresponding CRB.

lowed perfect a priori knowledge of both the number of sources and their number of harmonics, whereas the proposed method is allowed no such knowledge. As is clear from the figures, the HALO method offers a preferable performance as compared to the Sub estimator, and only marginally worse than the NLS estimator, in spite of both the latter being allowed perfect model orders information. Here, the number of sensors in the array was $M = 5$ and we used 20 ms of data sampled at $f_s = 8820$ Hz, i.e., $N = 176$ samples. Furthermore, $c = 343$ m/s and $d = c/f_s \approx 0.0389$ m. We proceed to consider the case of multi-pitch signals impinging on the array. Measuring as in the single-pitch case, we now form a multi-pitch signal with two pitches and fundamental frequencies $\{150,\ 220\}$ Hz containing $\{6,\ 7\}$ harmonics, coming from $\vartheta_1 = -30°$. Figure 3 shows the RMSE and PWL estimates, as obtained using 250 MC simulations, clearly showing that the HALO estimator is able to reach close to optimal performance also in this case. Here, no comparison is made with the NLS and Sub estimators of [15] as these are restricted to the single-pitch case. Throughout these evaluations, we have used $L_{\max} = 15$. Also, as the resulting estimates were found to be appropri-

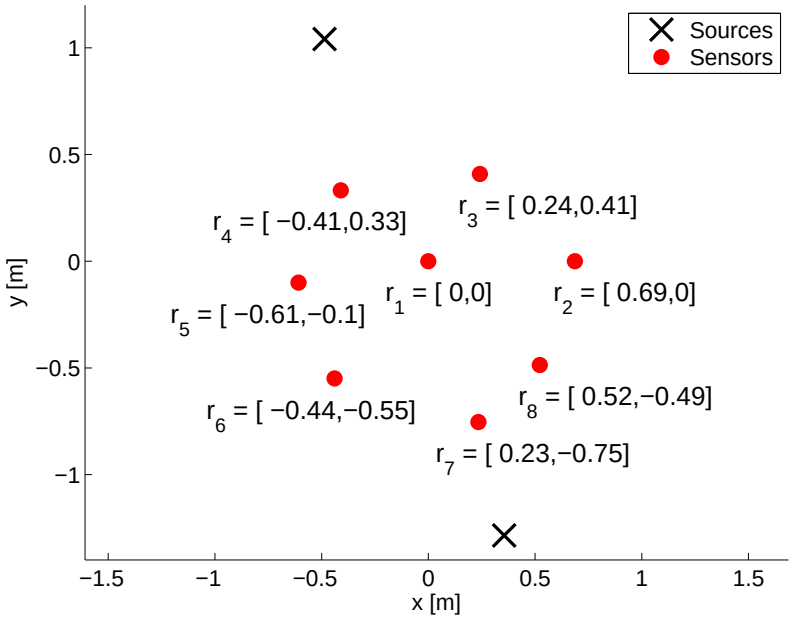


Figure 5: The two-source and eight-sensor layout in 2-D. The position of each sensor, shown in the plot with carthesian coordinates as $r_m = [x, y]$, was obtained in an *a priori* calibration step.

ately sparse when using only the convex penalties, and no reweighing steps were used. We next proceed to examine real measured signals. The measurements were made in an anechoic chamber, approximately $4 \times 4 \times 3$ meters in size, with the sensors and speakers located as shown in Figures 5 and 7. Two speakers were placed at locations (in polar coordinates) $\mathbf{s}_1 = [\vartheta_1, R_1] = [115.03^\circ, 1.15\,\mathrm{m}]$ and $\mathbf{s}_2 = [\vartheta_2, R_2] = [-74.53^\circ, 1.33\,\mathrm{m}]$, with respect to the central microphone, respectively. The positions of the sensors were determined by placing them together with the sources, using the acoustic method detailed in [39]. This is done by calibrating the sensors with a single moving source, using a correlation-based methodology. The positions were also confirmed via a computer vision approach were the positions were found by taking several photos and reconstructing the environment. The maximum deviation in position between these methods was

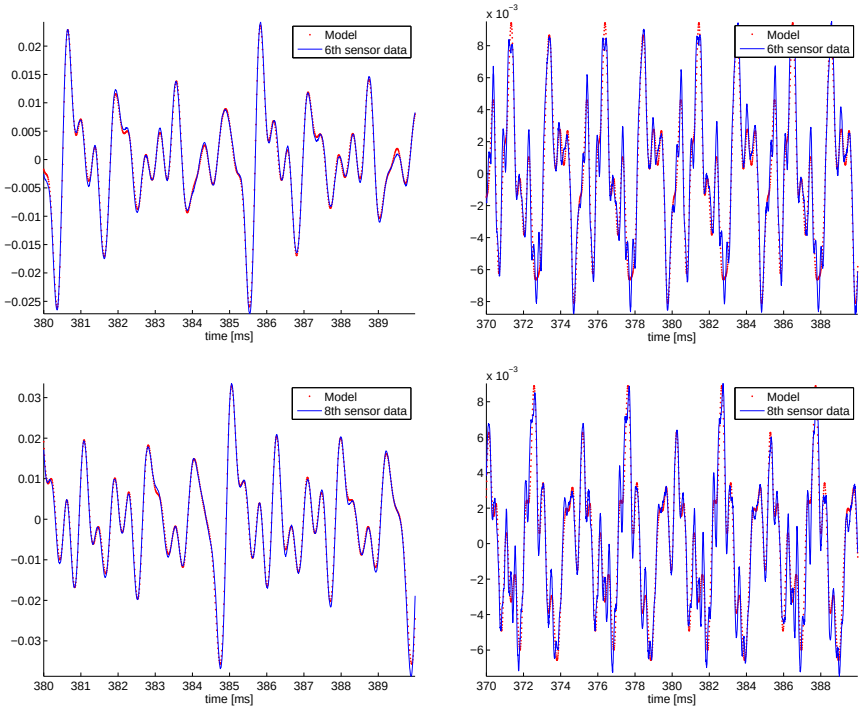


Figure 6: Time-domain data (lined) and estimated signal reconstruction (dotted) for the 6:th sensor (top two) and 8:th sensor (bottom two), for two different signals. The left two subfigures display a voice signal saying the phonetic 'a' in 'why', while the right two subfigures display a violin signal.

less than 1 cm. As the spatial impulse responses of the microphones were deemed to be reasonably omni-directional, as well as roughly the same for all the microphones, no further calibration of the sensor gains were performed. The positions were then projected onto a 2-D plane using principal component analysis. In order to illustrate the HALO estimator's ability to handle an environment with the same pitch signal originating from different sources, as a much simplified proof of concept for a reverberating room environment, we examine a case with two sources playing the same signal content. Both sources plays a (TIMIT) recording of a female voice saying 'Why were you away a year, Roy?', timing the source's playback so that the recording at each microphone sounds slightly echoic. The

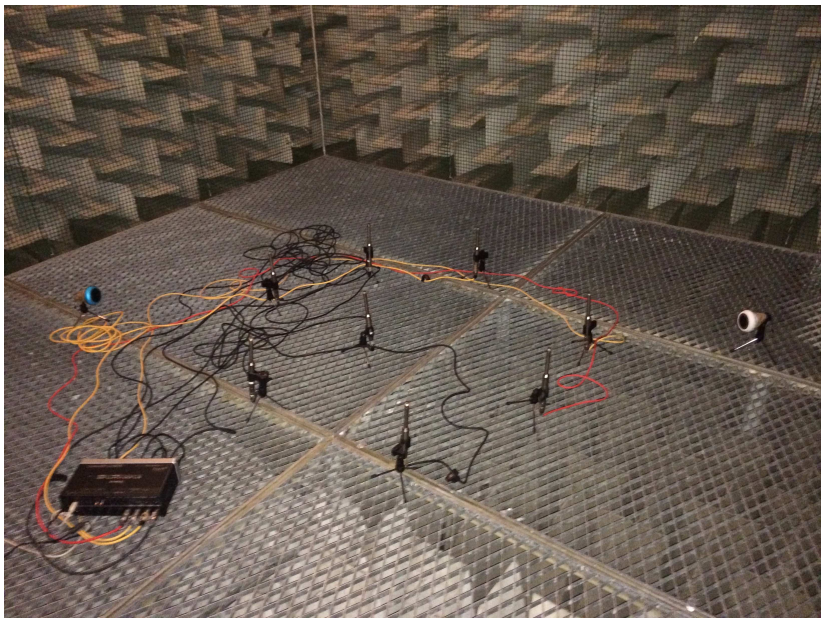


Figure 7: A photo showing the experimental setup in the anechoic chamber, where eight sensors are used to record two coherent sources.

eight microphones all record at a sample rate of $f_s = 96$ kHz. The data is then divided into time frames of 10 ms, i.e., $N = 960$ samples, which allow each frame to be well modelled as being stationary. Examining a part of the speech that is voiced, arbitrarily selected as the frame starting 380 ms into the recording, about when the voice is saying the voiced phonetic sound 'a' in 'why', Figure 4 show the signal measured at the 6th and 8th microphone, respectively, together with the reconstructed signal obtained from the pitch estimation step in HALO, obtained as

$$\hat{\mathbf{Y}} = \mathcal{W}\hat{\mathcal{B}} \tag{60}$$

using the resulting model orders and estimates. The estimator indicate that the signal contains a single pitch at $\hat{\omega}/2\pi = 193.5$ Hz, having $\hat{L} = 12$ overtones. As is clear from the figures, the estimator is well able to model the measured signal in spite of the presence of the reverberation. Comparing the figures, one may also note the time shift between the sensors, due to the additional time-delay for the

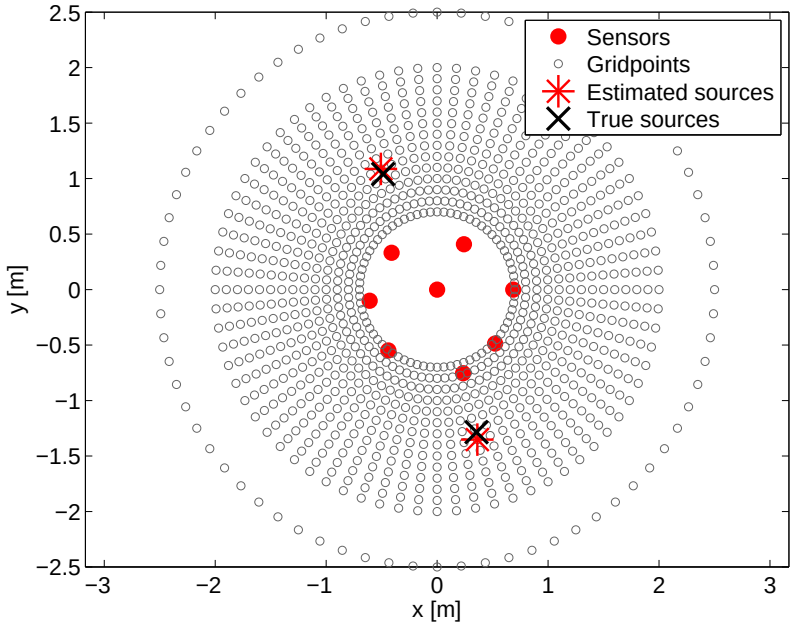


Figure 8: The experimental setup in the anechoic chamber, showing the sensor and loudspeaker locations, the considered dictionary grid, as well as the resulting estimated as obtained by the proposed algorithm.

wavefront traveling between them, corresponding to a linear combination of the two sources, each with their particular TDOA and attenuation. It should also be noted that the signals are not simply time-shifted versions of each other due to the room environment and the attenuation of the signal when propagating in space (which would thus create problems for an estimator based on the cross-correlation between the sensors). The same situation is illustrated in left two subfigures in Figure 4, showing the results when the signal source is replaced with that of a part of a (SQAM) violin signal. Again, the estimator can be seen to be able to well model the impinging signals, which is estimated as being a single pitch with the fundamental frequency $\hat{\omega}/2\pi = 198.0$ Hz, containing $\hat{L} = 14$ harmonics. In order to examine the location estimation, we construct a 2-D grid of feasible locations, chosen such that the space is discretized into 1008 points, consisting

of 72 directions between $[-180^\circ, 180^\circ)$, spaced every 5°, where each direction allows for ranges $R \in [0.7, 2]$ m, spaced 10 cm apart. The resulting grid is shown in Figure 8, which is roughly covering the entirety of the anechoic chamber. To also allow for far-field sources, a range of $R = \infty$ is also added to the grid for each direction, which we have chosen to illustrate by the outer circle in Figure 8. For these far-field grid points, the time-delays are instead computed as (see also [9])

$$\tau_m = \frac{\min_{\mathbf{z}} \|\mathbf{r}_m - \boldsymbol{\ell}(\mathbf{z})\|_2}{c} \tag{61}$$

for a location $\mathbf{z}$ on the line $\boldsymbol{\ell}(\cdot)$, which is perpendicular to the DOA and goes through $\mathbf{r}_1$. The figure also shows the locations for the sensors and the sound sources, as well as the estimated locations, as obtained by the second step of the HALO estimator (the estimated locations were identical for both audio recordings). The errors in position were 5 cm in range for each source, where a bias, overestimating the range, accounts for almost all of the error. On the other hand, as shown in the figure, the angles of the sources ϑ were accurately estimated. The overestimation of the range may to a large extent likely be explained by poor scaling when calibrating the array. One may note that, for localization in 3-D, the size of the dictionary will increase significantly as compared to the 2-D case used for numerical illustration in this paper. For the case above, if also the elevation angle is to be considered, having the same resolution as for the azimuth, this would yield a dictionary of 72 576 atoms. Although much larger, a sparse modeling systems of this size is by no mean impractical to work with. Also, our investigations show that a less dense location grid may be used, whereafter a zooming step can be taken. Finally, we illustrate the algorithm's performance using MC simulations, using simulated sources, one near- and one far-field source, detailed with $\omega = [200, 270]$ Hz, $L = [15, 14]$ harmonics, impinging from $\vartheta = [110^\circ, -70^\circ]$ at $R = [1.3, \infty]$ m, respectively. The sensors are placed as a uniform circular array, with 7 sensor placed evenly at a 0.5 m radius, together with a sensor being placed in the center of the array. First, we examine the position estimates using a coarse spacing for the possible sources, spaced by 11 cm in angle for all angles $\vartheta \in [-180^\circ, 180^\circ)$, and spaced by 10 cm in range, at $R \in [0.7, 3]$ m. In each MC simulation, the true location of each source was offset by a (uniformly distributed) range offset of plus minus one half the grid spacing. In all simulations, we ensured that neither of the sources were placed on a dictionary grid point. Figure 9 shows the PWL for the angle and range estimates, where the limit is chosen to

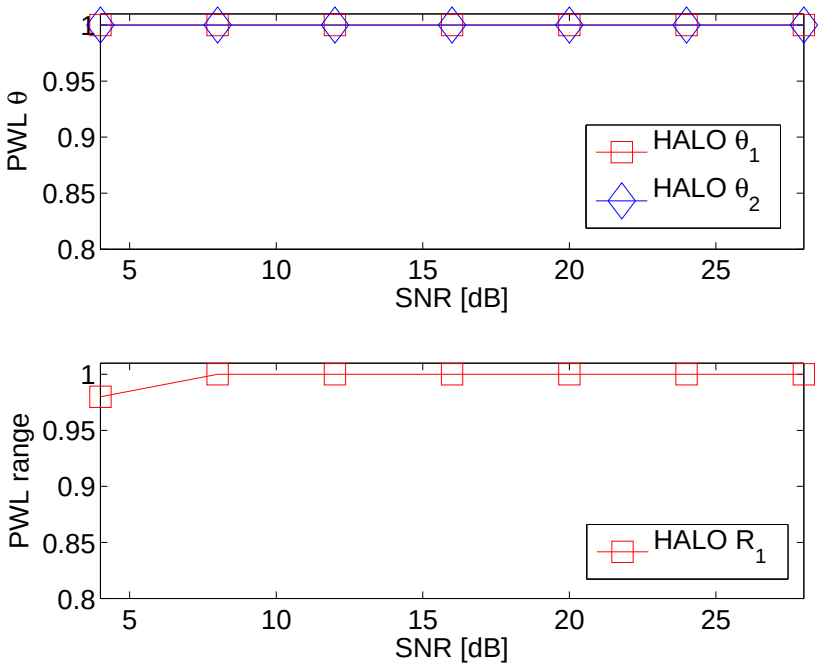


Figure 9: The PWL ratio for the angle and range estimates when using a coarsely spaced grid, indicating the ratio of estimates that are within ± 10 cm in range, and $\pm 5^\circ$ in angle.

