Model-based Analysis and Processing of Speech and Audio Signals

Christensen, Mads Græsbøll

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

MODEL-BASED ANALYSIS AND PROCESSING OF SPEECH AND AUDIO SIGNALS

BY
MADS GRÆSBØLLL CHRISTENSEN

DISSERTATION SUBMITTED 2020

AALBORG UNIVERSITY
DENMARK

Model-based Analysis and Processing of Speech and Audio Signals

Doctoral Dissertation

Mads Græsbølll Christensen

Dissertation submitted
September 15, 2020

Dissertation submitted: September, 2020

Assessment Committee: Professor Stefania Serafin (Chairman),
Aalborg University

Professor Patrick Naylor,
Imperial College

Professor Gaël Richard,
TELECOM Paris

Doctoral Thesis Series: Technical Faculty of IT and Design,
Aalborg University

ISSN: 1603-6204
ISBN: 978-87-85000-00-2

Printed in Denmark by Rosendahls, 2021

To my son, Erik

Biography

Mads Græsbøll Christensen received the M.Sc. and Ph.D. degrees in 2002 and 2005, respectively, from Aalborg University (AAU) in Denmark, where he is also currently employed at the Dept. of Architecture, Design & Media Technology as Full Professor in Audio Processing and is head and founder of the Audio Analysis Lab. He was formerly with the Dept. of Electronic Systems at AAU and has held visiting positions at Philips Research Labs, ENST, UCSB, and Columbia University. He has published 4 books and more than 250 papers in peer-reviewed conference proceedings and journals and has given tutorials at major international conferences such as ICASSP, EUSIPCO, and INTERSPEECH and a keynote talk at IWAENC. His research interests lie in audio and acoustic signal processing where he has worked on topics such as microphone arrays, noise reduction, signal modeling, speech analysis, audio classification, and audio coding. He has received several awards for his work, including best paper awards, the Spar Nord Foundation's Research Prize, a Danish Independent Research Council Young Researcher's Award, the Statoil Prize, and the EURASIP Early Career Award. He has received major grants from Independent Research Fund Denmark, the Villum Foundation, and Innovation Fund Denmark. Currently, he serves as Editor-in-Chief of EURASIP Journal on Audio, Speech, and Music Processing and as Senior Area Editor of IEEE Signal Processing Letters, and he has previously served several terms as Associate Editor of IEEE/ACM Trans. on Audio, Speech, and Language Processing and IEEE Signal Processing Letters. He is a Member of the IEEE Audio and Acoustic Signal Processing Technical Committee, and a founding Member and Vice-Chair of the EURASIP Technical Area Committee in Acoustic, Speech and Music Signal Processing. He is Senior Member of the IEEE, Member of EURASIP, and Fellow of the Danish Academy of Technical Sciences.

Abstract

This thesis is concerned with model-based analysis and processing of speech and audio signals, to which a number of scientific contributions are made in the form of new mathematical models and new methods for the processing of such signals. The thesis demonstrates how a number of models can be used for modeling speech and audio signals in different ways and for different purposes. It is shown how the problem of estimating the parameters of these models can be solved in a number of principled ways using methods such as maximum likelihood, subspace methods, and sparse approximations, whereby both very accurate and robust estimators that explicitly take the properties of speech and audio signals and the presence of noise into account are obtained. Among the parameter estimation problems considered are those of fundamental frequency estimation, linear prediction, source localization, and order selection, problems that have many important applications in speech and audio processing, including the analysis, coding, and enhancement of such signals. It is then shown how such models can be integrated in filtering methods to solve both signal enhancement and parameter estimation problems, such as noise statistics estimation and fundamental frequency estimation, and it is shown how these principles can be extended to multiple channels to solve the problems of beamforming and source localization. The results of the thesis as a whole demonstrate the benefits of the model-based approach compared to the typically non-parametric methods used in speech and audio processing, not only in terms of obtaining new and better methods but also advancing our understanding of both speech and audio signals and the associated estimation problems.

Resumé

Denne afhandling omhandler model-baseret analyse og processering af tale- og audiosignaler, hvortil en række vindenskabelige bridrag ydes i form af nye matematiske modeller og metoder til processering af sådanne signaler. Afhandling demonstrerer hvordan en række modeller kan bruges til at modellere tale- og audiosignaler i forskellige sammenhænge. Det vises hvordan problemet at estimere parametrene af disse modeller kan løses på en række matematisk funderede måder vha. af metoder såsom maksimum likelihood, underrumsmetoder, og sparse approksimationer. Herved opnås både meget præcise samt robuste estimatorer, der eksplicit tager højde for egenskaberne ved tale- og audiosignaler og tilstedeværelsen af støj. Blandt disse parameterestimeringsproblemer kan nævnes grundfrekvensestimering, lineær prædiktion, lokalisering samt orden-selektion, problemer der har mange vigtige anvendelser inden for tale- og audioprocessering, såsom signalanalyse, kodning og støjfjernelse. Dernæst vises det, hvordan modellerne kan indarbejdes i filtreringsmetoder til at løse både parameterestimerings- og støjfjernelsesproblemer, såsom estimering af støjstatistikker og grundfrekvensestimering, og det vises hvordan disse principper kan udvides til flere kanaler og anvendes til at løse beamforming- og lokaliserings-problemer. Resultaterne i afhandling som helhed viser fordelene ved den model-baserede tilgang sammenlignet med de i tale- og audio-processing typisk anvendte ikke-parametriske metoder, ikke kun mht. udviklingen af nye og bedre metoder men også at forbedre vores forståelse af både tale- og audiosignaler og de dertil hørende estimeringsproblemer.

List of publications

The main body of this thesis consists of the publications listed below. They can be downloaded using the links.

[A] **M. G. Christensen**, A. Jakobsson, and S. H. Jensen, *Sinusoidal Order Estimation using Angles between Subspaces*, EURASIP Journal on Advances in Signal Processing, Article ID 948756, pp. 1—11, 2009. Download: `https://tinyurl.com/yc5e3xhd`

[B] **M. G. Christensen** and A. Jakobsson, *Optimal Filter Designs for Separating and Enhancing Periodic Signals*, IEEE Transactions on Signal Processing, vol. 58, no. 12, pp. 5969–5983, 2010. Download: `https://tinyurl.com/yc865c4h`

[C] **M. G. Christensen**, *Accurate Estimation of Low Fundamental Frequencies*, IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 10, pp. 2042–2056, October 2013. Download: `https://tinyurl.com/ya4t7xpl`

[D] S. M. Nørholm, J. R. Jensen, and **M. G. Christensen**, *Instantaneous Pitch Estimation with Optimal Segmentation for Non-Stationary Voiced Speech*, IEEE/ACM Transactions on Audio, Speech, Language Processing, vol. 24, no. 12, pp. 2354—2367, December 2016. Download: `https://tinyurl.com/ycpfxn3z`

[E] S. M. Nørholm, J. R. Jensen, and **M. G. Christensen**, *Enhancement and Noise Statistics Estimation for Non-Stationary Voiced Speech*, IEEE/ACM Transacstions on Audio, Speech, and Language Processing, vol. 24, no. 4, pp. 645-–658, April 2016. Download: `https://tinyurl.com/y6veuu4d`

[F] D. Giacobello, **M. G. Christensen**, M. N. Murthi, S. H. Jensen, and M. Moonen, *Sparse Linear Prediction and Its Applications to Speech Processing*, IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 5, pp. 1644-–1657, July 2012. Download: `https://tinyurl.com/ycdle27r`

[G] J. R. Jensen, **M. G. Christensen**, J. Benesty and S. H. Jensen, *Joint Spatio-Temporal Filtering Methods for DOA and Fundamental Frequency Estimation*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 174-–185, January 2015. Download: `https://tinyurl.com/y9rkxzyl`

Contents

Preface

This thesis is submitted to the Technical Faculty of IT and Design at Aalborg University in fulfillment of the requirements for the degree doctor technices (dr.techn.). The thesis is written in the form of a collection of papers and a summary. None of the papers in this collection have appeared in a thesis written by me before. The summary presents my research as a whole and places the papers included in the thesis, which are then to be seen as example of my research, in a greater context. The selected papers are all journal papers, as these have been held to the highest standard of peer review. It should be noted that since it is tradition and common practice in signal processing to publish and present your work at conferences while also publishing a full version in a journal papers, the journal papers in this thesis were, when they were published, typically accompanied by a number of conferences papers on the same topic. The papers have been selected for a number of reasons. They are on different topics that represent my work broadly over the past 15 years, and they embody the spirit of my work. Finally, I feel a high degree of intellectual ownership of the contents of these particular papers. To complement the papers and complete the exposition, I will throughout the summary make references to other papers that are co-authored by me but are not included in the thesis.

The year 2020 marks 15 years since I defended my Ph.D. thesis and changed the direction of my research, as I started my post-doctoral life, and started down the path that I am still following today. In my Ph.D. studies, I was fortunate to be able to work on audio coding. It was the topic I would have picked, had I been in a position to freely select a topic, and a topic that had fascinated me since I first encountered mp3 when I was still in high school. Looking back, it may look like it was my destiny to work on audio coding in my Ph.D., but it was largely due to coincidences. Similarly, the topic that I would turn my attention to next was also because of a coincidence. In 2003, Andreas Jakobsson was giving a course on spectral analysis at Aalborg University that I was following and that is how I was introduced to estimation theory. It dawned on me that here was something very general that, if mastered, would enable one to study the mysteries of the universe, explore sub-atomic particles, make fortunes in finance, or, more importantly,

solve important problems in speech and audio processing! Quickly, I came to realize that here was a way of thinking that was a bit different from the typical speech and audio research, which was then, and still is, often dominated by heuristics, implicit assumptions, intuition, and perception, and that there was bound to be a big potential for making scientific progress, not only by deriving new solutions to engineering problem, but also by improving our understanding of these problems. Thus, I decided that my future was in estimation theory and its applications to speech and audio signals, and this thesis is the result of 15 years of research on this topic. Andreas Jakobsson would go on to become a big inspiration for me, and a personal friend, and we have worked together ever since. In short, my research has, in one way or another, revolved around answering the questions of how to model speech and audio signals, how to estimate their parameters under adverse conditions, and how to use these models for solving signal processing problems such as filtering, classification, and localization.

One may reasonably ask: why write a thesis such as this one? There is no requirement at Danish universities to write such a thesis, so surely writing such a thesis must be a sign of some kind of academic madness. There are several reasons for me writing this thesis, though. Firstly, it was for a long time my ambition to write such a thesis. In fact, the book entitled *Multi-Pitch Estimation* that I wrote more than ten years ago with Andreas Jakobsson after finishing my postdoc was intended as a starting point for such a thesis, but as time went on and I became preoccupied with other (more?) important things, it came to the point where I felt that the contents of that book had become a bit dated and I also, perhaps more importantly, felt that it did not fully demonstrate the potential behind the ideas that underlie my work. Secondly, I feel that I have an important contribution to make with this thesis, an important story to tell. Indeed, as conferences and journals within all branches of signal processing are being overwhelmed with one paper after another about how deep learning can be applied to solve various problems without really understanding neither the methods nor the problems, I believe that it is important that we ask ourselves what we mean by science and what standards our work should be held to. Is it enough to have good results? For me, it was never just about the results. It was also about the approach, the methodology, gaining a better of understanding both our tools and of the world around us. Indeed, I believe that the papers in this thesis demonstrate how a principled approach based on explicit signal models and assumptions can be used to solve difficult engineering problems. Often, critics have attacked our models and our assumptions saying that they are incorrect and that the state of the art does not make this or that assumption. However, even when our approach initially fails, it is due to the signal models and the assumptions that we know where to look and how to improve our models and methods. By contrast, it is difficult to improve on the non-parametric

methods that are commonplace in speech and audio processing where the assumptions are often implicit and not clear at all. In our work, we have seen this time and time again, and it is this point that I would like to make with this thesis.