be the same as the grid spacing, i.e., the ratio of estimates that are within ± 10 cm in range, and $\pm 5^\circ$ in angle. As seen from the figure, the both the range and the DOA of the sources are well determined, indicating that even with the use of a coarse grid, one is able to obtain reliable estimates. Proceeding to instead using a fine grid, the coarse estimates may then be refined by zooming in the grid over the found locations. Using a dictionary of the same size as the coarse grid, although centered around the found estimates, yields a resolution of ± 5 mm in range and $\pm 0.25^\circ$ in angle. Figure 10 shows the resulting RMSE for the angle and pitch estimates on the finer grid, as compared to the CRB (given in the Appendix). As can be seen from the figure, the RMSE (and the corresponding CRB) of the far-field source is somewhat lower than the near-field source, although both sources

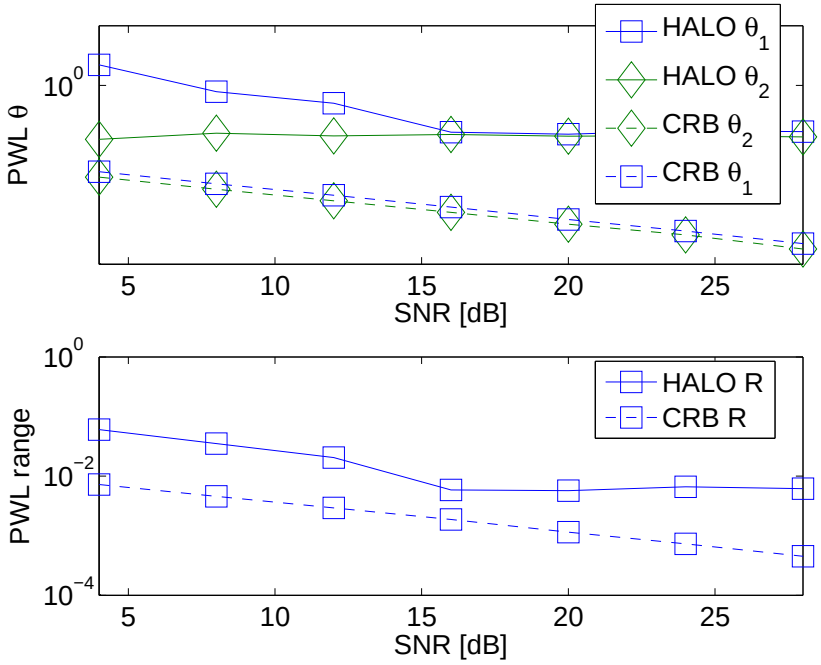


Figure 10: The RMSE for the angle and range estimates when using a finely spaced grid, indicating the ratio of estimates that are within ± 5 mm in range, and $\pm 0.25^\circ$ in angle.

are well estimated, yielding a performance close to being optimal. The slight offset from the CRB is deemed to be largely due to a small bias in the final estimates, resulting from the smoothness of the approximative cost function resulting from the additive convex constraints. As is clear from the above presentation, the HALO estimate exploits the harmonic structure in the received audio signals to position the sources, using the pitch estimates to form a sparse estimate over a wide range of feasible positions. Obviously, most audio signals are not harmonic at all times, and the estimator should thus be used in combination with a tracking technique, possibly using a methodology reminiscent to the one presented in [40, 41]. In such a tracking scheme, the estimated pitch amplitudes should be used as an indicator for the reliability of the obtained positioning, yielding poor or maybe even erroneous positioning for unvoiced or non-harmonic audio sig-

nals, whereas reasonably accurate positions may be expected for more harmonic signals.

6 Conclusions

In this paper, we have presented an efficient sparse modeling approach for localizing harmonic audio sources using a calibrated sensor array. Assuming that each harmonic components in each pitch can only come from one source, the localization estimate is based on the phase and attenuation information for all of the harmonics jointly. The resulting model phases and attenuation will then depend on the source location. By using sparse modeling, the method inherently estimates both the number of sources, the number of harmonics in each source, as well as the extent of a possibly occurring reverberation. The effectiveness of the resulting algorithm is shown using both simulated and measured audio sources.

7 Acknowledgements

The authors wish to express their gratitude to the Signal Processing Group at Electrical and Information Technology, Lund University, for allowing use of their experimental facilities, as well as to the authors of [15] for sharing their Matlab implementations.

8 Appendix: The Cramér-Rao lower bound

In this appendix, we briefly summarize the Cramér-Rao lower bound (CRB) for the examined localization problem. As is well known, under the assumption of complex circularly symmetric Gaussian distributed noise, the Slepian-Bangs formula yields [11, p. 382]

$$\left[P_{cr}^{-1}\right]_{ij} = \text{trace}\left[\mathbf{\Gamma}^{-1}\mathbf{\Gamma}_i'\mathbf{\Gamma}^{-1}\mathbf{\Gamma}_j'\right] + 2\mathcal{R}\left[\boldsymbol{\mu}_i'^H\mathbf{\Gamma}^{-1}\boldsymbol{\mu}_j'\right] \tag{62}$$

where $\mathcal{R}$ denotes the real part of a complex scalar, $\mathbf{\Gamma}$ the covariance matrix of the noise process, and $\boldsymbol{\mu}$ is the deterministic signal component, with $\mathbf{\Gamma}_i'$ and $\boldsymbol{\mu}_i'$ denoting the derivative of $\mathbf{\Gamma}$ and $\boldsymbol{\mu}$ with respect to element i of the parameter vector, respectively. For the case of uncorrelated noise with a known variance σ^2,

this simplifies to

$$\left[P_{cr}^{-1}\right]_{ij} = 2\mathcal{R}\left[\boldsymbol{\mu}_i'^H \boldsymbol{\mu}_j'\right]/\sigma^2 \tag{63}$$

Using the assumed signal model as measured at sensor m, stacking the the observations as in (2), and then using the vec operator on the resulting matrix results, one obtains the $\boldsymbol{\mu}$ function needed for the CRB calculations. Here, the parameters to be estimated are

$$\boldsymbol{\Delta} = \left\{ \left\{ a_{k,\ell}, \varphi_{k,\ell} \right\}_{\ell=1,\ldots,L_k}, \omega_k, \vartheta_{s,k}, R_{s,k} \right\}_{\substack{s=1,\ldots,S \\ k=1,\ldots,K}} \tag{64}$$

Clearly, the resulting function may easily be derivated with respect to the magnitude, frequency and phase parameters. However, since the location parameter, $\vartheta_{s,k}$ and $R_{s,k}$, enter into the expression in a complicated manner depending on the sensor geometry, the corresponding derivatives are not straight forward for an arbitrary array. For this reason, for the considered array geometries, we here simply approximate the resulting expressions using numerically differentiated expressions.

References

[1] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 2, pp. 148–152, Mar 1996.

[2] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberent rooms," in *Microphone Arrays: Techniques and Applications*, M. Brandstein and D. Ward, Eds., pp. 157–180. Springer-Verlag, New York, 2001.

[3] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 791–803, Nov 2003.

[4] E. Kidron, Y. Y. Schechner, and M. Elad, "Cross-modal localization via sparsity," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1390–1404, April 2007.

[5] M. D. Gillette and H. F. Silverman, "A linear closed-form algorithm for source localization from time-differences of arrival," *IEEE Signal Processing Letters*, vol. 15, pp. 1–4, 2008.

[6] K. C. Ho and M. Sun, "Passive source localization using time differences of arrival and gain ratios of arrival," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 464–477, Feb 2008.

[7] X. Alameda-Pineda and R. Horaud, "A geometric approach to sound source localization from time-delay estimates," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 6, pp. 1082–1095, June 2014.

[8] H. F. Silverman and S. E. Kirtman, "A two-stage algorithm for determining talker location from linear microphone array data," *Computer Speech & Language*, vol. 6, no. 2, pp. 129 – 152, 1992.

[9] H. Krim and M. Viberg, "Two Decades of Array Signal Processing Research," *IEEE Signal Process. Mag.*, pp. 67–94, July 1996.

[10] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part IV, Optimum Array Processing*, John Wiley and Sons, Inc., 2002.

[11] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Prentice Hall, Upper Saddle River, N.J., 2005.

[12] J. Benesty, M. Sondhi, M. Mohan, and Y. Huang, *Springer handbook of speech processing*, Springer, 2008.

[13] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, Springer-Verlag, New York, NY, 1988.

[14] M. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, Morgan & Claypool, 2009.

[15] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear Least Squares Methods for Joint DOA and Pitch Estimation," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 21, no. 5, pp. 923–933, 2013.

[16] S. Gerlach, S. Goetze, J. Bitzer, and S. Doclo, "Evaluation of joint position-pitch estimation algorithm for localising multiple speakers in adverse acoustical environments," in *Proc. German Annual Conference on Acoustics (DAGA)*, Düsseldorf, Germany, 2011, vol. Mar. 2011, pp. 633–634.

[17] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, "Joint DOA and Multi-pitch Estimation Based on Subspace Techniques," *EURASIP J. on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–11, 2012.

[18] J. J. Fuchs, "On the Use of Sparse Representations in the Identification of Line Spectra," in *17th World Congress IFAC*, Seoul, jul 2008, pp. 10225–10229.

[19] I. F. Gorodnitsky and B. D. Rao, "Sparse Signal Reconstruction from Limited Data Using FOCUSS: A Re-weighted Minimum Norm Algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, March 1997.

[20] M. D. Plumbley, S. A. Abdallah, T. Blumensath, and M. E. Davies, "Sparse representations of polyphonic music," *Signal Processing*, vol. 86, no. 3, pp. 417–431, March 2006.

[21] M. Genussov and I. Cohen, "Multiple fundamental frequency estimation based on sparse representations in a structured dictionary," *Digit. Signal Process.*, vol. 23, no. 1, pp. 390–400, Jan. 2013.

[22] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Estimating Multiple Pitches Using Block Sparsity," in *38th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancouver, May 26–31, 2013.

[23] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Multi-Pitch Estimation Exploiting Block Sparsity," *Elsevier Signal Processing*, vol. 109, pp. 236–247, April 2015.

[24] S. L. Marple, "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2600–2603, September 1999.

[25] T. Ballal and C.J. Bleakley, "DOA Estimation of Multiple Sparse Sources Using Three Widely-Spaced Sensors," in *Proceedings of the 17th European Signal Processing Conference*, 2009.

[26] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A Sparse-Group Lasso," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.

[27] M. Elad, *Sparse and Redundant Representations*, Springer, 2010.

[28] E. J. Candes, M. B. Wakin, and S. Boyd, "Enhancing Sparsity by Reweighted l_1 Minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, Dec. 2008.

[29] L. Qing, Z. Wen, and W. Yin, "Decentralized jointly sparse optimization by reweighted ell-q minimization," *Signal Processing, IEEE Transactions on*, vol. 61, no. 5, pp. 1165–1170, March 2013.

[30] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Comm. Pure Appl. Math.*, vol. 63, 2010.

[31] N. R. Butt, S. I. Adalbjörnsson, S. D. Somasundaram, and A. Jakobsson, "Robust Fundamental Frequency Estimation in the Presence of Inharmonicities," in *38th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancouver, May 26–31, 2013.

[32] O. Besson and P. Stoica, "Exponential signals with time-varying amplitude: parameter estimation via polar decomposition," *Signal Processing*, vol. 66, pp. 27–43, 1998.

[33] Inc. CVX Research, "CVX: Matlab Software for Disciplined Convex Programming, version 2.0 beta," `http://cvxr.com/cvx`, Sept. 2012.

[34] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pp. 95–110. Springer-Verlag Limited, 2008, http://stanford.edu/∼boyd/graph_dcp.html.

[35] J. F. Sturm, "Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11-12, pp. 625–653, August 1999.

[36] R. H. Tutuncu, K. C. Toh, and M. J. Todd, "Solving semidefinite-quadratic-linear programs using SDPT3," *Mathematical Programming Ser. B*, vol. 95, pp. 189–217, 2003.

[37] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[38] N. Parikh and S. Boyd, "Proximal Algorithms," *Found. Trends Optim.*, vol. 1, pp. 127–239, 2014.

[39] Z. Simayijiang, F. Andersson, Y. Kuang, and K. Åström, "An automatic system for microphone self-localization using ambient sound," in *European Signal Processing Conference (Eusipco 2014)*, 2014.

[40] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 520–529, Sept 2004.

[41] D. Gatica-Perez, G. Lathoud, J. Odobez, and I. McCowan, "Audiovisual Probabilistic Tracking of Multiple Speakers in Meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 601–616, Feb 2007.

C

Paper C

Joint DOA and Multi-Pitch Estimation via Block Sparse Dictionary Learning

Ted Kronvall, Stefan Ingi Adalbjörnsson,
and Andreas Jakobsson

Centre for Mathematical Sciences, Lund University, Lund, Sweden

Abstract

In this paper, we introduce a novel sparse method for joint estimation of the direction of arrivals (DOAs) and pitches of a set of multi-pitch signals impinging on a sensor array. Extending on earlier approaches, we formulate a novel dictionary learning framework from which an estimate is formed without making assumptions on the model orders. The proposed method alternatively uses a block sparse approach to estimate the pitches, using an alternating direction method of multipliers framework, and alternatively a nonlinear least squares approach to estimate the DOAs. The preferable performance of the proposed algorithm, as compared to earlier methods, is shown using numerical examples.

Keywords: multi-pitch estimation, block sparsity, dictionary learning, ADMM, direction-of-arrival.

1 Introduction

The estimation of fundamental frequencies, or pitches, of harmonically related, and often acoustic, signals is a common problem occurring in various forms of applications, and perhaps most notably so in audio processing (see, e.g., [1] and the references therein). Due to the importance of such applications, there have been notable contributions on pitch estimation for signals containing both single and multiple pitches (see e.g., [2–5]). By using an array of several sensors, one may exploit the relative time-delay information at the different sensors to determine the location of the impinging sound sources. Commonly, existing techniques, as the ones in, e.g., [6–8], make strong *a priori* assumptions on the model structure of the impinging signals, such as the number of pitches, as well as the number of harmonics in each pitch. Alternatively, model order information criterias may be used to determine the appropriate model order, such as in [9, 10], or by applying an optimal filtering approach reminiscent to the one proposed in [11]. In this work, we extend on the method presented in [5], and propose a novel joint DOA and pitch estimation technique, formed by using a novel sparse signal reconstruction framework. The technique is reminiscent to the one presented in [12], wherein the solution space is expanded to a large dictionary of candidate fundamental frequencies, from where a small number of pitches which have the best fit to the data are chosen. As the data is measured with several sensors, where each has a phase offset according the specific geometry of the array and the location of the sound source, both the pitches and the sensor phases must be estimated jointly. Such a joint estimation typically requires solving a non-convex optimization problem. Herein, we avoid this difficult by applying a dictionary learning technique, reminiscent to the ones presented in [13, 14]. We thereby split the problem into two subproblems, allowing for an iterative refinement of the pitch estimates, formed using an alternating direction method of multipliers (ADMM) framework, and of the DOA estimates, using a nonlinear least squares (NLS) formulation. The method allows for the estimation of the DOAs and pitches from multi-pitch signals originating from one or more locations, without having to know the number of sources, pitches, or their respective number of harmonics. Our claims are illustrated using numerical simulations of audio signals, comparing the achieved performance to other recent estimators.