I owe many people my sincere gratitude because the research reported is thesis would not have been possible without them. First, I would like to thank my mentor, Søren Holdt Jensen, who was a big influence on my early research career. Next, I would like to thank my long-time collaborator Andreas Jakobsson for all the research we have done together, for his idealistic nature and approach to research, and for being a generally positive influence on my life. Jacob Benesty is another person whom I would like to thank for our collaborations and for what he has done for me. I have the greatest admiration for his imagination, his never-ending search for new mathematical ideas an principles and for his uncompromising pursuit of excellence. I would also like to thank my various co-authors, particularly Marc Moonen and Manohar Murthi, along with my former Ph.D. students Daniele Giacobello and Sidsel Marie Nørholm who all contributed to this work. Special thanks go to Jesper Rindom Jensen and Jesper Kjær Nielsen, with whom I have worked closely now for many years and have made many important contributions to our field. Their combination of good spirits, hard work, and talent makes it a privilege to work with them! I would also like to thank some people who did not contribute directly to this work, yet have been important influences, namely Petra Stoica, Jingdong Chen, Richard Heusdens, Toon van Waterschoot, Jesper Boldt, Barry Quinn, and Max Little.

Mads Græsbøll Christensen
Aalborg, September 15, 2020

Summary

1 Introduction

Audio and speech processing, which is the topic of this thesis, has a rich history with many important contributions that influence our daily lives and define who we are and how we live. Modern inventions such as digital assistants, hearing aids, internet telephony, multimedia streaming, smart phones, etc., are all enabled by speech and audio processing. These inventions are enabled by progress in the development of methods for solving the underlying scientific problems of automatic speech recognition, speaker recognition, compression, noise reduction, dereverberation, beamforming, localization, source separation, etc. Speech and audio processing can, though, broadly be characterized as being concerned with one or more of three tasks: analysis (extracting information from signals), synthesis (generating signals), or transformation (changing signals). To perform these tasks it is important that we understand how signals are naturally generated by the source, e.g., as standing waves on a string, how they are propagated from a source to a receiver through the air and reflected off surfaces, and how they are perceived by a listener, imagined or real. The first and the second are studied in physics and acoustics while the latter is studied in human sound perception and cognition. Speech and audio processing thus draws upon knowledge from several scientific disciplines to understand the problems we seek to solve. Moreover, to actually analyse, synthesize or transform signals, we draw on mathematics for inspiration on how this can be achieved. More specifically, mathematical disciplines such as linear algebra, convex optimization, function analysis, statistics, and machine learning are the ones that we lean on to solve these problems. Speech and audio processing is thus a melting pot for a number of different scientific fields and disciplines, and time and time again important fundamental new insights and methods are discovered in this intersection between fields. Many of the unsolved problems of the field are those that are ill-posed and often also nonlinear, meaning that there either does not exist a solution to the problem in its most native form, and, if there does, we do not known how to find it! An example of such a problem

is source separation, which is the problem of finding the individual sources of a mixture. When only a single observation, which is the sum of a number of sources, is available, it is not, without further knowledge, possible to recover the individual sources. Indeed, such a problem has an infinite number of solutions even though only one or a few would be considered proper solutions. In such cases, we must impose constraints on the problem to obtain a meaningful solution. One way to do this, is to incorporate models into the problem definition. This could be models of the sources, how they are mixed, or how they propagate in space. Finding good models of speech and audio signals is, however, difficult. While it is possible derive models based on differential equations and boundary conditions in isolated, specific cases, such as for specific musical instruments, this is generally not possible for audio. The sources themselves are too diverse and too complex, and so are the ways in which they can be modified by the acoustic environment. They can degraded by background noise, changed by the reverberation of the surroundings, or distorted by nonlinear phenomena. Moreover, even if an exact model is known to hold, it may still be too complex to be of any practical use. Instead, models that are well-known to be approximations to nature are often used. Indeed, an approximate model may still prove good enough to solve the problem at hand. Given a model, the next question is how to estimate its parameters, and, depending on the nature of the parameters, this may be relatively simple or very difficult. Speech and audio signals are often recorded in adverse conditions, meaning that the signals of interest are degraded by various phenomena, such as background noise and reverberation. In that case, statistical principles that explicitly take the presence of these phenomena into account are preferable. Surprisingly, there are many examples within speech and audio of methods that do not explicitly take the presence of such degradations into account. Fortunately, there exists many mathematically well-founded principles for parameter estimation, some examples being maximum likelihood, Bayesian, and subspace methods. Once the parameters of the model have been found, a final question presents itself, namely how to use the models and their parameters. An example would be how to use a model of the signal of interest to perform noise reduction or source localization. These are the questions that my work has revolved around which also form the basis of this thesis. In summary, my work seeks to answer the following questions:

What are good models of speech and audio signals recorded in adverse conditions, how do we find their parameters, and how can we use them?

It must be stressed that there is a symbiotic relationship between these three questions concerning the models, the estimators, and their use. In answering the question above for different kinds of signals and for different

purposes, this thesis seeks to demonstrate the merits of the model-based approach to solving problems in speech and audio processing. Accordingly, the summary of the thesis is structured in the following way. First, signal modeling will be discussed and signal models of particular importance to this thesis will be presented. This is followed by an overview of parameter estimation methods for speech and audio signals, after which model-based filtering will be discussed. The last part serves to demonstrate how the model-based approach can be used in speech and audio processing to solve problems, in this case filtering, in a principled way and to demonstrate what can be achieved with signal models even if they are not always correct. Throughout all this, the contributions of this thesis and related work by the author will be highlighted. Finally, in the conclusion, a paper-wise summary of the contributions will be given and discussed before giving directions for future research.

2 Signal Modeling

In relation to signal models, such as those considered in this thesis, an interesting question is what exactly constitutes a good model. A good model, one could argue, is one that fits the observed data well. A model that accurately captures how the underlying physics that generated the signal is arguably also a good one. However, as George Box famously wrote, *all models are wrong, but some are useful* [1], and this applies very much to modeling of speech and audio signals. As previously mentioned, speech and audio signals are generally too complex to find exact models, e.g., by solving differential equations (something that can be done in some fields of engineering). What matters then is that the models capture the relevant aspects, which in turn depends on the application in question. The linear prediction model of speech is a good example of this [2]. It has clear physical meaning but it is well-known to be wrong, as it ignores the properties of voiced speech and the presence of the nasal cavity (see, e.g., [3]), yet it proves tremendously useful in many applications [2]. Another example is the models used in source separation which are frequently too poor to be used as fully generative models but accurate enough that a Wiener filter that can be used to estimate different sources from the observed signal. It is not enough that the model fits the data well, however. It is well-known that more complex models will generally fit noisy data better than simpler ones, even if the data was generated by a simple model. This is easy to explain using the polynomial model. A high-order polynomial can always fit the data better than a low-order polynomial model, since the low-order polynomial is contained in the higher-order polynomial. The parsimony principle, also known as Occam's razor, states that the simplest model that can be used to explain some phenomenon should be preferred over more complex ones. Although, this sounds like an appealing

idea, it is actually quite difficult to implement. In the context of signal modeling, the underlying problem is that of model selection or order selection in the case that we are choosing between subsets of the same model, which is a notoriously difficult problem to which many different approaches have been proposed over the years, e.g., [4–7]. It is also an extremely important problem, however. Sparsity is another way of incorporating or enforcing simplicity in a model. By imposing that a linear model only has a few non-zero coefficients, we obtain a simple model. Finally, for many purposes it is also desirable that that the model is physically meaningful, meaning that the parameters can be related to the physics that generated the signal. The absence of such meaning or interpretability is by many considered a huge problem for the models obtained in deep learning, although efforts have been made to address this also in speech and audio processing [8, 9].

While models are common in some fields of science and engineering, they are quite rare in speech and audio processing and most state-of-the-art methods are based on vague models. For example, most speech enhancement methods based on optimal filtering characterize both the signal of interest and the noise using second-order statistics, and many noise statistics estimators rely on proper initialization to form an initial model of the noise from which speech presence is then detected [10–12]. Similarly, beamforming and localization methods for microphone arrays frequently make use of propagation models but rarely make use of any more detailed knowledge about the signals than their second-order statistics [13]. There are some examples, of such models in speech and audio, though. Classical examples include sinusoidal models [14], which are sometimes used for problems such as time- and pitch-scale modification [15] and source coding [16], source-filter models, which formed the basis of early speech coders and have been used for speech analysis and transformation, and physical models of musical instruments are frequently used for synthesis. It might be argued that methods based on non-negative matrix factorization (NMF), which have found many uses in speech and audio processing (e.g., [17–20]), also fall into this category, and there might be some merit to this argument, but the vast majority of these methods impose rather vague assumptions and structure on the dictionaries and activation matrices that the so-obtained models lack physical meaning, much like the codebook-based methods based on vector quantization that have frequently been employed with much success in speech processing (see, e.g., [21]). Moreover, they are typically not parametric. The lack of fully generative, parametric signal models in speech and audio processing could be interpreted as a lack of faith in the correctness of such models. Indeed, the usage of parametric models is often met with some skepticism.

2.1 Basic Model and Assumptions

We will now introduce the basic signal models that are used throughout the thesis. The observed signal, to be analysed or processed, is denoted by $x(n)$ for $n = 0, 1, \ldots, N-1$ and consists of a signal of interest $s(n)$ in background noise $w(n)$. Define a vector containing M consecutive samples of the observed signal, termed a snapshot, as $\mathbf{x}(n) = [x(n)\ x(n+1)\ \cdots\ x(n+M-1)]^T$ where $\cdot^T$ denotes the transpose and similarly for $s(n)$ and $w(n)$. The models in this thesis can generally be described by the following equation:

$$\mathbf{x}(n) = \mathbf{s}(n) + \mathbf{w}(n) \tag{1}$$

$$= \mathbf{Z}(n)\mathbf{a} + \mathbf{w}(n), \tag{2}$$

where $\mathbf{Z}(n)$ is an $M \times L$ matrix whose columns are used to model the signal, $\mathbf{a}$ is a coefficient vector of length L, and $\mathbf{w}(n)$ is an additive noise vector of length M. Depending on the context, these quantities may all be real or complex in this thesis. Let $\mathrm{E}\{\cdot\}$ denote the expectation operator. Then, the covariance matrix of $\mathbf{x}(n)$ is defined as $\mathbf{R}_\mathbf{x} = \mathrm{E}\{\mathbf{x}(n)\mathbf{x}^H(n)\}$ (and similarly for other quantities) where $\cdot^H$ denotes the Hermitian transpose. With the definition of a snapshot $\mathbf{x}(n)$ above, it means that we can form a collection of $N - M + 1$ snapshots $\{\mathbf{x}(n)\}_{n=0}^{N-M+1}$ with $M \leq N$ from N observations of $x(n)$. From these snapshots, the covariance matrices for $\mathbf{x}(n)$ can be estimated using the sample covariance matrix estimate as $\widehat{\mathbf{R}}_\mathbf{x} = 1/(N-M+1)\sum_{n=M-1}^{N-1}\mathbf{x}(n)\mathbf{x}^H(n)$. When the time index n of the observed signal is not of interest, we shall use the following shorthand notation: $\mathbf{x} = \mathbf{x}(0)$, $\mathbf{Z} = \mathbf{Z}(0)$, and $\mathbf{w} = \mathbf{w}(0)$. For the case that $\mathbf{Z}(n)$ is given, the model is also known as the linear model while if its columns are given by some nonlinear unknown parameters, $\boldsymbol{\xi}$, the model is nonlinear. We refer to the model in (2) as being parametric, as it is determined via parameters, namely $\mathbf{a}$ and $\boldsymbol{\xi}$, which combined form the parameter vector $\boldsymbol{\theta}$. Moreover, when the model, or parts thereof, is further described in terms of statistical properties, such as the probability density function (pdf), it is also statistical. This would be the case, for example, if the noise is assumed to be zero-mean Gaussian[1] in which case it is characterized by the covariance matrix $\mathbf{R}_\mathbf{w}$, i.e., $\mathbf{w}(n) \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_\mathbf{w})$. If the term $\mathbf{Z}(n)\mathbf{a}$ is considered deterministic, the observed signal is distributed as $\mathbf{x}(n) \sim \mathcal{N}(\mathbf{Z}(n)\mathbf{a}, \mathbf{R}_\mathbf{w})$. Concerning the noise term $w(n)$ some further comments are in order, as it is frequently misunderstood. From the previous discussion, it should be clear that $\mathbf{Z}(n)\mathbf{a}$ represent what may be thought of as a deterministic part of the model in (2). Then $w(n)$ represents anything that is stochastic. Indeed, this would include observation noise, which is generally considered a nuisance to be dealt with, but it may also represent other

[1] The notation $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$ means that $\mathbf{x}$ is Gaussian distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{R}$.

stochastic parts of the signal, including some that may be of interest. For example, speech signals include both voiced speech and unvoiced speech. While voiced speech due to its periodic nature may, as we shall see, be modeled with $\mathbf{Z}(n)\mathbf{a}$, unvoiced speech cannot due to its stochastic nature. Then, $w(n)$ may represent an integral part of the speech signal, the unvoiced signal. Similar observations hold for many signals produced by musical instruments. The model in (2) is a fully generative in that given parameters $\mathbf{a}$ and $\boldsymbol{\xi}$, the signal of interest, $\mathbf{s}(n)$, can be generated.