2 Pitch-DOA signal model

Consider K complex-valued and harmonically related acoustic signals impinging on an array of sensors, corrupted by additive noise and interference, such that the signal measured at the mth sensor may be well modelled as [6, 15]

$$y_m(t) = \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} c_m d_{k,\ell} e^{j\omega_k \ell(t+\tau_{k,m})} + e_m(t) \tag{1}$$

where $d_{k,l}$ is the complex-valued amplitude of the ℓth harmonic of the kth pitch, whereas L_k and ω_k are the number of harmonics and the pitch of the kth signal source, respectively. Furthermore, let $e_m(t)$ denote the additive noise term, c_m the sensor gain, and $\tau_{k,m}$ the time-of-arrival for the kth signal source. Define the measurement matrix

$$\mathbf{Y} = \left[\begin{array}{ccc} \mathbf{y}(1) & \dots & \mathbf{y}(N) \end{array} \right]^T \tag{2}$$

where, at each time point, $n = 1, \dots, N$, the data snapshot is found as

$$\mathbf{y}(t) = \left[\begin{array}{ccc} y_0(t) & \dots & y_{M-1}(t) \end{array} \right]^T$$

with $(\cdot)^T$ denoting the transpose. Then, (2) may be concisely expressed as

$$\mathbf{Y} = \sum_{k=1}^{K} \mathbf{W}_k \mathrm{diag}(\mathbf{d}_k) \mathbf{F}_k(\boldsymbol{\tau}_k) \mathbf{C} + \mathbf{E} \tag{3}$$

where $\mathbf{E}$ denotes the combined noise term constructed in the same manner as $\mathbf{Y}$, and

$$\mathbf{W}_k = \left[\begin{array}{ccc} \mathbf{w}_k & \dots & \mathbf{w}_k^{L_k} \end{array} \right] \tag{4}$$

$$\mathbf{w}_k = \left[\begin{array}{ccc} e^{j\omega_k} & \dots & e^{j\omega_k N} \end{array} \right]^T \tag{5}$$

$$\mathbf{d}_k = \left[\begin{array}{ccc} d_{k,1} & \dots & d_{k,L_k} \end{array} \right]^T \tag{6}$$

$$\mathbf{F}_k(\boldsymbol{\tau}_k) = \left[\begin{array}{ccc} e^{j\omega_k \tau_{k,1}} & \dots & e^{j\omega_k \tau_{k,M}} \\ e^{j\omega_k 2\tau_{k,1}} & \dots & e^{j\omega_k 2\tau_{k,M}} \\ \vdots & \ddots & \vdots \\ e^{j\omega_k L_k \tau_{k,1}} & \dots & e^{j\omega_k L_k \tau_{k,M}} \end{array} \right] \tag{7}$$

$$\boldsymbol{\tau}_k = \begin{bmatrix} \tau_{k,1} & \dots & \tau_{k,M} \end{bmatrix}^T \tag{8}$$

$$\mathbf{C} = \operatorname{diag}\left(\begin{bmatrix} c_1 & \dots & c_M \end{bmatrix}\right) \tag{9}$$

such that diag$(\cdot)$ is a diagonal matrix. One may note that $\mathbf{W}_k$, for $k = 1, \dots, K$, consists of stacked Fourier vectors, for each harmonic of a pitch in the temporal domain, whereas $\mathbf{F}_k$ consists of stacked Fourier vectors (or array transfer vectors) in the spatial domain with respect to the time-of-arrivals, $\boldsymbol{\tau}_k$, repeated for each pitch k and its L_k harmonics. We proceed to reformulating the problem in (3) using a sparse estimation framework, reminiscent to the one presented in [12], extending the representation of the K pitches onto a large dictionary of P candidate fundamental frequencies, $\omega_1, \dots, \omega_P$, where $P \gg K$, chosen so large that K of these will reasonably well coincide with the true pitches in the signal. In the same fashion, the number of harmonics of each pitch, L_p, is extended to an arbitrary upper level, say $L_{\max}$, for all dictionary elements, $p = 1, \dots, P$. One can, without loss of generality, assume $\mathbf{C} = \mathbf{I}$, i.e., that the data measurement matrix has been preconditioned to account for different gain at different sensors. The signal model may thus be expressed as

$$\begin{aligned} \mathbf{Y} &= \sum_{p=1}^{P} \mathbf{W}_p \operatorname{diag}(\mathbf{d}_k) \mathbf{F}_p(\tau_p) + \mathcal{E} && (10) \\ &= \mathcal{W} \operatorname{diag}(\mathbf{d})\, \mathcal{F}(\boldsymbol{\tau}) + \mathcal{E} && (11) \end{aligned}$$

where the block dictionary matrices are formed by stacking the matrices such that

$$\begin{aligned} \mathcal{W} &= \begin{bmatrix} \mathbf{W}_1 & \dots & \mathbf{W}_P \end{bmatrix} && (12) \\ \mathcal{F}(\boldsymbol{\tau}) &= \begin{bmatrix} \mathbf{F}_1(\tau_1)^T & \dots & \mathbf{F}_P(\tau_P)^T \end{bmatrix}^T && (13) \end{aligned}$$

where $\mathcal{W} \in \mathbb{C}^{N \times PL_{\max}}$, $\mathcal{F}(\boldsymbol{\tau}) \in \mathbb{C}^{PL_{\max} \times M}$, and

$$\begin{aligned} \mathbf{d} &= \begin{bmatrix} \mathbf{d}_1{}^T & \dots & \mathbf{d}_P{}^T \end{bmatrix}^T && (14) \\ \boldsymbol{\tau} &= \begin{bmatrix} \boldsymbol{\tau}_1 & \dots & \boldsymbol{\tau}_P \end{bmatrix}^T && (15) \end{aligned}$$

with $\mathbf{d} \in \mathbb{C}^{PL_{\max} \times 1}$ and $\boldsymbol{\tau} \in \mathbb{R}^{P \times M}$. The resulting signal formulation provides a more structured framework than the one presented in [15], separating the complex-valued amplitudes, $\mathbf{d}$, and the sensor offsets in $\mathcal{F}(\boldsymbol{\tau})$. If the sensor array is assumed to be a uniform linear array (ULA), the time-of-arrivals may be related to the corresponding DOA as [9]

$$\tau_{k,m} = (m-1)\delta \sin(\vartheta_k)\gamma^{-1} \tag{16}$$

with δ, γ, and ϑ denoting the uniform distance between sensors, the wave propagation velocity, and the DOA respectively. The $P \times M$ time-of-arrivals may thus be expressed as a function of the set of DOAs

$$\boldsymbol{\vartheta} = \begin{bmatrix} \vartheta_1 & \dots & \vartheta_P \end{bmatrix}^T \tag{17}$$

In the interest of notational simplicity, we hereafter use only the dependency of $\boldsymbol{\vartheta}$ instead of $\boldsymbol{\tau}(\boldsymbol{\vartheta})$. For other array geometries, one may replace (16) with another function mapping from directionality or location to the time-of-arrival.

3 Dictionary learning approach

In order to form the estimate of the unknown DOAs and pitches, we formulate the estimates as the solution to a group sparse minimization reminiscent to the scheme presented in [5], such that

$$\begin{aligned} \underset{\boldsymbol{\vartheta},\mathbf{d}}{\text{minimize}} \;\; & \frac{1}{2}\left\| \mathbf{Y} - \mathcal{W}\,\text{diag}(\mathbf{d})\,\mathcal{F}(\boldsymbol{\vartheta}) \right\|_{\mathcal{F}}^2 \\ & + \lambda\mu \sum_{p=1}^{P} \|\mathbf{d}_k\|_2 + \lambda(1-\mu)\|\mathbf{d}\|_1 \end{aligned} \tag{18}$$

where block sparsity is imposed on $\mathbf{d}$, such that the number of pitches, as well as the number of harmonics within each pitch, are sparse. Here, we set $\lambda > 0$ as a parameter weighting the degree of sparsity to the fit of the solution, while $\mu \in [0\,,\ 1]$ prioritizes between sparsity and block sparsity. In order to simplify the minimization, one may formulate (18) equivalently as

$$\begin{aligned} \underset{\boldsymbol{\vartheta},\mathbf{d}}{\text{minimize}} \;\; & \frac{1}{2}\sum_{m=1}^{M} \left\| \mathbf{y}_m - \mathcal{W}\,\text{diag}\left(f_m(\boldsymbol{\vartheta})\right)\mathbf{d} \right\|_2^2 \\ & + \lambda\mu \sum_{p=1}^{P} \|\mathbf{d}_k\|_2 + \lambda(1-\mu)\|\mathbf{d}\|_1 \end{aligned} \tag{19}$$

such that the minimization is formed by summing the squared residual errors sensor by sensor, where $f_m(\cdot)$ is the mth column of $\mathcal{F}(\cdot)$, and where we have used that $\text{diag}\left(f_m(\boldsymbol{\vartheta})\right)\mathbf{d} = \text{diag}(\mathbf{d})f_m(\boldsymbol{\vartheta})$. However, solving (19) is a hard problem, as $f(\cdot)$ is a non-convex function of $\boldsymbol{\vartheta}$, as is its product with $\mathbf{d}$. On the other

Algorithm 1 The IAPEBS algorithm

1: Initiate $\mathbf{d}^{(0)}$ by taking steps 4-11 for data $\mathbf{y}_1$ only.
2: Set $k = 0$
3: **repeat** {Dictionary learning scheme}
4: Take NLS step $\vartheta^{(k+1)} = \underset{\vartheta}{\operatorname{argmin}}\, q(\vartheta, \mathbf{d}^{(k)})$
5: Initiate $\mathbf{u}^{(0)} = \mathbf{d}^{(k)}, \mathbf{z}^{(0)} = \mathbf{z}^{(\text{save})}, i = 0$
6: **repeat** {ADMM scheme}
7: $\mathbf{z}^{(i+1)} = \underset{\mathbf{z}}{\operatorname{argmin}}\, L_\varkappa(\mathbf{z}, \mathbf{u}^{(i)}, \mathbf{d}^{(i)})$
8: $\mathbf{u}^{(i+1)} = \underset{\mathbf{u}}{\operatorname{argmin}}\, L_\varkappa(\mathbf{z}^{(i+1)}, \mathbf{u}, \mathbf{d}^{(i)})$
9: $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - (\mathbf{z}^{(i+1)} - \mathbf{u}^{(i+1)})$
10: $i \leftarrow i + 1$
11: **until** convergence
12: Set $\mathbf{d}^{(k+1)} = \mathbf{u}^{(\text{end})}$, and $\mathbf{z}^{(\text{save})} = \mathbf{z}^{(\text{end})}$
13: $k \leftarrow k + 1$
14: **until** convergence

hand, for a fixed ϑ, the minimization is the ordinary LASSO with block sparsity for complex sinusoids (see, e.g., [16]), where $\mathcal{W}\,\text{diag}\left(f_m(\vartheta)\right)$ may be seen as a phase-shifted dictionary at sensor m with respect to the corresponding DOA. Adopting a dictionary learning framework reminiscent to the one used in [13,14], the problem is split in two sub-problems. In the first, we fix the DOAs, and (19) may be solved via one of the freely available interior point solvers, such as SeDuMi [17] and SDPT3 [18]. However, such solvers will typically scale poorly with increasing data length, the use of a finer grid of candidate pitches, and/or the number of sensors. Such methods may thus in many cases be computationally cumbersome, and we here introduce an efficient ADMM-based formulation of (19). To do so, one splits the objective function into two parts, where we let one contain the squared residual error, and the second the sparsity constraints, whereafter an auxiliary variable is introduced, such that

$$\underset{\mathbf{z},\mathbf{u}}{\text{minimize}}\; g_1(\mathbf{z}) + g_2(\mathbf{u}) \text{ subj. to } \mathbf{z} - \mathbf{u} = \mathbf{0} \tag{20}$$

since only $\mathbf{z} = \mathbf{u}$ is a feasible point, and where

$$g_1(\mathbf{z}) = \frac{1}{2}\sum_{m=1}^{M}\left\|\mathbf{y}_m - \mathcal{W}\,\mathrm{diag}\left(f_m(\vartheta)\right)\mathbf{z}\right\|_2^2 \tag{21}$$

$$g_2(\mathbf{u}) = \lambda\mu\sum_{p=1}^{P}\|\mathbf{u}_k\|_2 + \lambda(1-\mu)\|\mathbf{u}\|_1 \tag{22}$$

are convex functions. Under the assumption that there is no duality gap, which, for a fixed ϑ, is true for (18), one may solve the optimization problem via the dual function, defined as the infimum of the augmented Lagrangian with respect to $\mathbf{z}$ and $\mathbf{u}$, i.e., [19]

$$L_\varkappa(\mathbf{z},\mathbf{u},\mathbf{x}) = g_1(\mathbf{z}) + g_2(\mathbf{u}) + \mathbf{x}^T(\mathbf{z}-\mathbf{u}) + \frac{\varkappa}{2}||\mathbf{z}-\mathbf{u}||_2^2$$

where $\mathbf{x}$ is the dual variable. The ADMM method solves this iteratively by, at step $i+1$, minimizing the Lagrangian for one primal variable while holding the other fixed at its previous value, and then updating the dual variable by taking a gradient ascent step and maximizing the dual function, i.e.,

$$\mathbf{z}^{(i+1)} = \arg\min_{\mathbf{z}} L_\varkappa(\mathbf{z},\mathbf{u}^{(i)},\mathbf{d}^{(i)}) \tag{23}$$

$$\mathbf{u}^{(i+1)} = \arg\min_{\mathbf{u}} L_\varkappa(\mathbf{z}^{(i+1)},\mathbf{u},\mathbf{d}^{(i)}) \tag{24}$$

$$\tilde{\mathbf{x}}^{(i+1)} = \tilde{\mathbf{x}}^{(i)} - \varkappa(\mathbf{z}^{(i+1)} - \tilde{\mathbf{x}}^{(i+1)}) \tag{25}$$

where $\varkappa$ is the step size for maximizing the dual function, and $\tilde{\mathbf{x}} = \mathbf{x}/\varkappa$ is the scaled version of the dual variable, which is more convenient for implementation (see [19] for further details on these aspects). The function in (23), which is quadratic, can be solved in closed form as

$$\mathbf{z}^{(i+1)} = \left(\sum_{m=1}^{M}\tilde{\mathcal{W}}_m^H\tilde{\mathcal{W}}_m + \varkappa\mathbf{I}_{PL_{\max}}\right)^{-1} \times \left(\sum_{m=1}^{M}\tilde{\mathcal{W}}_m^H\mathbf{y}_m + \mathbf{u}^{(i)} + \mathbf{x}^{(i)}\right) \tag{26}$$

where $\tilde{\mathcal{W}}_m = \mathcal{W}\,\mathrm{diag}\left(f_m(\vartheta)\right)$ denotes the phase-shifted dictionary at sensor m. The function in (23), i.e., the primal variable for the sparsity constraints, is

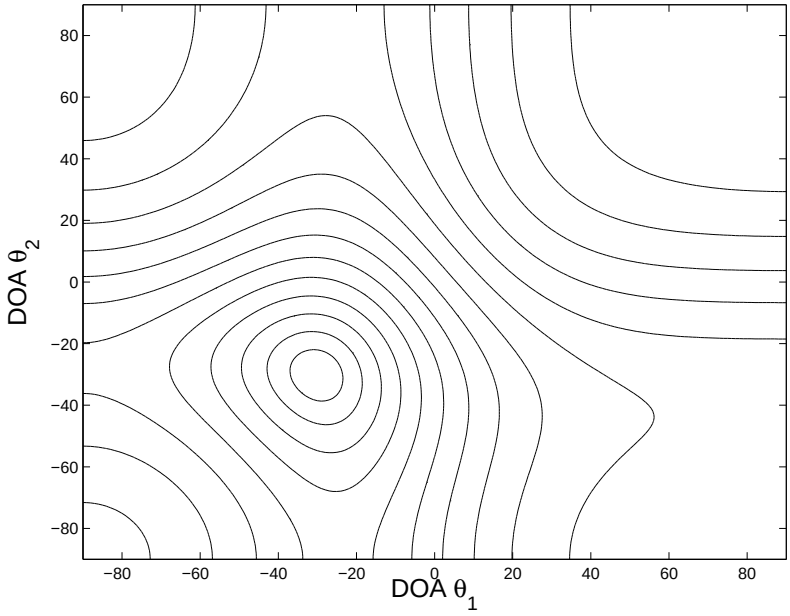


Figure 1: Level curves for the function in (28), for a multi pitch signal containing two pitches, both originating from $-30°$.

obtained by solving sub-differential equations, yielding

$$\mathbf{u}^{(i+1)} = h\left(h'\left(\mathbf{z}^{(i+1)} - \mathbf{x}^{(i)}, \lambda\mu\right), \lambda(1-\mu)\right) \tag{27}$$

where $h\left(\mathbf{b}, \xi\right) = \mathbf{b}\left(1 - \xi / \left\|*\right\| \mathbf{b}_2\right)^+$, for a vector $\mathbf{b}$ and a positive scalar ξ, with $(\cdot)^+$ denoting the identity function for finite values and zero otherwise, and $h'(\cdot)$ defined similarly but operate element-wise on $\mathbf{b}$ (see also [5]). The resulting estimate of $\mathbf{d}^{(k)}$ is then inserted into the second subproblem of the dictionary learning scheme, i.e.,

$$q(\vartheta, \mathbf{d}^{(k)}) = \frac{1}{2}\left\|\mathbf{Y} - \mathcal{W}\,\mathrm{diag}(\mathbf{d}^{(k)})\,\mathcal{F}(\vartheta)\right\|_{\mathcal{F}}^2 \tag{28}$$

which is minimized for ϑ, and is equivalent to performing a dictionary learning update to the phase-shifted dictionary, $\tilde{\mathcal{W}}_m$, which was used in the ADMM

procedure, i.e., (20)-(27). Figure 1 illustrates the cost function in (28) after a few dictionary learning iterations of the proposed algorithm, showing that although the cost function will not be convex, it is unimodal for DOAs in the range $[-90, 90]^\circ$ and may thus be easily solved using a few iterations of, for instance, Newton-Raphson's method. To summarize, an algorithm outline of the proposed metod is stated in Algorithm 1, where it may be noted that the inner ADMM scheme takes fewer and fewer steps at every iteration of the outer dictionary learning scheme, until convergence is reached and only a single ADMM step is taken. The sparsity parameter λ is chosen with cross validation in a similar fashion as performed in [20], but the estimates are rather unsensitive with respect to this choice. The proposed method requires estimating a total of $PL_{\max} + M$ parameters, which is considerably fewer than the recent sparse method presented in [15], which required estimating $PL_{\max}M$ parameters.