2.2 Sinusoidal Models

We will now introduce and discuss some models of particular interest and relevance to this thesis. Indeed, these are the models that the papers in this thesis are based on. A general model which can be encountered throughout many fields of science and engineering, where it is perhaps the most commonly used regression model, is the polynomial model[2] where the matrix $\mathbf{Z}$ is given by:

$$\mathbf{Z}(n) = \begin{bmatrix} z_1^n & z_2^n & \cdots & z_L^n \\ z_1^{n+1} & z_2^{n+1} & \cdots & z_L^{n+1} \\ \vdots & \vdots & & \vdots \\ z_1^{n+M-1} & z_2^{n+M-1} & \cdots & z_L^{n+M-1} \end{bmatrix}. \tag{3}$$

For this model, the linear parameters, $\mathbf{a}$, are then the coefficients of the polynomial while the nonlinear parameters, $\boldsymbol{\xi}$, are $\{z_l\}_{l=1}^L$. Several of the models considered herein are special cases of this model. Indeed, a model that is used in the papers in this thesis, particularly in [A], is the sinusoidal model where $z_l = e^{j\omega_l}$, i.e., z_l is constrained to lie on the unit circle. In that case the matrix $\mathbf{Z}(n)$ is given by

$$\mathbf{Z}(n) = \begin{bmatrix} e^{j\omega_1 n} & e^{j\omega_2 n} & \cdots & e^{j\omega_L n} \\ e^{j\omega_1 (n+1)} & e^{j\omega_2 (n+1)} & \cdots & e^{j\omega_L (n+1)} \\ \vdots & \vdots & & \vdots \\ e^{j\omega_1 (n+M-1)} & e^{j\omega_2 (n+M-1)} & \cdots & e^{j\omega_L (n+M-1)} \end{bmatrix}. \tag{4}$$

It is now a matrix containing complex exponential functions having frequencies $\{\omega_l\}_{l=1}^L$, which then define $\boldsymbol{\xi}$, and the corresponding vector $\mathbf{a}$ comprising the linear parameters is given by

$$\mathbf{a} = \begin{bmatrix} A_1 e^{j\phi_1} & A_2 e^{j\phi_2} & \cdots A_L e^{j\phi_L} \end{bmatrix}, \tag{5}$$

which, in words, is formed from the real and positive amplitudes $\{A_l\}_{l=1}^L$ and phases $\{\phi_l\}_{l=1}^L$ with $0 \leq \phi_l < 2\pi$. When $\mathbf{x}(n)$ is real, the frequen-

[2]The term model polynomial model is perhaps more commonly associated with the special case $\mathbf{Z} = \mathbf{Z}(0)$, but for the sake of the discussion here, the more general form is preferred.

cies and complex amplitudes occur in complex-conjugate pairs. The sinusoidal model has a long history in speech and audio processing where it can be used in many problems and in a wide range of applications, including speech coding [16], speech analysis [22], audio coding [23–25], time- and pitch-scale modification [14], music synthesis and modification [26,27], interpolation/extrapolation [28–30], enhancement [31], and separation [32].

The model in (4) is more general than it may seem at first sight, as the Fourier transform matrix, $\mathbf{F} \in \mathbb{C}^{M \times M}$, can be seen to be a special case of this model. More specifically, if we define $\mathbf{F} = \mathbf{Z}$ with $L = M$ and $\omega_l = 2\pi\frac{l-1}{M}$ for $l = 1, \ldots, M-1$, then $\mathbf{F}^H\mathbf{x}(n)$ is the discrete Fourier transform of the snapshot $\mathbf{x}(n)$ and the entries in the vector $\mathbf{a}$ then form the complex spectrum of $\mathbf{x}(n)$ evaluated in frequencies $\omega_l = 2\pi\frac{l-1}{L}$ for $l = 1, \ldots, M-1$. This means that the model in (2) also can be used to represent short-time Fourier transform (STFT) based modeling and processing of signals (see also [33]).

A particularly important special case of this model that much of the work in this thesis has been devoted to is the harmonic model, which is a good model of signals that are periodic or approximately so. In this model, which is studied in papers [B], [C], and [G], the frequencies are all integral multiples of a fundamental frequency, ω_0, i.e., $\omega_l = \omega_0 l$, meaning that there is only one nonlinear parameter in $\boldsymbol{\xi}$ for this model. For a complex signal $\mathbf{x}(n) \in \mathbb{C}^M$, the matrix $\mathbf{Z}(n)$ is then given by

$$\mathbf{Z}(n) = \begin{bmatrix} e^{j\omega_0 n} & e^{j\omega_0 2n} & \ldots & e^{j\omega_0 Ln} \\ e^{j\omega_0 (n+1)} & e^{j\omega_0 2(n+1)} & \ldots & e^{j\omega_0 L(n+1)} \\ \vdots & \vdots & & \vdots \\ e^{j\omega_0 (n+M-1)} & e^{j\omega_0 2(n+M-1)} & \ldots & e^{j\omega_0 L(n+M-1)} \end{bmatrix}. \tag{6}$$

This thesis makes several contributions to modeling of speech and audio signals using the harmonic model in (6). In [C], it is demonstrated how the harmonic model can be used to derive a number of exact methods for estimating the fundamental frequency of periodic signals, where exact here means that no approximations (such as asymptotic approximations assuming infinite or large N) are used. This way, it is possible to estimate low fundamental frequencies in noisy signals, an otherwise difficult problem, and it is possible to take the real nature of speech and audio signals into account. This clearly demonstrates the merits of the model-based approach, as non-parametric methods have a difficult time dealing with these problems. In other work, it has been shown how this model can be extended to multiple channels in several ways. In [G] (and [34]) it is shown how it can be used in uniform linear arrays, in [35] how it can be used for multi-channel pitch estimation in a general setting, for example allowing for different SNRs in different channels, and in [36] how it can be used for estimation of fundamental frequencies in stereophonic mixtures by exploiting panning laws. This work was later ex-

tended to allow for separation and re-panning of individual sources [37]. Also, [38] shows how the harmonic model can be extended to the multi-pitch case where multiple periodic signals, termed sources, are present at the same time and how the corresponding covariance matrix model looks. The paper also demonstrates how difficult it actually is to model such signals, particularly due to model order, which can be different for different sources. Finally, [B], [C], and [G] demonstrate that the harmonic model is amenable to different kinds of optimal filtering that exploit the model and its properties explicitly.

2.3 Properties of Sinusoidal Models

We will now proceed to explore some interesting properties of the sinusoidal models. For the sinusoidal model (4) and the harmonic model (6), the model in (2) can be written in a number of ways that prove useful in different contexts:

$$\mathbf{x}(n) = \mathbf{Z}(n)\mathbf{a} + \mathbf{w}(n) \tag{7}$$

$$\mathbf{x}(n) = \mathbf{Z}\mathbf{D}(n)\mathbf{a} + \mathbf{w}(n) \tag{8}$$

$$\mathbf{x}(n) = \mathbf{Z}\mathbf{a}(n) + \mathbf{w}(n), \tag{9}$$

where $\mathbf{D}(n) = \text{diag}\left(\begin{bmatrix} e^{j\omega_1 n} & e^{j\omega_2 n} & \cdots & e^{j\omega_L n} \end{bmatrix}\right)$ is a diagonal matrix that accounts for the time-shift associated with n. Some interesting observations can be made from these ways of writing the model. As can be seen, the time-varying nature of the model, i.e., the dependency on the time index n can be taken into account in several ways. We can think of it as being part of the matrix $\mathbf{Z}(n)$. However, due to the structure of the model, we can also think of the complex amplitudes as being time-varying with $\mathbf{a}(n) = \mathbf{D}(n)\mathbf{a}$ or the dependency can be explained by $\mathbf{D}(n)$. As we shall see throughout this thesis, each of these have their own merits and applications. Often, it is useful to characterize the observed signal using its covariance matrix. For the model in (9), assuming that the signal of interest $s(n)$ and the noise $w(n)$ are uncorrelated, the covariance matrix can be shown to be [39]

$$\mathbf{R}_\mathbf{x} = \mathbf{Z}\text{E}\left\{\mathbf{a}(n)\mathbf{a}^H(n)\right\}\mathbf{Z}^H + \text{E}\left\{\mathbf{w}(n)\mathbf{w}^H(n)\right\} \tag{10}$$

$$= \mathbf{Z}\mathbf{P}\mathbf{Z}^H + \mathbf{R}_\mathbf{w}, \tag{11}$$

which is called the covariance matrix model. $\mathbf{Z}\mathbf{P}\mathbf{Z}^H$ is the covariance matrix of $\mathbf{s}(n)$, i.e., $\mathbf{R}_\mathbf{s} = \mathbf{Z}\mathbf{P}\mathbf{Z}^H$ and $\mathbf{R}_\mathbf{w}$ is the covariance matrix of the noise $w(n)$. Meanwhile, $\mathbf{P}$ is the covariance matrix of the amplitudes, which can be shown to be (under certain conditions, see [B])

$$\mathbf{P} = \text{diag}\left(\begin{bmatrix} A_1^2 & \cdots & A_L^2 \end{bmatrix}\right). \tag{12}$$

Concerning (11) and (12) some remarks are in order. This model was originally developed in the context of array signal processing where the amplitudes represent the signal of interest, while $\mathbf{Z}$ matrix represents a propagation model (see, e.g., [40,41]. This means that what is here considered a parametric model for the signal of interest, is actually not considered a parametric model in array signal processing. Moreover, this also raises some questions about the validity of the model, because $\mathbf{ZPZ}^H$ is derived based on the amplitudes being stochastic whereas these would be considered deterministic for models such as the harmonic model, and this is particularly so when seen as a sequence of snapshots $\mathbf{x}(n)$, $\mathbf{x}(n+1)$, ..., where the complex amplitudes, assuming that the observed signal is stationary, would be completely predictable. However, as shown in [B], the covariance matrix holds asymptotically also for deterministic signals.

A useful property of the sinusoidal and harmonic models in (4) and (6), respectively, is that the columns of $\mathbf{Z} \in \mathbb{C}^{M\times L}$ are asymptotically orthogonal, i.e., we have that

$$\lim_{M\to\infty} \frac{1}{M}\mathbf{Z}^H\mathbf{Z} = \mathbf{I}_L. \tag{13}$$

Another property is based on the observation that the covariance matrix of a time series, say $\mathbf{R}_\mathbf{x} \in \mathbb{R}^{M\times M}$, which exhibits a Toeplitz structure, is asymptotically equivalent to a circulant matrix [42]. This in turn means that such covariance matrices, for large M, can be diagonalized approximately by the Fourier transform matrix, $\mathbf{F} \in \mathbb{C}^{M\times M}$, i.e.,

$$\lim_{M\to\infty} \frac{1}{\sqrt{M}}\|\mathbf{R}_\mathbf{x} - \frac{1}{M}\mathbf{F}\mathbf{\Gamma}_\mathbf{x}\mathbf{F}^H\|_F = 0, \tag{14}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\mathbf{F} \in \mathbb{C}^{M\times M}$ is defined as in (4) with $\mathbf{F} = \mathbf{Z}$ for $L = M$ and $\omega_l = 2\pi\frac{l-1}{L}$ for $l = 1,\ldots,M-1$, and $\mathbf{\Gamma}_\mathbf{x} = \text{diag}\left([P_x(\omega_1)\ P_x(\omega_1)\ \cdots P_x(\omega_M)]\right)$ where $P_x(\omega)$ is the power spectral density[3] (psd) of $x(n)$.

From the approximation above, it follows that for the model in (4), where $\mathbf{Z}$ is generated from the frequencies $\{\omega_l\}_{l=1}^L$, we have that

$$\frac{1}{M}\mathbf{Z}^H\mathbf{R}_\mathbf{x}\mathbf{Z} \approx \text{diag}\left([P_x(\omega_1)\ P_x(\omega_2)\ \cdots P_x(\omega_M)]\right), \tag{15}$$

that is, the result is given by the psd of $x(n)$ evaluated in the frequencies $\{\omega_l\}_{l=1}^L$. The same arguments can be applied to $\mathbf{Z}^H\mathbf{R}_\mathbf{x}^{-1}\mathbf{Z}$ where the result is then given by the reciprocal of the psd of $x(n)$.