4 Numerical results

We proceed to illustrate the performance of our proposed method, as compared to other recent methods, using synthetic audio signals. As the fundamental frequencies are estimated on a discrete dictionary grid, comparison is made using a percentage within limits (PWL) metric, defined as the ratio of pitch estimates within a range of $\pm 1/4$ Hz from the true values. For DOA comparison, the total root mean square error (RMSE) is used for all sources, defined as

$$\mathrm{RMSE}_\vartheta = \sqrt{\frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\left(\hat{\vartheta}_{k,i} - \vartheta_k\right)^2} \tag{29}$$

where n is the number of Monte Carlo (MC) simulation estimates, and K is the number of pitches in the signal. Figure 2 shows the PWL of the fundamental frequency, as well as the RMSE for the DOA, for a signal containing a single pitch with $f_1 = 220$ Hz and $L_1 = 7$ harmonics, impinging on a 5-sensor ULA from direction $\vartheta_1 = -30^\circ$. These results have been computed using 250 MC simulations, assuming a sampling frequency of $f_s = 8820$ Hz, a sound wave propagation velocity of $\gamma = 324.3$ m/s, and a sensor spacing of $\delta = \gamma/f_s = 3.84$ cm. The sensor gains may be obtained from a covariance matrix estimate on the measurement matrix $\mathbf{Y}$, but are, in these simulations and without loss of generality, set to $c_1 = \cdots = c_M = 1$. The figures show the performance for

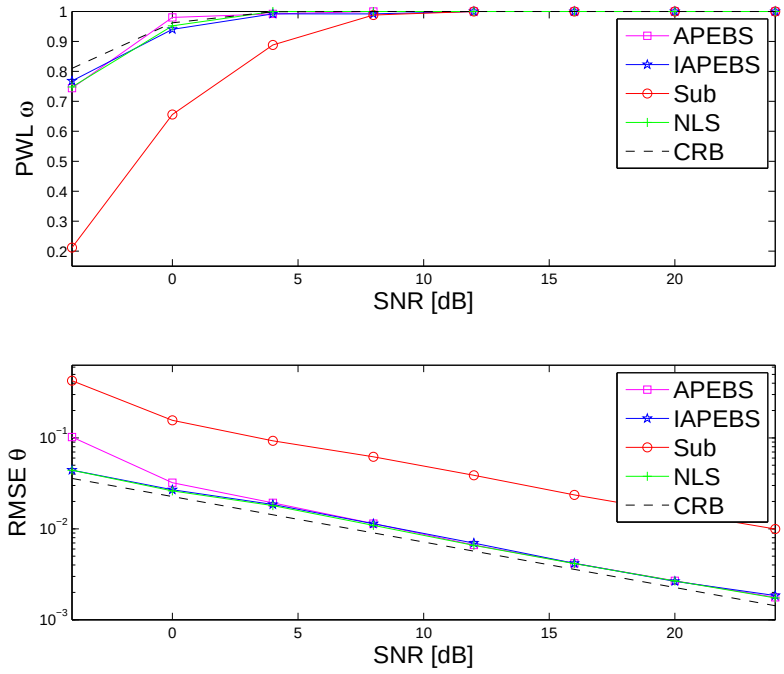


Figure 2: The PWL and RMSE for a single-pitch signal as compared with the optimal performance of an estimator reaching the CRB.

growing signal-to-noise ratios (SNRs), defined as

$$\mathrm{SNR} = 10 \cdot \log\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right) \quad \text{(dB)} \tag{30}$$

As is clear from the figure, the proposed method, here termed the iterative array DOA and pitch estimator using block sparsity (IAPEBS), performs similarly to the recently proposed APEBS estimator [15], and the NLS-based estimator proposed in [6], achieving a performance close to the Cramér-Rao Lower Bound (CRLB). The subspace-based method (Sub), also introduced in [6], is found to yield a somewhat lower performance. Figure 3 shows the corresponding performance for a multi-pitch signal consisting of two pitches, with $[\omega_1, \omega_2] = [150, 220]$ Hz, and with $[L_1, L_2] = [7, 6]$ harmonics, impinging from directions $[\vartheta_1, \vartheta_2] =$

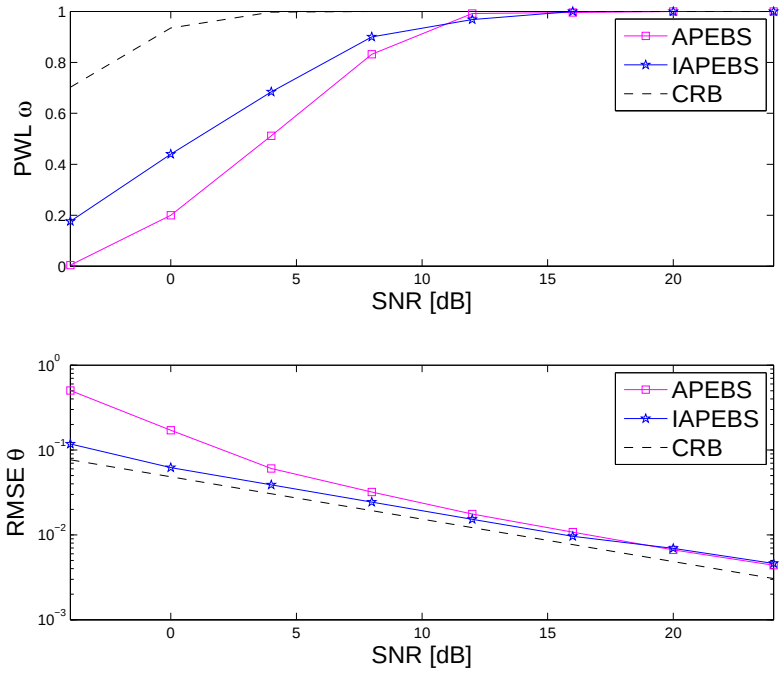


Figure 3: The PWL and RMSE for a multi-pitch signal with two pitches, as compared to the corresponding CRB.

$[-30, -30]^\circ$. As the NLS and Sub estimators only allow for single pitch signals, the figure only shows the performance of IAPEBS, as compared with APEBS and the corresponding CRB. As is clear from the figures, the IAPEBS estimator yields highly accurate parameter estimates, almost reaching the CRBs, notably improving the achievable performance as compared to the APEBS estimator, which decouples the estimation into first estimating the pitches, whereafter the DOAs are determined in a second step. This should be compared with the here proposed iterative estimation scheme, which enables a better joint estimation of pitch and DOA.

References

[1] M. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, Morgan & Claypool, 2009.

[2] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 25, no. 1, pp. 24–33, 1977.

[3] E. Benetos and S. Dixon, "Joint Multi-Pitch Detection Using Harmonic Envelope Estimation for Polyphonic Music Transcription," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1111–1123, Oct. 2011.

[4] C. Lee, Y. Yang, and H. H. Chen, "Multipitch Estimation of Piano Music by Exemplar-Based Sparse Representation," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 608–618, 2012.

[5] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Estimating Multiple Pitches Using Block Sparsity," in *38th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancouver, May 26–31, 2013.

[6] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear Least Squares Methods for Joint DOA and Pitch Estimation," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 21, no. 5, pp. 923–933, 2013.

[7] S. Gerlach, S. Goetze, J. Bitzer, and S. Doclo, "Evaluation of joint position-pitch estimation algorithm for localising multiple speakers in adverse acoustical environments," in *Proc. German Annual Conference on Acoustics (DAGA)*, Düsseldorf, Germany, 2011, vol. Mar. 2011, pp. 633–634.

[8] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, "Joint DOA and Multi-pitch Estimation Based on Subspace Techniques," *EURASIP J. on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–11, 2012.

[9] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Prentice Hall, Upper Saddle River, N.J., 2005.

[10] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Sinusoidal Order Estimation using Angles between Subspaces," *EURASIP Journal on Advances in Signal Processing*, 2009, Article ID 948756, 11 pages.

[11] M. G. Christensen, J. .L Højvang, A. Jakobsson, and S. H. Jensen, "Joint fundamental frequency and order estimation using optimal filtering," *EURASIP Journal on Advances in Signal Processing*, vol. 13, pp. 1–18, 2011.

[12] J. J. Fuchs, "On the Use of Sparse Representations in the Identification of Line Spectra," in *17th World Congress IFAC*, Seoul, jul 2008, pp. 10225–10229.

[13] C. D. Austin, J. N. Ash, and R. L. Moses, "Dynamic Dictionary Algorithms for Model Order and Parameter Estimation," *IEEE Transactions on Signal Processing*, vol. 61, no. 20, pp. 5117–5130, October 2013.

[14] I. Tosic and P. Frossard, "Dictionary Learning," *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 27–38, March 2011.

[15] T. Kronvall, S. I. Adalbjörnsson, and A. Jakobsson, "Joint DOA and Multi-Pitch Estimation Using Block Sparsity," in *39th IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Florence, May 4-9 2014.

[16] J. Friedman, T. Hastie, and R. Tibshirani, "A Note on the Group Lasso and a Sparse Group Lasso," Unpublished manuscript, 2010.

[17] J. F. Sturm, "Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11-12, pp. 625–653, August 1999.

[18] R. H. Tutuncu, K. C. Toh, and M. J. Todd, "Solving semidefinite-quadratic-linear programs using SDPT3," *Mathematical Programming Ser. B*, vol. 95, pp. 189–217, 2003.

[19] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[20] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society B*, vol. 58, no. 1, pp. 267–288, 1996.

D

Paper D

Sparse Modeling of Chroma Features

Ted Kronvall, Maria Juhlin, Johan Swärd,
Stefan Ingi Adalbjörnsson, and Andreas Jakobsson

Centre for Mathematical Sciences, Lund University, Lund, Sweden

Abstract

This work treats the estimation of chroma features for harmonic audio signals using a sparse reconstruction framework. Chroma has been used for decades as a key tool in audio analysis, and is typically formed using a periodogram-based approach that maps the fundamental frequency of a musical tone to its corresponding chroma. Such an approach often leads to problems with tone ambiguity, which we adress via sparse modeling, allowing us to appropriately penalize ambiguous estimates while taking the harmonic structure of tonal audio into account. Furthermore, we also allow for signals to have time-varying envelopes. Using a spline-based amplitude modulation of the chroma dictionary, the presented estimator is able to model longer frames than what is conventional for audio, as well as to model highly time-localized signals, and signals containing sudden bursts, such as trumpet or trombone signals. Thus, we may retain more signal information as compared to alternative methods. The performance of the proposed methods is evaluated by analyzing average estimation errors for synthetic signals, as compared to the Cramér-Rao lower bound, and by visual inspection for estimates of real instrument signals, showing strong visual clarity, as compared to other commonly used methods.

Keywords: Chroma, multi-pitch estimation, sparse modeling, amplitude modulation, block sparsity, ADMM

1 Introduction

Music is an art-form that people have enjoyed for millennia. Perhaps music is even enjoyed more today, as the development of personalized computers and smart telephones have enabled ubiquitous music listening, automatic identification of songs, or even the chance for anyone to be a self-made DJ. When listening, learning, composing, mixing, and identifying music, there are a number of musical features one may utilize (see, e.g. [1]). One of the fundamental building blocks in music, the musical note, is a periodic sound, typically characterized by its pitch, timbre, intensity, and duration. For transcription purposes, i.e., to separate one tone from another, pitch serves as the common descriptor. From a conventional perspective, pitch is measured on an ordinal scale, at which a pitch is humanly perceived as either higher, lower, or the same as another pitch. However, from a perspective of scientific audio analysis, pitches are quantified using an interval scale, in which its spectral distribution of energy is modeled. A single pitch may be seen as a superposition of several narrowband spectral peaks, which are approximately integer multiples of a fundamental frequency. Thus, we refer to the group of frequencies as the pitch, and to each frequency component as the harmonic, or, alternatively, as the partial harmonic. As to the fundamental frequency, it is either the lowest partial, or, if that partial is missing, the smallest spectral distance between adjacent partials. The number of harmonics in a certain pitch, as well as the relative power between these, varies greatly between different sounds, as well as over time. Identifying pitches in a way similar to our human perception has proved to be a difficult estimation problem. Partly, this difficulty is due to coinciding frequency components between certain pitches. For instance, two pitches, where one has exactly twice the fundamental frequency of the other, are referred to as being octave equivalent, as the relative distance by a factor of two is called an octave. These will typically share a large number of partials, often making an estimation procedure ambiguous between octaves. To further complicate matters, other pairs of pitches may also have many coinciding partials, and these are typically found together in audio, an aspect which is referred to as harmony, since they are perceptually pleasant to hear [2]. Jointly estimating several pitches in a signal, i.e., multi-pitch estimation, has been thoroughly examined in the literature (see e.g., [3–5], and the references therein). However, separating intricate combinations of frequency components into multiple pitches often proves difficult, even if the harmonic structure of each musical tone is taken into account. Typically, issues arise when the complexity of the audio signal increases, such that

there are simultaneously two or more pitches with overlapping spectral content present, for instance played by two or more instruments. In the Western musicological system, the frequency interval corresponding to an octave is discretized into twelve intervals, called semi-tones. By gathering all pitches with octave equivalence to their respective semi-tone, these form twelve groups of pitches, called chroma. As octave equivalent pitches share a large number of harmonics, the notion of chroma is thus a method for grouping together those pitches which are perceived as most similar. Therefore, chroma features are widely used in applications such as cover song detection, transcription, and recommender systems (see, e.g. [6–8]). Most methods for chroma estimation begin by obtaining estimates of the pitches in a signal, which are then mapped into their respective chroma. Some of these take the harmonic structure into account, and others do not. The commonly used method by Ellis [9] is formed via a time-smoothed version of the Short-Time Fourier Transform (STFT), whereas the CP and CENS methods by Müller and Ewert [10] use a filterbank approach. The method in [11] uses a sparse methodology, and the method in [12] uses a non-negative least squares approach. Neither of these take the harmonic structure of pitches into account. Other approaches instead allow for the harmonic structure, such as the method presented in [13], which uses a comb filtering technique, and the method in [14], in which post-processing on the periodogram is performed. Most existing methods have in common that their estimates are not directly formed from the actual data, but rather on a representation of these measurements, such as, for instance, using the STFT or the magnitude of the periodogram. Herein, we propose to estimate the chroma using a sparse model reconstruction framework, where explicit model orders are not required. The estimate is found as the solution to a convex optimization problem, where the solution is obtained as a linear combination of an over-complete chroma-based set of Fourier basis functions. Overfitting is avoided by introducing convex penalties promoting solutions having the sought chroma structure. The model orders are thus set implicitly, using tuning parameters which may be obtained using cross-validation, or by utilizing some simple heuristics. In this paper, we generalize upon the work in [5], taking into account the chroma structure, as well as allowing the frequency components to have time-varying amplitudes. The proposed extension increases robustness, as it allows for highly non-stationary signals, or signals with sudden bursts, like trumpets, whose nature may easily be misinterpreted when using ordinary chroma selection techniques. As in [15], the extended model uses a spline basis to detail the time-

varying envelope of the signal, thereby enabling the amplitudes to evolve smoothly with time. The theoretical performance of the proposed estimator is verified using synthetic signals, which are compared to the Cramér-Rao Lower Bound (CRLB), which we here present for the chroma signal model. The practical use of the proposed estimator is illustrated using some excerpts from a recorded trumpet signal, showing an increased visual performance, as compared to some typical reference methods.

2 The chroma signal model

A sound signal typically contains a broad band of frequency content. However, for tonal audio, it is well-known that a predominant part of the spectral energy is confined to a small number of frequency locations. Let $\psi(f,\ell)$ denote the function which describes the frequency of the ℓ:th component. If this function is known, the entire group of components, or partials, representing a musical tone may be described by their fundamental frequency, f. Many oscillating sources, such as, for instance, the human vocal tract and stringed, or wind, instruments, emit tonal audio where the partials are integer multiples of the fundamental, i.e.,

$$\psi(f,\ell) = f\ell, \quad \ell \in \mathcal{L} \subseteq \mathbb{N} \tag{1}$$

where $\mathcal{L}$ denotes the index set of partials present in the signal. However, for an arbitrary $\mathcal{L}$, the definition in (1) is not sufficient to uniquely describe a pitch, as the set of frequencies may map to infinitely many combinations of f and $\mathcal{L}$. For example, for any $n \in \mathbb{N}$, the two pitches

$$\Psi = \{\psi(f,\ell) : f \in \mathbb{R}, \ell \in \mathcal{L} \subseteq \mathbb{N}\} \tag{2}$$

$$\Psi' = \left\{\psi(f',\ell') : f' = \frac{f}{n}, \ell' \in \mathcal{L}' = \{n\ell : \ell \in \mathcal{L}\}\right\} \tag{3}$$

have identical frequency components. Therefore, some constraints need to be imposed on $\mathcal{L}$. A common assumption for pitches is spectral smoothness of the harmonics, i.e., that adjacent harmonics should be of comparable magnitude [16]. This implies that $\mathcal{L}$ typically has few missing harmonics, and that n is as small as possible. However, in some signals, the first harmonic might be missing, so rather than defining the pitch as the signal's smallest frequency component, we define the fundamental frequency more rigorously. If the set of frequencies in a pitch

may be described by (2), then for any $n \in \mathbb{Q}$, the fundamental frequency is the largest $f' = f/n$ which fulfill (3), i.e., which ensures that $\mathcal{L}' = \{n\ell, \ell \in \mathcal{L}\} \subseteq \mathbb{N}$. The index set therefore plays a vital role in the definition of the pitch frequency. Furthermore, because of the harmonic structure, many different pitches will have coinciding partials. To illustrate this, consider two pitches

$$\Psi = \{\psi(f, \ell) : f \in \mathbb{R}, \ell \in \{1, 2, \ldots, L\}\} \tag{4}$$

$$\Psi' = \left\{ \psi(f', \ell') : f' = \frac{f}{n}, \ell' \in \{1, 2, \ldots, nL\} \right\} \tag{5}$$

which consist of all harmonics from $\ell = 1$ up to L and nL, respectively. Here, n may be a rational number, as long as (5) is fulfilled. Indeed, both pitches are unique according to our definition. Still, they will share a large number of harmonics, in fact L of them, as Ψ forms a perfect subset of Ψ', i.e., $\Psi \in \Psi'$, and they will also, as sounds, be perceived as being similar, especially if n is small. This motivates the introduction of chromas, which are also referred to as pitch classes. The chroma, which means 'color' in greek, is the collection of pitches which are an integer number of octaves apart, meaning that n in (5) fulfills

$$n = 2^{-m}, m \in \mathbb{Z} \tag{6}$$

with $m \in \mathbb{N}$ denoting the octave, which implies that $n \in \mathbb{Q}$. The fundamental frequency may thus be modeled in terms of its chroma, $\tilde{c}$, and its octave, m, as (see also, e.g., [1])

$$f = f_b \, 2^{\tilde{c}+m} \tag{7}$$

where $\tilde{c} \in [0, 1)$ and f_b denote the chroma class and a base (tuning) frequency, respectively. Using this formulation, the parametric pitch model presented in [17] may be extended into a parametric chroma model. Thus, the frequency peaks in a complex-valued[1] noise-free musical tone may be modeled as

$$x(t) = \sum_{\ell=1}^{L} a_\ell(t) e^{i2\pi f_b 2^{\tilde{c}+m} \ell t} \tag{8}$$

for a time-frame $t = 1, ..., N$, where $a_\ell(t)$ denotes the complex-valued amplitude of the ℓ:th harmonic, which may be either constant over the time-frame, or may

[1] In order to simplify notation, we here examine the discrete-time analytic signal version (see, e.g., [3, 18]) of the measured audio signal.

vary slowly. Here, $\tilde{c}$, m, and L denote the chroma, octave, and the number of sinusoids of the tone, respectively. It may be noted that the data is thus modeled in the time domain, as this is shown to render more efficient estimates than using the magnitude STFT [3]. In most Western music, there are twelve chroma classes, defined as the twelve semitones

$$C, C\#, D, D\#, E, F, F\#, G, G\#, A, A\#, \text{ and } B \tag{9}$$

and the concatenation of a chroma with its octave number, e.g., A4, denotes a musical tone. Here, two adjacent semitones are relatively spaced by $2^{1/12}$. Thus, the chroma parameter $\tilde{c}$ is discretized into twelve values, uniformly spaced on $[0, 1)$, i.e.,

$$\tilde{c} \in \left\{0, \frac{1}{12}, \frac{2}{12} \ldots, \frac{11}{12}\right\} \tag{10}$$

The tuning parameter f_b often varies somewhat amongst musicians, but a common standard sets 'A4' to 440 Hz [19]. This corresponds to $\tilde{c} = 9/12$, and $m = 4$, yielding the (normalized) tuning frequency

$$f_b = \frac{440}{f_s} 2^{-(9/12+4)} \tag{11}$$

where f_s denotes the sampling frequency. Our auditory system does not only perceive tones with these chroma as being distinctly different from each other, but also as equally spaced, which gives credit to the idea that our hearing is log-tempered. Furthermore, coinciding harmonics are not restricted to pitches within the same chroma, as pitches in different chromas may yield coinciding harmonics. For instance, for $n = 3/2 \approx 2^{7/12}$, the two pitches in our example will have many coinciding partials; two such tones are referred to as fifths. Fifths are thus spaced by approximately seven semitones and are commonly used together in musical compositions, as the overlapping spectral content is often deemed perceptually pleasant. Thus, if assuming that a polyphonic audio signal consists of K superimposed musical tones, the signal may be well modeled as

$$y(t) = x(t) + e(t) \tag{12}$$

where

$$x(t) = \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} a_{k,\ell}(t) e^{i2\pi f_b 2^{c_k + m_k} \ell t} \tag{13}$$

with the subscript k denoting the parameter of the k:th tone, and where e(t) is some form of additive noise. As (13) only models the sinusoidal part of the signal $y(t)$, any other features, such as, e.g., the timbre, will, without any loss of generality, be modeled as a part of the noise. In this work, the amplitude is allowed to be either constant, i.e., $a_{k,\ell}(t) = a_{k,l}, \forall(k,\ell)$, or slowly varying within each considered time-frame of N samples. Reminiscent to the approach in [15], we model the amplitude's time-varying nature using a spline basis with uniformly spaced knots (see, e.g., [20, p. 151]), i.e., such that the amplitudes in the time-frame follow a superposition of R B-spline bases,

$$a_{k,\ell}(t) = \sum_{r=1}^{R} \gamma_r(t) s_{k,\ell,r} \tag{14}$$

where the r:th spline base is weighted by an unknown complex amplitude, $s_{k,\ell,r}$.

3 Sparse chroma modeling and estimation

One way of estimating the unknown parameters in (13) may be to form the estimate as the one minimizing the (possibly weighted) squared estimation residuals, e.g., by using the non-linear least squares (NLS) algorithm. However, such an estimate requires precise knowledge about the model orders, something which generally is unknown. Such model orders are typically difficult to estimate for multi-pitch signals, as both the number of pitches and the number of harmonics in each pitch must then be determined. Furthermore, even if the true model orders are known, the NLS estimate will still require solving a multidimensional minimization over a typically multimodal cost function, thus necessitating an accurate search initialization [21]. On the other hand, if one tries to estimate the tonal content using, for instance, a periodogram-based approach, where the spectral peaks are estimated without taking the chroma structure into account, and thereafter grouping together the resulting estimates, this yields an involved combinatorial problem, as a number of frequency components typically belong in several tones, due to harmony. Instead, in this work, we construct an estimator based on the assumption that any given frequency component will be part of an ordered group of harmonic frequencies, i.e., a pitch. To achieve this, we propose to use a sparse modeling approach, reminiscent of the one presented in [5], where the non-linear model in (13) is replaced by a linear approximation of it, consisting of a highly overdetermined linear system, where the number of non-zero parameters

in the sought solution should be few, i.e., the solution should be sparse. Thereby, one may take the spectral structure of musical tones into account, while circumventing the need for explicitly estimating the model orders. Thus, consider the linear approximation

$$x(t) \approx \tilde{x}(t) = \sum_{c=0}^{11} \sum_{m=M_{\min}}^{M_{\max}} \sum_{\ell=1}^{L_{\max}} a_{c,m,\ell} e^{i2\pi f_b \ell t 2^{(c/12+m)}} \tag{15}$$

where $\tilde{x}(t)$ denotes the signal model representing the chromas in the Western musicological system, as described in (9)-(10). By denoting the twelve semitones using $c = 12\tilde{c}$, ordered as in (9), (15) includes all candidate tones within a range of octaves, from $M_{\min}$ to $M_{\max}$. Furthermore, $L_{\max}$ denotes the maximal number of harmonics considered, and $a_{c,m,\ell}$ the (complex-valued) amplitude for the ℓ:th harmonic in the m:th octave of pitch class c. From this approximation, it is clear that the spectral content is discretized into $Q = 12(M_{\max} - M_{\min})L_{\max}$ feasible frequencies, grouped into pitches of the same chroma. Also, as noted above, many of the harmonics between tones typically coincide, and it is therefore insufficient to simply map individual frequencies to a chroma, as they will likely map to several other chromas as well. To illustrate the sought sparsity structure of the solution, let

$$\mathbf{\Psi} = \left\{ \{a_{c,m,1}, \ldots, a_{c,m,L_{\max}}\}_{m=M_{\min},\ldots,M_{\max}} \right\}_{c=0,\ldots,11} \tag{16}$$

be the set of linear amplitude parameters for all possible frequencies in the overcomplete model. As the set $\mathbf{\Psi}$ is much larger than the actual solution set, most amplitudes, $a_{c,m,\ell}$, in (16) should be equal to zero, i.e., $\mathbf{\Psi}$ should be sparse. If, for instance, only the key C#5 is played, then all amplitudes, except $a_{1,5,\ell}$, for those ℓ present in this tone, should be zero. To measure the fit of the selected and estimated non-zero parameters, one may examine the minimum of the squared model residuals, by solving

$$\underset{\mathbf{\Psi}}{\text{minimize}} \sum_{t=1}^{N} \left| y(t) - \tilde{x}_{\mathbf{\Psi}}(t) \right|^2 \tag{17}$$

However, such a minimization will not promote the sought sparsity structure, and we therefore impose constraints to ensure a more desirable sparsity structure. In principle, we will do so by adding penalties to (17), reminiscent to the ones

used in [22–24], which add cost to non-desirable solutions that violate the sought sparsity pattern. The use of these will be somewhat different depending on if the amplitudes are allowed to vary or not; in the next two sections, we will deal with the two approaches separately.

3.1 Promoting sparsity when the amplitudes are constant

We proceed by first detailing the proposed chroma estimation procedure for the case without amplitude modulation. To simplify the exposition, consider the signal model in (15) for the entire time-frame expression on vector form as

$$\mathbf{y} = \left[\begin{array}{ccc} y(1) & \ldots & y(N) \end{array} \right]^T \tag{18}$$

$$= \sum_{c=0}^{11} \mathbf{W}_c \, \mathbf{a}_c + \mathbf{e} \triangleq \mathbf{W}\mathbf{a} + \mathbf{e} \tag{19}$$

where $(\cdot)^T$ denotes the transpose, and where

$$\mathbf{W} = \left[\begin{array}{ccc} \mathbf{W}_0 & \ldots & \mathbf{W}_{11} \end{array} \right]^T \tag{20}$$

$$\mathbf{W}_c = \left[\begin{array}{ccc} \mathbf{W}_{c,M_{\min}} & \ldots & \mathbf{W}_{c,M_{\max}} \end{array} \right]^T \tag{21}$$

$$\mathbf{W}_{c,m} = \left[\begin{array}{ccc} \mathbf{w}_{c,m}^1 & \ldots & \mathbf{w}_{c,m}^{L_{\max}} \end{array} \right]^T \tag{22}$$

$$\mathbf{w}_{c,m} = \left[\begin{array}{ccc} e^{i2\pi 2^{(c/12+m)}} & \ldots & e^{i2\pi N 2^{(c/12+m)}} \end{array} \right]^T \tag{23}$$

denote the dictionary of candidate tones and their partials, respectively. Also, let

$$\mathbf{a} = \left[\begin{array}{ccc} \mathbf{a}_c^T & \ldots & \mathbf{a}_c^T \end{array} \right]^T \tag{24}$$

$$\mathbf{a}_c = \left[\begin{array}{ccc} \mathbf{a}_{c,M_{\min}}^T & \ldots & \mathbf{a}_{c,M_{\max}}^T \end{array} \right]^T \tag{25}$$

$$\mathbf{a}_{c,m} = \left[\begin{array}{ccc} a_{c,m,1} & \ldots & a_{c,m,L_{\max}} \end{array} \right]^T \tag{26}$$

denote the linear amplitude parameters, $\mathbf{\Psi}$, of the over-complete dictionary on vector form. Thus, the blocks-within-blocks dictionary, $\mathbf{W} \in \mathbb{C}^{N \times Q}$, consists of twelve blocks of candidate chroma, such that each chroma is a block of ($M_{\max} - M_{\min}$) octave equivalent pitches, where each of these, in turn, consists of a block of $L_{\max}$ Fourier vectors. Our proposed method obtains the sought sparsity structure

by minimizing

$$||\mathbf{y} - \mathbf{W}\mathbf{a}||_2^2 + \lambda_2||\mathbf{a}||_1 + \lambda_3 \sum_{c=0}^{11} ||\mathbf{a}_c||_2 + \lambda_4||\mathbf{F}\mathbf{a}||_1 \tag{27}$$

where λ_i, for $i = 2, 3, 4$, denotes the user-defined sparse regularizers which weigh the importance between the different terms in (27), and where $\mathbf{F} \in \mathbb{C}^{(Q-1)\times Q}$ denotes the first order difference matrix, having elements $\mathbf{F}_{i,i} = 1$ and $\mathbf{F}_{i,i+1} = -1$ for $i = 1, \ldots, Q-1$, and zeros elsewhere. The first term in (27) penalizes the distance between the model and the measured signal, whereas the second term governs the overall sparsity of the amplitudes, thus forcing small values of $\mathbf{a}$ to be zero, affecting all indices equally. The third term is a group sparsity penalty, promoting sparsity between chromas, thereby countering the contributions from other chromas with partially overlapping spectral content. The last term in (27) is a total variation penalty which will penalize non-zero amplitudes at wrong octaves within the chroma, so that they will be efficiently clustered.

3.2 Promoting sparsity while allowing for time-varying amplitudes

To also allow for time-varying amplitudes, one has to consider some additions as well as some alterations to the earlier described method. Firstly, to allow for amplitude modulation, one has to extend the original problem with an additional parameter dimension. Using (14), the amplitudes' time-varying nature may be expressed on vector form as

$$\mathbf{a}_{k,l} = \sum_{r=1}^{R} \boldsymbol{\gamma}_r s_{r,k,l} = \Gamma \mathbf{s}_{k,l} \tag{28}$$

so that the amplitude vector, $\mathbf{a}_{k,l}$, is a linear combination of the $\boldsymbol{\gamma}_r \in \mathbb{R}^{N\times 1}$, for $r = 1, \ldots, R$, spline basis vectors, and where $s_{r,k,l}$ denotes the corresponding complex amplitude at spline point r of the l:th harmonic for the k:th pitch, and with

$$\mathbf{a}_{k,l} = \begin{bmatrix} a_{k,l}(1) & a_{k,l}(2) & \cdots & a_{k,l}(N) \end{bmatrix}^T \tag{29}$$

$$\mathbf{s}_{k,l} = \begin{bmatrix} s_{1,k,l} & s_{2,k,l} & \cdots & s_{R,k,l} \end{bmatrix}^T \tag{30}$$

$$\Gamma = \begin{bmatrix} \gamma_1 & \gamma_2 & \cdots & \gamma_R \end{bmatrix} \tag{31}$$

Using this formulation, the signal model for the time dependent amplitude becomes

$$\mathbf{y} = \sum_{m=M_{\min}}^{M_{\max}} \sum_{c=0}^{11} \operatorname{diag}(\mathbf{\Gamma}\mathbf{S}_{c,m}\mathbf{W}_{c,m}^{T}), \tag{32}$$

where

$$\mathbf{S}_{c,m} = \begin{bmatrix} \mathbf{s}_{c,m,1} & \cdots & \mathbf{s}_{c,m,L_{max}} \end{bmatrix} \tag{33}$$

$$\mathbf{s}_{c,m,l} = \begin{bmatrix} s_{1,c,m,l} & \cdots & s_{R,c,m,l} \end{bmatrix}^{T} \tag{34}$$

As a result, the sought chroma features of the considered signal frame may be found as the parameters minimizing

$$\underset{S_{0,M_{\min}} \cdots S_{11,M_{\max}}}{\text{minimize}} \frac{1}{2} \left|\left| \mathbf{y} - \sum_{c=0}^{11} \sum_{m=M_{\min}}^{M_{\max}} \operatorname{diag}(\mathbf{\Gamma}\mathbf{S}_{c,m}\mathbf{W}_{c,m}^{T}) \right|\right|_2^2 \tag{35}$$

where $\mathbf{y}$ denotes the vector containing the measured signal. To promote a sparse solution, one may rewrite and extend (35) as

$$\underset{S_P}{\text{minimize}} \frac{1}{2} \left|\left| \mathbf{y} - \sum_{p=0}^{P} \operatorname{diag}(\mathbf{\Gamma}\mathbf{S}_{p}\mathbf{W}_{p}^{T}) \right|\right|_2^2 \tag{36}$$

$$+ \lambda_2 \sum_{p=0}^{P} \sum_{l=1}^{L_{max}} \left|\left|\mathbf{s}_{p,l}\right|\right|_2 + \lambda_3 \sum_{c=0}^{11} \left|\left|\tilde{\mathbf{S}}_{c}\right|\right|_F \tag{37}$$

where the reparametrization from c, m to p is $p = 12(m - M_{\min}) + c$, with P denoting the total number of chroma-octave pairs in the dictionary, and with

$$\tilde{\mathbf{S}}_{c} = \begin{bmatrix} \mathbf{S}_{c,M_{\min}} & \cdots & \mathbf{S}_{c,M_{\max}} \end{bmatrix} \tag{38}$$

The first term in (37) measures the distance between the signal model and the measured data, the second term in (37) has the effect of setting columns in $\mathbf{s}_{p,l}$ with small l_2-norm to zero, whereas the third term promotes the sparsity of the resulting chroma estimate.

4 Efficient implementations

The optimization problems in (27) and (37) are convex, and may thus be solved using one of the many freely available interior point methods, such as, e.g., SeDuMi [25] and SDPT3 [26]. However, these methods typically scale poorly with increasing data lengths or with increasing dictionary sizes. To allow for a more efficient implementation, we here propose an implementation based on the Alternating Direction Method of Multipliers (ADMM), splitting the optimization into two or more simpler optimizations, which are then solved iteratively. Depending on the complexity of these sub-problems, the ADMM in general reaches a good approximate solution very fast, while thereafter converging more slowly to a really accurate solution [27]. For sparse modeling, this becomes evident as the ADMM converges quickly to the correct set of non-zero variables, while convergence to the correct relative amplitudes requires some further iterations. For the constant amplitude case in (27), the generalized ADMM (for more than two functions) is used, reminiscent to the approach proposed in [28]; this case is detailed in the following.

4.1 Chroma estimation with constant amplitudes via ADMM

The ADMM considers convex optimization problem which can be expressed as the sum of two convex functions by separating the variable into two parts

$$\underset{\mathbf{z},\mathbf{u}}{\text{minimize}} \quad f(\mathbf{z}) + g(\mathbf{u}) \quad \text{subject to} \quad \mathbf{u} - \mathbf{G}\mathbf{z} = \mathbf{0} \tag{39}$$

whereafter the augmented Lagrangian, i.e.,

$$L_\rho(\mathbf{z}, \mathbf{u}, \mathbf{d}) = f(\mathbf{z}) + g(\mathbf{u}) + \frac{\rho}{2} ||\mathbf{G}\mathbf{z} - \mathbf{u} + \mathbf{d}||_2^2 \tag{40}$$

can be used to find a solution to the original problem by iteratively solving

$$\mathbf{z}(\ell+1) = \arg\min_{\mathbf{z}} L_\rho(\mathbf{z}, \mathbf{u}(\ell), \mathbf{d}(\ell)) \tag{41}$$

$$\mathbf{u}(\ell+1) = \arg\min_{\mathbf{u}} L_\rho(\mathbf{z}(\ell+1), \mathbf{u}, \mathbf{d}(\ell)) \tag{42}$$

$$\mathbf{d}(\ell+1) = \mathbf{G}\mathbf{z}(\ell+1) - \mathbf{u}(\ell+1) + \mathbf{d}(\ell) \tag{43}$$

To cast (27) in this framework we use the generalization idea proposed in [27] to extend the ADMM to problems with more than two convex function. This is

done by assuming that $f = 0$, and defining g as the sum of the functions in the original problem, i.e.,

$$\underset{\mathbf{u}}{\text{minimize}} \quad \sum_{i=1}^{3} g_i(\mathbf{H}_i\mathbf{u}) \tag{44}$$

with $\mathbf{H}_1 = \mathbf{W}$, $\mathbf{H}_2 = \mathbf{I}$, $\mathbf{H}_3 = \mathbf{F}$, and

$$g_1(\mathbf{W}\mathbf{u}) = ||\mathbf{y} - \mathbf{W}\mathbf{u}||_2^2 \tag{45}$$

$$g_2(\mathbf{u}) = \lambda_2||\mathbf{u}||_1 + \lambda_3 \sum_{c=0}^{11} ||\mathbf{u}_c||_2 \tag{46}$$

$$g_3(\mathbf{F}\mathbf{u}) = \lambda_4||\mathbf{F}\mathbf{u}||_1 \tag{47}$$

The augmented Lagrangian of (27) is

$$\begin{aligned} L(\mathbf{z}, \mathbf{u}, \mathbf{d}) = g_1(\mathbf{u}_1) + g_2(\mathbf{u}_2) + g_3(\mathbf{u}_3) + \frac{\mu}{2}||\mathbf{W}\mathbf{z} - \mathbf{u}_1 - \mathbf{d}_1||_2^2 \\ + \frac{\mu}{2}||\mathbf{z} - \mathbf{u}_2 - \mathbf{d}_2||_2^2 + \frac{\mu}{2}||\mathbf{F}\mathbf{z} - \mathbf{u}_3 - \mathbf{d}_3||_2^2 \end{aligned} \tag{48}$$

where

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_1^T & \mathbf{u}_2^T & \mathbf{u}_3^T \end{bmatrix}^T \tag{49}$$

$$\mathbf{d} = \begin{bmatrix} \mathbf{d}_1^T & \mathbf{d}_2^T & \mathbf{d}_3^T \end{bmatrix}^T \tag{50}$$

denote the additional variables used to rewrite the optimization problem, and the dual variables, respectively. Thus, for the ℓ:th iteration,

$$\mathbf{z}(\ell + 1) = \arg\min_{\mathbf{z}} L(\mathbf{z}, \mathbf{u}(\ell), \mathbf{d}(\ell)) \tag{51}$$

which has the solution

$$\mathbf{z}(\ell + 1) = \left(\mathbf{G}^H\mathbf{G}\right)^{-1} \mathbf{G}^H \left(\mathbf{u}(\ell) + \mathbf{d}(\ell)\right) \tag{52}$$

where

$$\mathbf{G} = \begin{bmatrix} \mathbf{W}^T & \mathbf{I} & \mathbf{F}^T \end{bmatrix}^T \tag{53}$$

For $\mathbf{u}_1$,

$$\mathbf{u}_1(\ell+1) = \arg\min_{\mathbf{u}_1} L(\mathbf{z}(\ell+1), \mathbf{u}_1, \mathbf{d}_1(\ell)) \tag{54}$$

which may be solved as

$$\mathbf{u}_1(\ell+1) = \frac{\mathbf{y} + \mu\left(\mathbf{W}\mathbf{z}(\ell+1) - \mathbf{d}_1(\ell)\right)}{1+\mu} \tag{55}$$

For the remaining variables,

$$\mathbf{u}_2(\ell+1) = \arg\min_{\mathbf{u}_2} L\left(\mathbf{z}(\ell+1), \mathbf{u}_2, \mathbf{d}_2(\ell)\right) \tag{56}$$

$$\mathbf{u}_3(\ell+1) = \arg\min_{\mathbf{u}_3} L(\mathbf{z}(\ell+1), \mathbf{u}_3, \mathbf{d}_3(\ell)) \tag{57}$$

which have the solutions (see, e.g., [29])

$$\mathbf{u}_2(\ell+1) = \mathbf{T}\left(\mathbf{t}\left(\mathbf{z}(\ell+1) - \mathbf{d}_2(\ell), \frac{\lambda_2}{\mu}\right), \frac{\lambda_3\sqrt{(M)}}{\mu\sqrt{(12)}}\right) \tag{58}$$

$$\mathbf{u}_3(\ell+1) = \mathbf{t}\left(\mathbf{F}\mathbf{z}(\ell+1) - \mathbf{d}_3(\ell), \frac{\lambda_3\sqrt{(M)}}{\mu\sqrt{(12)}}\right) \tag{59}$$

where the shrinkage mappings $\mathbf{T}(\cdot)$ and $\mathbf{t}(\cdot)$ are defined as

$$\mathbf{t}(\mathbf{x}, \varkappa) = \frac{\mathbf{x}_k}{|\mathbf{x}_k|}\max\left(|\mathbf{x}_k| - \varkappa, 0\right), \text{ for all elements in } \mathbf{x} \tag{60}$$

$$\mathbf{T}(\mathbf{x}, \varkappa) = \frac{\mathbf{x}}{||\mathbf{x}||_2}\max\left(||\mathbf{x}||_2 - \varkappa, 0\right) \tag{61}$$

The augmented dual variable is updated as

$$\mathbf{d}(\ell+1) = \mathbf{d}(\ell) - (\mathbf{G}\mathbf{z}(\ell+1) - \mathbf{u}(\ell+1)) \tag{62}$$

The final chroma estimate is then found as setting $\hat{\mathbf{a}} = \mathbf{z}(\ell_{\text{final}})$. The resulting estimator is termed Chroma Estimation using Block Sparsity (CEBS). A summary of CEBS is shown in Algorithm 1.

Algorithm 1 The proposed CEBS algorithm

1: Initiate $\mathbf{z} = \mathbf{z}(0), \mathbf{u} = \mathbf{u}(0), \mathbf{d} = \mathbf{d}(0)$, and $\ell = 0$
2: **repeat**
3: $\quad \mathbf{z}(\ell+1)$ is updated as (52)
4: $\quad \mathbf{u}_1(\ell+1)$ is updated as (55)
5: $\quad \mathbf{u}_2(\ell+1)$ is updated as (58)
6: $\quad \mathbf{u}_3(\ell+1)$ is updated as (59)
7: $\quad \mathbf{d}(\ell+1)$ is updated as (62)
8: **until** convergence

4.2 Chroma estimation with amplitude modulation via ADMM

After the addition of amplitude modulation to the signal model, the problem is still convex, and we make use, once again, of the ADMM formulation, reminiscent to the approach proposed in [27]. The derivation becomes some what different to that in the previous section, since the amplitude modulated chroma model is more intricate. Denoting $\mathbf{S} = \begin{bmatrix} \mathbf{S}_1 & \cdots & \mathbf{S}_P \end{bmatrix}$, (37) may be rewritten as

$$\underset{X,Z}{\text{minimize}} \quad f(\mathbf{X}) + g(\mathbf{Z}) \quad \text{subject to} \quad \mathbf{X} - \mathbf{Z} = 0 \tag{63}$$

where

$$\begin{aligned} f(\mathbf{X}) &= \frac{1}{2} \left|\left| \mathbf{y} - \sum_{p=1}^{P} \text{diag}(\mathbf{\Gamma} \mathbf{X}_p \mathbf{W}_p) \right|\right|_2^2 \\ g(\mathbf{Z}) &= \lambda \sum_{p=1}^{P} \sum_{l=1}^{L_{max}} \|\mathbf{z}_{p,l}\|_2 + \gamma \sum_{c=0}^{11} \|\mathbf{Z}_c\|_F \end{aligned} \tag{64}$$

with $\mathbf{X}$ and $\mathbf{Z}$ having the same structure as $\mathbf{S}$. It is worth noting that the ADMM separates the sought variable into two unknown variables, here denoted $\mathbf{X}$ and $\mathbf{Z}$, enabling the original problem to be decomposed into easier sub-problems. These are in turn solved iteratively until convergence. The augmented Lagrangian of (63) becomes

$$L_\rho(\mathbf{X}, \mathbf{Z}, \mathbf{D}) = f(\mathbf{X}) + g(\mathbf{Z}) + \frac{\rho}{2} ||\mathbf{X} - \mathbf{Z} + \mathbf{D}||_2^2 \tag{65}$$

where $\mathbf{D}$ represents the scaled dual variable (see also [27]), which allows (65) to be solved iteratively as

$$\mathbf{X}(\ell+1) = \arg\min_{\mathbf{X}} L_\rho(\mathbf{X}, \mathbf{Z}(\ell), \mathbf{D}(\ell)) \tag{66}$$

$$\mathbf{Z}(\ell+1) = \arg\min_{\mathbf{Z}} L_\rho(\mathbf{X}(\ell+1), \mathbf{Z}, \mathbf{D}(\ell)) \tag{67}$$

$$\mathbf{D}(\ell+1) = \mathbf{X}(\ell+1) - \mathbf{Z}(\ell+1) + \mathbf{D}(\ell) \tag{68}$$

at the ℓ:th iteration. To solve (66), one differentiates $f(\mathbf{X}) + \frac{\rho}{2}||\mathbf{X} - \mathbf{Z} + \mathbf{D}||_2^2$ with respect to $\mathbf{X}_p$ and sets the result equal to zero, which yields

$$-\sum_{n=1}^{N} y(n)\mathbf{\Gamma}(n,\cdot)^H \mathbf{W}_p(\cdot,n)^H + \frac{\rho}{2}(\mathbf{X}_p - \mathbf{Z}_p + \mathbf{D}_p)$$
$$+\sum_{u=1}^{P}\sum_{n=1}^{N} \mathbf{\Gamma}(n,\cdot)^H \mathbf{\Gamma}(n,\cdot)\mathbf{X}_u \mathbf{W}_u(\cdot,n)\mathbf{W}_p(\cdot,n)^H = 0$$

By stacking all columns in $\mathbf{X}$ on top of each other, this may be represented as

$$\sum_{n=1}^{N} \mathbf{a}(p,n)^H y(n) + \frac{\rho}{2}(\mathbf{z}_p - \mathbf{d}_p) = \sum_{n=1}^{N}\sum_{u=1}^{P} \mathbf{a}(p,n)^H \mathbf{a}(u,n)\mathbf{x}_u + \frac{\rho}{2}\mathbf{x}_p \tag{69}$$

where

$$\mathbf{a}(u,n) = \mathbf{W}_u(\cdot,n)^T \otimes \mathbf{\Gamma}(n,\cdot) \tag{70}$$

$$\mathbf{x}_u = \mathrm{vec}(\mathbf{X}_u) \tag{71}$$

$$\mathbf{z}_u = \mathrm{vec}(\mathbf{Z}_u) \tag{72}$$

$$\mathbf{d}_u = \mathrm{vec}(\mathbf{D}_u) \tag{73}$$

with $\otimes$ denoting the Kronecker product, and $\mathbf{W}_u(\cdot,n)$ and $\mathbf{\Gamma}(n,\cdot)$ denoting the n:th column in $\mathbf{W}_u$ and the n:th row $\mathbf{\Gamma}$, respectively. Let

$$\mathbf{A}(p,u) = \sum_{n=1}^{N} \mathbf{a}(p,n)^H \mathbf{a}(u,n) \tag{74}$$

$$\tilde{\mathbf{y}}(p) = \sum_{n=1}^{N} \mathbf{a}(p,n)^H y(n) \tag{75}$$

Algorithm 2 The proposed CEAMS algorithm

1: Initiate $\mathbf{X} = \mathbf{X}(0), \mathbf{Z} = \mathbf{Z}(0), \mathbf{D} = \mathbf{D}(0)$, and $\ell = 0$
2: **repeat**
3: $\quad \mathbf{X}(\ell+1) = (\mathbf{A}^H\mathbf{A} + \frac{\rho}{2}\mathbf{I})^{-1}\mathbf{A}^H\tilde{\mathbf{Y}}$
4: $\quad \mathbf{Z}(\ell+1) = \boldsymbol{\mathcal{T}}(\mathbf{T}(\mathbf{X}_p(\ell+1) + \mathbf{D}_p(\ell), \beta/\rho), \alpha/\rho), \forall p$
5: $\quad \mathbf{D}(\ell+1) = \mathbf{X}(\ell+1) - \mathbf{Z}(\ell+1) + \mathbf{D}(\ell)$
6: $\quad \ell \leftarrow \ell + 1$
7: **until** convergence

$$\tilde{\mathbf{Y}} = \begin{bmatrix} \tilde{\mathbf{y}}(1) & \cdots & \tilde{\mathbf{y}}(P) \end{bmatrix}^T \tag{76}$$

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}(1,1) & \cdots & \mathbf{A}(1,P) \\ \vdots & \ddots & \vdots \\ \mathbf{A}(P,1) & \cdots & \mathbf{A}(P,P) \end{pmatrix} \tag{77}$$

This yields the proposed algorithm, which is summarized in Algorithm 2, where $\mathbf{T}(\cdot)$ is defined as in (60), and $\boldsymbol{\mathcal{T}}(\cdot)$ is defined as

$$\boldsymbol{\mathcal{T}}(\mathbf{X}, \varkappa) = \frac{\mathbf{X}}{||\mathbf{X}||_F} \max\left(||\mathbf{X}||_F - \varkappa, 0\right) \tag{78}$$

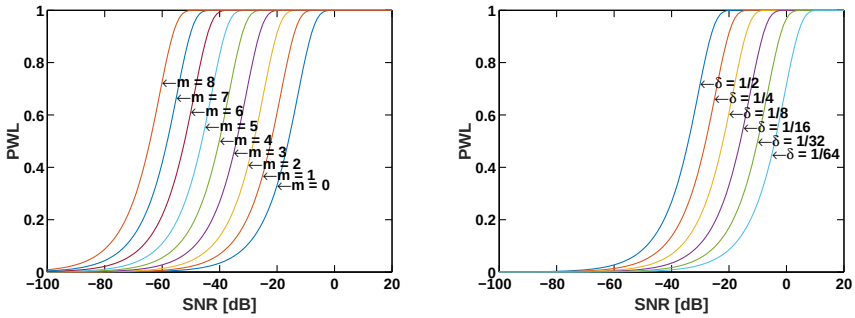
and is interpreted column wise, with $\boldsymbol{\mathcal{T}}(\cdot)$ operating over each part of $\mathbf{X}_p + \mathbf{D}_p$ that corresponds to $\tilde{\mathbf{S}}_{\tilde{c}}$. We term the resulting algorithm the Chroma Estimation of Amplitude Modulated Signals (CEAMS) method.

5 Numerical results

We proceed to examine the performance of the proposed estimators as a function of the Signal-to-Noise Ratio (SNR), measured in dB, defined as

$$\text{SNR} = 20\log_{10}\frac{\sigma_x}{\sigma_e} \tag{79}$$

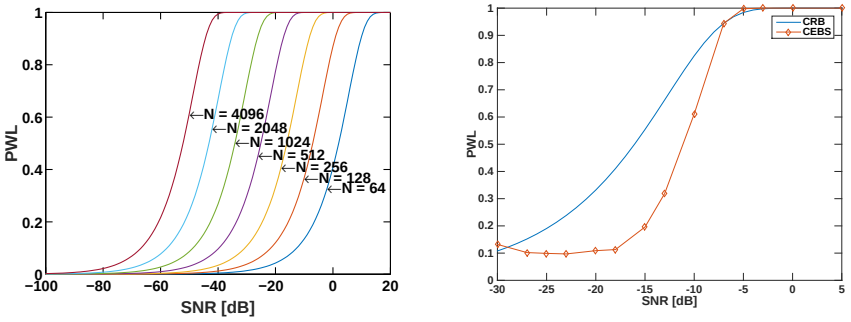
where σ_x and σ_e denote the power of the noise-free signal and the noise, respectively. As noted, the noise signal is here considered to consist of both the actual background noise and of any non-harmonic components in the recording. Therefore, in the case of strong formants, inharmonicity, or other musical features not



(a) (b)

Figure 1: Percentage of estimates within $c \pm 1/2$ from the true tone, when using twelve chromas, corresponding to the twelve semi-tones. Here, (a) is evaluated for the note C at different octaves, m, whereas (b) is evaluated for the note C3 when $c \in [0, 12)$ is discretized into $6/\delta$ points. For both, $N = 1024$ and $f_s = 20$ KHz (which equals a signal of approximately 51 ms).

modeled in this work, this signal might be quite strong. To examine the statistical limitations of chroma estimates, we initially examine the estimation limits, as obtained by the CRLB, which is derived in the appendix. As chroma is conventionally not considered a continuous variable, but rather as a number of grid points corresponding to some musicological system, we examine the achievable performance using the percentage-within-limits (PWL). This measures the number of estimates which are expected to fall within some pre-defined limit from the true value, i.e., $c \pm \delta$. For $\delta = 1/2$, this corresponds to the probability of obtaining estimates within the correct semi-tone, as $c = 0, \ldots, 11$. For $\delta = 1/4$, the PWL instead determines the likelihood of correctly estimating each quarter tone, and so forth. Figure 1(a) illustrates the performance of C notes at octaves $m = 0$ through $m = 8$, illustrating how the estimation problem becomes more difficult as the frequencies move closer to zero. The note is here formed from $N = 1024$ samples of a three-harmonic single pitch signal, measured at $f_s = 20$ KHz, which corresponds to a signal of approximately 51 ms. As can be seen from the figure, the PWL will reach 100% for the lowest note, i.e., being the most difficult estimation problem, at an SNR of approximately 0 dB. Figure 1(b) further illustrates the estimation limit for half tones up to the 64th tones, for a C3 tone, again reaching a perfect PWL at an SNR of approximately 0 dB, even for the



(a) (b)

Figure 2: Percentage of estimates within $c \pm 1/2$ from the true tone, when using twelve chromas. Here, (a) is evaluated for the note C3, for different data lengths, using $f_s = 20$ KHz, which implies a signal of N/f_s seconds, i.e., being (from the left) approximately 205 ms, 102 ms, 51 ms, 26 ms, 13 ms, 6 ms, and 3 ms. In (b), the estimated PWL for the CEBS estimator is compared to the CRLB for the C0 note, using $N = 1024$.

64th tone. Figure 2(a) similarly illustrates the estimation bounds as a function of the data length for the C3 note, using $\delta = 1/2$. All three figures thus indicate that one may expect a statistically efficient estimator to have no problems in correctly estimating the chromas, even in cases of SNR being significantly lower than expected for most audio recordings. However, due to the introduced penalties in the proposed estimators, one cannot expect these to be statistically efficient, even if the noise signal was a white sequence. This as the penalties will introduce an estimation bias, that although minor for most cases, will prevent the estimators to reach the CRLB. This is illustrated in Figure 2(b), showing the estimated PWL for the CEBS estimator, as obtained using 1000 Monte Carlo simulations, as compared to the corresponding CRLB. As may be seen in the figure, the actually achieved performance is, as expected, somewhat worse than predicted by the CRLB, although the latter gives a good indication of the achievable performance. Next, we proceed to examine the clarity of the proposed estimates, as compared to the (publicly available) estimators in [9, 10], using two audio signals from [30], namely a two channel FM-violin playing a middle C scale (all tones from C4 to C5), and a C-major chord, both in equal temperament, sampled at $f_s = 22$

KHz, mixed to a single channel using the method detailed in [10]. Figure 3 illustrate the resulting log-chromagrams for the Ellis, the Müller and Ewert, and the CEBS estimators. We have here divided the signal in segments of length $N = 1024$ samples (about 46 ms), having an overlap of 50%. For CEBS, we set $\lambda_2 = 0.05$, $\lambda_3 = 2.3$, and $\lambda_4 = 0.1$, for the chord, and $\lambda_2 = 0.05$, $\lambda_3 = 4$, and $\lambda_4 = 0.1$ for the scale, which are chosen using a simple heuristics from the FFT (see, e.g., [5]). The tuning frequency is here set to $f_{\text{base}} = 440$, and results remain quite unchanged at ± 3 Hz. As can be seen in the figures, the CEBS estimator yields a preferable estimate, suffering from noticeably less leakage and spurious estimates. Continuing, we examine the performance of the proposed estimators using a concert C-scale played by a trumpet acquired from [31], i.e., a highly non-stationary signal. Figure illustrates the resulting chromagrams, as obtained using the estimators in [10], [9], the CEBS estimator and the CEAMS estimator, respectively. For the CEAMS, we use $\lambda = 0.3$ and $\gamma = 193$, a window length of 1024 samples, a sampling frequency of 22050 Hz, $L_{max} = 9$ overtones, and 9 spline points. As is clear from the figure, both the estimators in [9, 10] suffer from apparent problems in choosing the correct chroma-bin for the scale. The CEBS estimate is notably cleaner, but still suffers from some spurious chroma features due to the inharmonicity of the signal. These spurious peaks have almost completely vanished in the CEAMS estimate. Here, we have used the same basic settings for CEBS as for CEAMS, and with $\lambda_2 = 0.05$, $\lambda_3 = 3$ and $\lambda_4 = 0.1$ (in setting these parameters, we have taken care to find the best possible setting for CEBS). It may be noted that the G in the scale is not detected by any method. This is because the fundamental frequency found in those time frames is 808 Hz, which is slightly closer to $G\#5$ than to $G5$, using concert tuning. To illustrate the difference in time-localization between CEBS and CEAMS, Figure show the 3-D chromagrams, where it once again can be noted that CEBS fails to identify the chroma-bin at $G\#$. Moreover, one may note the spurious peaks produced in CEBS, compared to the rest of the chromagram. This is in contrast to CEAMS, where none of the above mentioned behavior is present. Finally, we examine how well the proposed estimators capture the actual signal dynamics, by studying the envelopes of the reconstructed signals, formed from the respective estimates. Figure 6 illustrates how the amplitude modulation introduced in the CEAMS estimator has an advantage over the CEBS estimator. The CEAMS estimator captures both the shape and magnitude of the true signal envelope, whereas the CEBS estimator captures the shape reasonably, but fails to capture the amplitude.

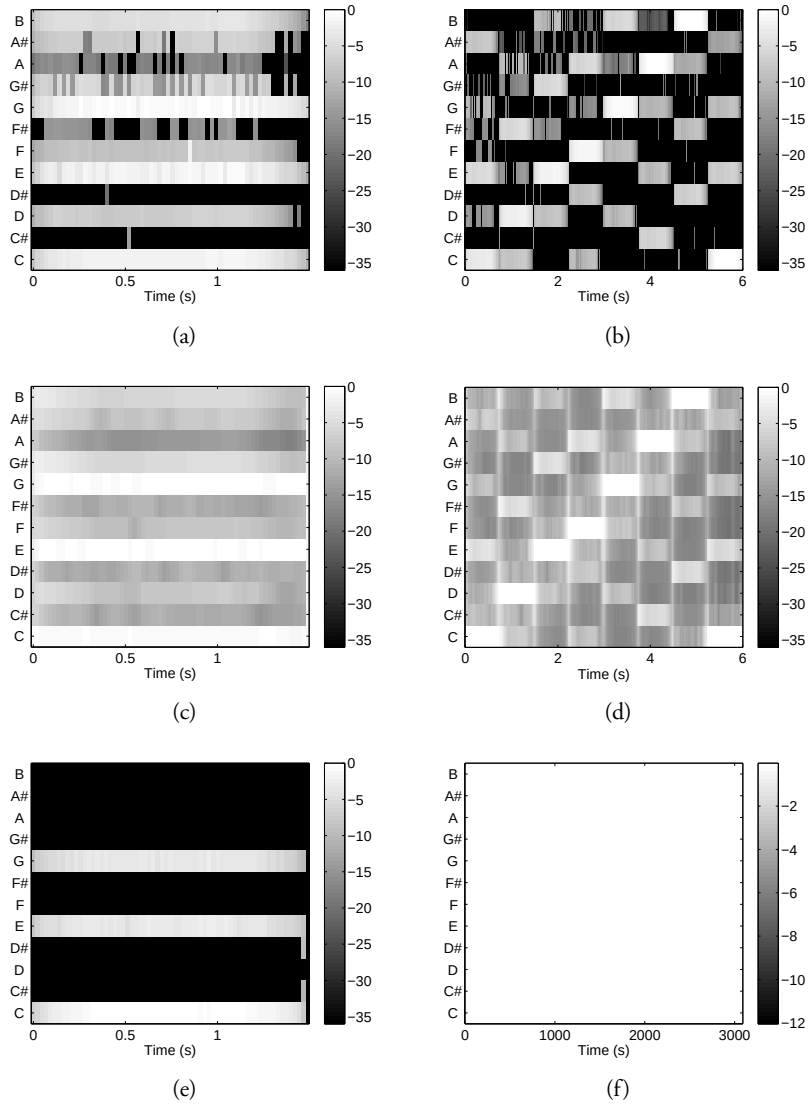


Figure 3: The performance for the (a,b) Ellis's method, (c,d) the Müller and Ewert method, and (e,f) the proposed CEBS algorithm, when evaluated on a C-chord (left), and a C-scale (right).

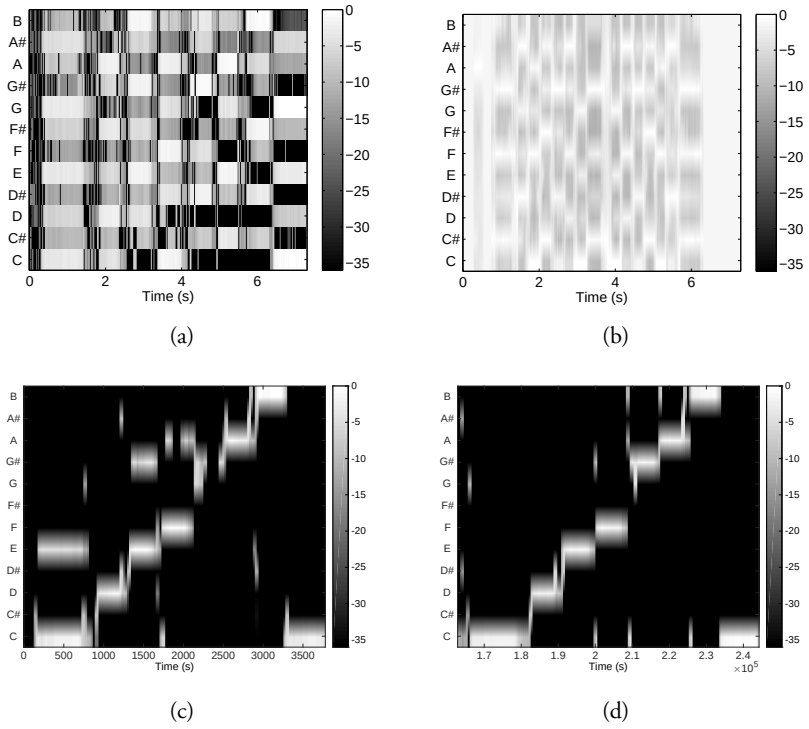


Figure 4: The figures above display the chromagrams for the trumpet scale, obtained using (a) Ellis's method, (b) the Müller and Ewert method, (c) CEBS, and (d) CEAMS.

6 Conclusions

In this article, we have presented two new methods for chroma estimation based on a sparse modeling reconstruction framework. The first method, CEBS, is designed to handle stationary time signals, and uses a fixed amplitude dictionary to model the measured signal. The method was further extended to also allow for time-varying signals, using a a spline-base model to capture the time-localization of the signal; the resulting estimator was termed the CEAMS method. The performance of the proposed estimators are compared both to the CRLB, presented herein for the problem at hand, as well as to two well-known chroma estimators using both real audio signals. It was found that the proposed estimators offer

a notable performance gain as compared to the comparable methods, with the CEAMS method being the better at capturing both the time-varying nature of the signal and the overall signal envelope.

7 Appendix: The Cramér-Rao lower bound

In this appendix, we present the Cramér-Rao Lower Bound (CRLB) for the chroma estimation problem. The signal in (15) may be equivalently be expressed as

$$x(t) = \sum_{k=1}^{K} \sum_{m=1}^{M_k} \sum_{l=1}^{L_k} a_{c_k,m,l} e^{j(2\pi f_b l t 2^m e^{(\ln(2)c_k/12)} + \varphi_{c_k,m,l})} \tag{80}$$

where M_k and L_k denote the highest octave and the highest harmonic for chroma class k, respectively. The the unknown parameters of the model are

$$\vartheta = [c_k, a_{c_k,1,1}, \varphi_{c_k,1,1} \cdots a_{c_k,m,l}, \varphi_{c_k,m,l}, c_{k+1}, a_{c_{k+1},1,1}, \varphi_{c_{k+1},1,1} \cdots] \tag{81}$$

The variance of the k:th parameter, ϑ_k, will thus be bound as

$$\text{var}(\vartheta_k) \geq [\mathbf{B}(\vartheta)]_{k,k} \tag{82}$$

where $\mathbf{B}(\vartheta)$ denotes the CRLB matrix. Let

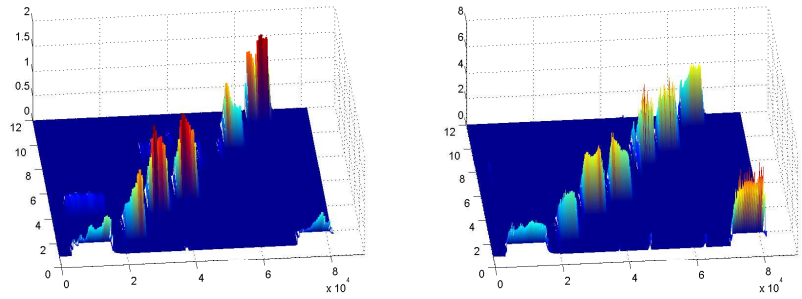
$$\hat{x}(\vartheta) = \begin{bmatrix} \hat{x}(0,\vartheta) & \cdots & \hat{x}(N-1,\vartheta) \end{bmatrix}^T \tag{83}$$

Assuming that the noise is independent of the parameters to be estimated, as well as having a Gaussian distribution with covariance matrix $\mathbf{Q}$, the Slepian-Bangs formula yields (see, e.g., [32, 33])

$$\mathbf{B}^{-1}(\vartheta) = 2\,\text{Re}\left\{ \frac{\partial \hat{\mathbf{x}}^H(\vartheta)}{\partial \vartheta} \mathbf{Q}^{-1} \frac{\partial \hat{\mathbf{x}}(\vartheta)}{\partial \vartheta^T} \right\} \tag{84}$$

Introduce

$$\nu_{c_k,m,l} = 2\pi f_b l t 2^m \frac{\ln(2)}{12} e^{\ln(2)c_k/12} a_{c_k,m,l} \tag{85}$$



(a) (b)

Figure 5: The chromagrams with time localization for the (a) CEBS and (b) CEAMS methods.

and form the partial derivatives with respect to the parameters as

$$\frac{\partial x(t,\vartheta)}{\partial \vartheta} = \begin{bmatrix} \sum_{m=1}^{M_k} \sum_{l=1}^{L_k} j\nu_{c_k,m,l} e^{j(2\pi f_b l t 2^m e^{(\ln(2)c_k/12)} + \varphi_{c_k,m,l})} \\ e^{j(2\pi f_b l t 2^m e^{(\ln(2)c_k/12)} + \varphi_{c_k,m,l})} \\ j a_{c_k,m,l} e^{j(2\pi f_b l t 2^m e^{(\ln(2)c_k/12)} + \varphi_{c_k,m,l})} \\ \vdots \end{bmatrix} \tag{86}$$

Making the further assumption that the noise is white, i.e., $\mathbf{Q} = \sigma^2 \mathbf{I}$, the CRLB matrix may be written as

$$\mathbf{B}^{-1}(\vartheta) = \frac{2}{\sigma^2} C \tag{87}$$

where $\mathbf{C}$ is defined as

$$\mathbf{C} = \mathrm{Re}\left\{ \frac{\partial \hat{\mathbf{x}}^H(\vartheta)}{\partial \vartheta} \frac{\partial \hat{\mathbf{x}}(\vartheta)}{\partial \vartheta^T} \right\} \tag{88}$$

Next, define

$$\chi_k = \left[\frac{\partial \hat{\mathbf{x}}(0,\vartheta)}{\partial c_k} \cdots \frac{\partial \hat{\mathbf{x}}(N-1,\vartheta)}{\partial c_k} \right]^T \tag{89}$$

$$\Psi_{c_k,m,l} = \begin{bmatrix} \frac{\partial \hat{\mathbf{x}}(0,\vartheta)}{\partial a_{c_k,m,l}} & \cdots & \frac{\partial \hat{\mathbf{x}}(N-1,\vartheta)}{\partial a_{c_k,m,l}} \\ \frac{\partial \hat{\mathbf{x}}(0,\vartheta)}{\partial \varphi_{c_k,m,l}} & \cdots & \frac{\partial \hat{\mathbf{x}}(0,\vartheta)}{\partial \varphi_{c_k,m,l}} \end{bmatrix} \tag{90}$$

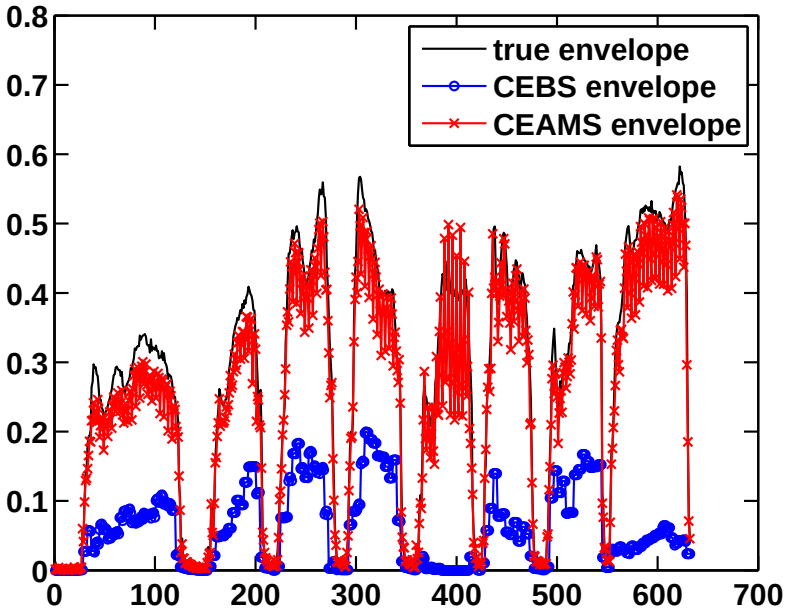


Figure 6: The figure above displays the time envelopes for the original signal (black) and the reconstructed signals.

Then, using $\sum_{p=1}^{P} = \sum_{m=1}^{M_k} \sum_{l=1}^{L_k}$,

$$\mathbf{C}_{1,1} = \begin{bmatrix} \chi_1^H \chi_1 & \chi_1^H \Psi_{1,1} & \chi_1^H \Psi_{1,2} & \cdots & \chi_1^H \Psi_{1,P} \\ \Psi_{1,1}^H \chi_1 & \Psi_{1,1}^H \Psi_{1,1} & \Psi_{1,1}^H \Psi_{1,2} & \cdots & \Psi_{1,1}^H \Psi_{1,P} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \Psi_{1,P}^H \chi_1 & \Psi_{1,P}^H \Psi_{1,1} & \Psi_{1,P}^H \Psi_{1,2} & \cdots & \Psi_{1,P}^H \Psi_{1,P} \end{bmatrix} \tag{91}$$

and, analogously,

$$\mathbf{C}_{2,1} = \begin{bmatrix} \chi_2^H \chi_1 & \chi_2^H \Psi_{1,1} & \chi_2^H \Psi_{1,2} & \cdots & \chi_2^H \Psi_{1,P} \\ \Psi_{2,1}^H \chi_1 & \Psi_{2,1}^H \Psi_{1,1} & \Psi_{2,1}^H \Psi_{1,2} & \cdots & \Psi_{2,1}^H \Psi_{1,P} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \Psi_{2,P}^H \chi_1 & \Psi_{2,P}^H \Psi_{2,1} & \Psi_{2,P}^H \Psi_{2,2} & \cdots & \Psi_{2,P}^H \Psi_{1,P} \end{bmatrix} \tag{92}$$

Thus,

$$\mathbf{C} = \mathrm{Re}\left\{ \begin{bmatrix} \mathbf{C}_{1,1} & \mathbf{C}_{1,2} & \mathbf{C}_{1,3} & \cdots & \mathbf{C}_{1,k} \\ \mathbf{C}_{2,1} & \mathbf{C}_{2,2} & \mathbf{C}_{2,3} & \cdots & \mathbf{C}_{2,k} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{C}_{k,1} & \mathbf{C}_{k,2} & \mathbf{C}_{k,3} & \cdots & \mathbf{C}_{k,k} \end{bmatrix} \right\} \tag{93}$$

with

$$\mathrm{Re}\{\chi_k^H \chi_k\} = \sum_{m=1}^{M_k} \sum_{l=1}^{L_k} \frac{a_{c_k,m,l}^2 (2\pi 2^m f_b l \frac{\ln(2)}{12} e^{\ln(2)c_k/12})^2}{6/\left(N(N+1)(2N+1)\right)}$$

$$\mathrm{Re}\{\Psi_{c_k,m,l}^H \Psi_{c_k,m,l}\} = \begin{bmatrix} N & 0 \\ 0 & N a_{c_k,m,l}^2 \end{bmatrix} \tag{94}$$

$$\mathrm{Re}\{\Psi_{c_k,m,l}, \chi_k\} = \begin{bmatrix} 0 \\ a_{c_k,m,l}^2 2\pi f_b l 2^m \frac{\ln(2)}{12} e^{\ln(2)c_k/12} \frac{N(N-1)}{2} \end{bmatrix} \tag{95}$$

$$\mathrm{Re}\{\Psi_{k,m,l}, \Psi_{k,m,r}\} = 0 \text{ for } \mathrm{l} \neq \mathrm{r} \tag{96}$$

If there is a spectral overlap between the chroma groups, and/or when the octaves considered have overlapping harmonics, the matrices $\mathbf{C}_{k,r}$, with k $\neq$ r will have non-zero entries. However, for the case considered herein, using 12 distinct chroma classes and only one tone, the following simplifications may be made:

$$\mathrm{Re}\{\chi_k \chi_r\} = 0 \text{ for } \mathrm{k} \neq \mathrm{r} \tag{97}$$

$$\mathrm{Re}\{\Psi_{k,p}, \Psi_{k,q}\} \approx 0 \tag{98}$$

$$\mathrm{Re}\{\Psi_k, \chi_r\} \approx 0, \tag{99}$$

implying that $\mathbf{C}$ will be a block-diagonal matrix, with all off diagonal blocks being zero, such that

$$\mathbf{C}^{-1} = \mathrm{Re}\left\{ \begin{bmatrix} \mathbf{C}_{1,1}^{-1} & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{C}_{2,2}^{-1} & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{C}_{k,k}^{-1} \end{bmatrix} \right\} \tag{100}$$

Partitioning the matrix $\mathbf{C}_{k,k}$ as

$$\mathbf{C}_{k,k} = \begin{bmatrix} c & \mathbf{d}^H \\ \mathbf{d} & \mathbf{E} \end{bmatrix} \tag{101}$$

where c is a constant, $\mathbf{d}$ is a vector, and $\mathbf{E}$ is a diagonal matrix, one may use the matrix inversion lemma to form the inverse matrix $[\mathbf{C}_{k,k}^{-1}]_{1,1}$ as

$$[\mathbf{C}_{k,k}^{-1}]_{1,1} = (c - \mathbf{d}^H \mathbf{E}^{-1} \mathbf{d})^{-1} \tag{102}$$

yielding the bound

$$\mathrm{var}(c_k) \geq \frac{6\sigma^2}{\sum_{m=1}^{M_\mathrm{k}} \sum_{l=1}^{L_\mathrm{k}} (a_{c_k,m,l} 2\pi f_b l 2^m \frac{\ln(2)}{12} e^{\ln(2) c_k / 12})^2 N(N-1)^2} \tag{103}$$

References

[1] M. Müller, D. P. W. Ellis, A. Klapuri, and G. Richard, "Signal Processing for Music Analysis," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1088–1110, 2011.

[2] R. Shepard, "Circularity in Judgements of Relative Pitch," *Journal of Acoustical Society of America*, vol. 36, no. 12, pp. 2346–2353, Dec. 1964.

[3] M. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, Morgan & Claypool, 2009.

[4] A. Klapuri and M. Davy, *Signal Processing Methods for Music Transcription*, Springer, 2006.

[5] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Multi-Pitch Estimation Exploiting Block Sparsity," *Elsevier Signal Processing*, vol. 109, pp. 236–247, April 2015.

[6] M. A. Bartsch and G. H. Wakefield, "Audio Thumbnailing of Popular Music Using Chroma-based Representations," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 96–104, Feb. 2005.

[7] S. Kim and S. Narayanan, "Dynamic Chroma Feature Vectors with Applications to Cover Song Identification.," in *10th IEEE Workshop on Multimedia Signal Processing*, 2008, pp. 984–987.

[8] T.-M. Chang, E.-T. Chen, C.-B. Hsieh, and P.-C. Chang, "Cover Song Identification with Direct Chroma Feature Extraction from AAC Files," in *IEEE 2nd Global Conference on Consumer Electronics*, Oct. 2013, pp. 55–56.

[9] D. P. W. Ellis, "Chroma Feature Analysis and Synthesis," `http://www.ee.columbia.edu/~dpwe/resources/matlab/chroma-ansyn/`, accessed Sept. 2014.

[10] M. Müller and S. Ewert, "Chroma Toolbox: MATLAB Implementations for Extracting Variants of Chroma-based Audio Features," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, 2011.

[11] J.S. Jacobson, *L1 Minimization for Sparse Audio Processing*, Ph.D. thesis, University of California, 2012.

[12] M. Mauch and S. Dixon, "Approximate Note Transcription for the Improved Identification of Difficult Chords," in *11th Int. Soc. Music Inf. Retrieval Conf.*, 2010, pp. 135–140.

[13] E. Gómez, *Tonal Description of Music Audio Signals*, Ph.D. thesis, Universitat Pompeu Fabra, 2006.

[14] M. Varewyck, J. Pauwels, and J.-P. Martens, "A Novel Chroma Representation of Polyphonic Music Based on Multiple Pitch Tracking Techniques," in *16th ACM International Conference on Multimedia*, New York, NY, USA, 2008, pp. 667–670, ACM.

[15] S. I. Adalbjörnsson, J. Swärd, T. Kronvall, and A. Jakobsson, "A Sparse Approach for Estimation of Amplitude Modulated Signals," in *48th Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, USA, Nov. 2-5 2014.

[16] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 11, no. 6, pp. 804–816, 2003.

[17] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Signal Processing*, vol. 88, no. 4, pp. 972–983, April 2008.

[18] S. L. Marple, "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2600–2603, September 1999.

[19] ISO, "Acoustics - Standard Tuning Frequency (Standard Musical Pitch)," Standard ISO 16:1975, International Organization for Standardization, Geneva, CH, 1975, ISO/TC 43, stage 90.93 (2011-12-22), ICS: 17.140.01.

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2 edition, 2009.

[21] P. Stoica, R. Moses, B. Friedlander, and T. Söderström, "Maximum Likelihood Estimation of the Parameters of Multiple Sinusoids from Noisy Measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 3, pp. 378–392, March 1989.

[22] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society B*, vol. 58, no. 1, pp. 267–288, 1996.

[23] M. Yuan and Y. Lin, "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[24] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society B*, vol. 67, no. 1, pp. 91–108, January 2005.

[25] J. F. Sturm, "Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11-12, pp. 625–653, August 1999.

[26] R. H. Tutuncu, K. C. Toh, and M. J. Todd, "Solving semidefinite-quadratic-linear programs using SDPT3," *Mathematical Programming Ser. B*, vol. 95, pp. 189–217, 2003.

[27] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[28] M. A. T. Figueiredo and J. M. Bioucas-Dias, "Algorithms for imaging inverse problems under sparsity regularization," in *Proc. 3rd Int. Workshop on Cognitive Information Processing*, May 2012, pp. 1–6.

[29] R. Chartrand and B. Wohlberg, "A Nonconvex ADMM Algorithm for Group Sparsity with Sparse Groups," in *38th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 26-31 2013.

[30] M. Romain, "Sound examples," https://ccrma.stanford.edu/~mromaine/220a/fp/sound- examples.html, accessed Sept. 2014.

[31] Mrs. Thomas, "Sound examples," http://www.hffmcsd.org/webpages/ arushkoski/nyssma.cfm, accessed Feb. 2015.

[32] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*, Prentice-Hall, Englewood Cliffs, N.J., 1993.

[33] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Prentice Hall, Upper Saddle River, N.J., 2005.