[3]The psd of $x(n)$ is here defined as $P_x(\omega) = \sum_{\tau=-\infty}^{\infty} r_{xx}(\tau)e^{-j\omega\tau}$ with $r_{xx}(\tau) = \text{E}\{x(n)x^*(n-\tau)\}$ which is consistent with our definition of $\mathbf{R}_\mathbf{x}$. For different definitions of psds and a discussion thereof, we refer the interested reader to [43].

2.4 Harmonic Chirp Model

A major criticism of the harmonic model, and the sinusoidal model, is that they are based on the assumption that the signal of interest is stationary for $n = 0, \ldots, N-1$. Speech and audio signals are inherently non-stationary. In speech, the pitch is a piece-wise smooth but continuously varying phenomenon while in music, the pitch may be more stationary, but glissando, vibrato, etc. render it non-stationary, as the frequency is continuously time-varying in these cases. In addition to this, the envelope of the signal of interest is also often time-varying. In fact, the envelope is an important part of the timbre of musical instruments [44]. To deal with this, speech and audio signal are analyzed over short segments. Speech is typically processed in segments of 20–30 ms while some types of music analysis uses much longer segments. While there is an abundance of models for analyzing and modeling non-stationary phenomena of both speech and audio signals in different contexts, often in the form of polynomial phase/amplitude models [45–48], AM-FM decompositions [49–51], or time-frequency analysis [52, 53], comparably few have focused on the problem in relation to fundamental frequency estimation [54, 55].

To investigate the need for taking such non-stationary phenomena into account explicitly and understanding its effects on fundamental frequency estimation, the harmonic chirp model was proposed in [56] which then led to [D] and [E]. Unlike the polynomial phase and amplitude models that have been proposed, or the chirp sinusoidal model [57], this model introduces only a minimum of new parameters, namely a linear change to the fundamental frequency within a segment. More specifically, the fundamental frequency is modeled as $\omega_0(n) = \omega_0 + \alpha_0 n$ where α_0 is the normalized fundamental chirp rate, which means that $\boldsymbol{\xi}$ is now comprised of ω_0 and α_0. This model leads to a (continuous) model of the instantaneous phase of the lth harmonic which is $\phi_l(n) = \frac{1}{2} l \alpha_0 n^2 + l\omega_0 n + \phi_l$. This leads to the following model for the matrix $\mathbf{Z} \in \mathbb{Z}^{M \times L}$ with $N = M$:

$$\mathbf{Z}^T = \begin{bmatrix} 1 & e^{j(\frac{1}{2}\alpha_0 + \omega_0)} & \ldots & e^{j(\frac{1}{2}\alpha_0 (N-1)^2 + \omega_0 (N-1))} \\ 1 & e^{j(\frac{1}{2} 2\alpha_0 + 2\omega_0)} & \ldots & e^{j(\frac{1}{2}\alpha_0 2(N-1)^2 + 2\omega_0 (N-1))} \\ \vdots & \vdots & & \vdots \\ 1 & e^{j(\frac{1}{2} L\alpha_0 + L\omega_0)} & \ldots & e^{j(\frac{1}{2}\alpha_0 L(N-1)^2 + L\omega_0 (N-1))} \end{bmatrix}. \tag{16}$$

It should be noted that unlike the previously considered models, this model does not lead to a simple decoupling of the time-dependency in (8), and in this sense the model is more complex and difficult to deal with. As a consequence, no simple covariance matrix model exists for the harmonic chirp model. This also nicely illustrates the point made earlier, namely that while complex models are good in the sense of describing the observations and

even the physics well, it may be difficult or even impossible to derive fast estimators and perform processing, such as filtering, on them. Papers [D] and [E] show how the harmonic chirp model takes non-stationary nature of voiced speech into account can be used for improved and robust speech analysis and for model-based noise reduction, respectively. In later work, a fast estimator for finding the parameters of the harmonic chirp model was proposed in [58]. The harmonic chirp model, and its close relative [57], are examples of models where we encounter a somewhat peculiar phenomenon, namely that the otherwise seemingly unimportant choice of time reference is important. As shown in [57], the choice of starting and end point actually affects the accuracy at which it is possible to estimate the model parameters. More specifically, the minimum error is obtain for a time indices symmetric around zero. Moreover, the choice of starting and end points also affects the correlation between the errors of the different parameters.

2.5 Linear Prediction

Another model that is ubiquitous in speech processing is that of linear prediction. In the linear prediction model, which in the statistical literature more commonly is referred to as the auto-regressive (AR) model, a sample of the observed signal is modeled as a linear combination of past samples plus an excitation which is often modeled as white Gaussian noise, i.e., $x(n) = \sum_{l=1}^{L} a_l x(n-l) + w(n)$. This model can also be cast in the form of (2) where $\mathbf{a} \in \mathbb{R}^L$ then contains the real linear prediction coefficients $\{a\}_{l=1}^{L}$ and the noise $w(n)$ is then the excitation. While often used as a parametric spectral estimator, the linear prediction model can also be interpreted as a generative model wherein the signal $w(n)$ excites an all-pole filter to produce the observation $x(n)$. The linear prediction model and related models been extensively used for speech coding, speech analysis, speech recognition, etc. It is interesting to note that despite its widespread use, there are many well-documented problems with the linear prediction model (see [2, 59]) and the most commonly used estimators for finding the coefficients, which are typically least squares methods or variants thereof [43]. In [F], sparse linear prediction is proposed with the aim of overcoming the problems associated with traditional linear prediction in connection with speech signals. In sparse linear prediction, the model is as follows:

$$\underbrace{\begin{bmatrix} x(N_1) \\ \vdots \\ x(N_2) \end{bmatrix}}_{\mathbf{x}\in\mathbb{R}^M} = \underbrace{\begin{bmatrix} x(N_1-1) & \cdots & x(N_1-L) \\ \vdots & & \vdots \\ x(N_2-1) & \cdots & x(N_2-L) \end{bmatrix}}_{\mathbf{Z}\in\mathbb{R}^{M\times L}} \underbrace{\begin{bmatrix} a_1 \\ \vdots \\ a_L \end{bmatrix}}_{\mathbf{a}\in\mathbb{R}^L} + \underbrace{\begin{bmatrix} w(N_1) \\ \vdots \\ w(N_2) \end{bmatrix}}_{\mathbf{w}\in\mathbb{R}^M}. \tag{17}$$

where N_1 and N_2 are the starting and end points with $M = N_2 - N_1 + 1$, respectively, which can then be chosen differently, whereby different methods

are obtained [43]. Typically, $x(n)$ is assumed to be zero outside an interval, e.g., for $n < 0$ or $n > N - 1$. Where $\mathbf{w}$ is assumed to be Gaussian distributed in traditional linear prediction, it is assumed to be sparse in sparse linear prediction, meaning that $\mathbf{w}$ only has a few large values whose magnitude is large while the rest are zero or small. A way of implementing this idea is via the assumption that $\mathbf{w}$ is Laplace distributed which leads to tractable estimation problems that can be solved using convex optimization. Furthermore, the notion that the predictor itself, i.e. the vector $\mathbf{a}$, can also be sparse was also proposed in [F]. This is the case, for example, for voiced speech where long-term correlations exist due to the pitch, and this correlation can be captured by a high-order but sparse predictor. These ideas lead to problems that can be solved with convex optimization at a much higher complexity than Levinson-Durbin algorithm which is used for solving thee traditional linear prediction problem (see, e.g., [43]). However, in follow-up work, notably [60,61], it has been shown how these problems can be solved efficiently with specialized algorithms. In later works, sparse Bayesian learning has also been investigated [62,63]. Another issue is that, unlike certain forms of traditional linear prediction, the stability of the corresponding auto-regressive filter is not guaranteed in sparse linear prediction, but, as shown in [64], this can be solved in a number of ways. It should be noted that since $\mathbf{w}$ here plays the role of excitation, it does not account for additive, background noise in this case. To do this, a more complicated model is needed, namely a state-space model.

The linear prediction model also has a clear physical and, in the case of speech, physiological interpretation. It can be interpreted as the lungs and vocal folds producing the signal $w(n)$ which is then filtered by the vocal tract. It is interesting to note that while early speech coders, such as LPC-10 [65], were very much thought of as model-based, later speech coders, such as CELP [65], are really just waveform approximating coders with little resemblance of a physical model (see, e.g., [66]). This point of view has been confirmed by recent developments, where after decades of little progress, deep learning has proven to lead to big improvements in coding gain in predictive coding at low bit-rates [67,68]. In these applications, the predictor is a essentially a nonlinear one, which means that there is no simple source-filter interpretation, only a mathematical, abstract, nonlinear mechanism for predicting speech samples from past samples.

3 Parameter Estimation

Having discussed some models of speech and audio signals, including also the models of interest to this thesis, the next question is how to estimate their parameters. Parameter estimation is, generally speaking, the art of estimating

a number of parameters from a number of observations, which are typically degraded in some way, for example by additive noise. Common parameter estimation problems in speech and audio processing include the estimation of pitch, relative transfer functions, direction-of-arrivals, frequencies, auto-regressive and moving average parameters, activation matrices, speech presence probability, noise statistics, etc. (see, e.g., [69]). This thesis is, however, primarily concerned with the models presented in Subsection 2.2, namely the harmonic model, the harmonic chirp model, and the linear prediction model. Concerning the sinusoidal model, the estimation of whose parameters is a classical problem in signal processing, there exists a vast body of work (see, e.g., [43, 70]), and we will refrain from any further discussion of this, except when relevant to the estimation problems considered in this thesis.

There exists many different methodologies for estimating parameters, including deterministic least squares, maximum likelihood, variational Bayes, sampling methods, sparse approximation methods, etc., some of which share commonalities or even coincide in special cases. While many of these have been explored in the context of speech and audio processing, e.g., [71–74], the predominant methodology continues to be least squares and variations thereof with some departures, in certain areas, into stochastic processes and minimum mean squared error. Very often, the underlying cost functions are implicit and estimators are essentially asserted and not derived from any fundamental principle or analyzed analytically. Example of this are widely used pitch estimators such as YIN [75], RAPT [76], PEFAC [55], SWIPE [77] which are all so-called non-parametric methods (in [78] and [79] it is shown how a number of these methods relate). Similar observations hold for the much-used localization methods GCC-PHAT and SRP-PHAT [80, 81], the good performance of which were later explained in [82] (see also [33]). As we shall see, when combining signal models and explicit assumptions with principled and statistical ways of solving problems, solutions are obtained whose properties can be analyzed and understood and whose weaknesses can be mitigated without resorting to guessing and trial-and-error. A notable exception to the above is the work on NMF which has led to important progress within speech and audio signals over the past two decades, particularly within source separation [83]. Indeed, that work has benefited from a statistical and principled way of solving estimation and modeling problems and has contributed significantly to the development of new methods for processing of speech and audio processing also outside of source separation. It is worth noting that in early work on NMF, the problem was seen as a deterministic matrix factorization problem, a way of looking at the problem that is highly problematic for speech and audio processing since spectrograms, to which these methods were applied, are not additive in a deterministic sense. However, this was later resolved by casting the factorization problems as statistical modeling and estimation problems [84].

3.1 Cramér-Rao Lower Bound

In terms of the signal model in (2), the associated parameter estimation problem can be defined formally as that of estimating the parameter vector $\boldsymbol{\theta}$, comprising both linear parameters $\mathbf{a}$ and nonlinear parameters $\boldsymbol{\xi}$, from the observations $x(n)$ for $n = 0, \ldots, N-1$, i.e., the observation vector $\mathbf{x}$ with $M = N$. Formally, an estimator can be defined as a function $f(\cdot) : \mathbb{R}^N \to \mathbb{R}^K$, which is typically achieved via a cost function which is minimized or maximized, and the argument for this is achieved is then the estimate. For example, given the cost function $J(\cdot)$, an estimate of $\boldsymbol{\theta}$ from $\mathbf{x} \in \mathbb{R}^N$ would then be $\hat{\boldsymbol{\theta}} = f(\mathbf{x}) = \arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$.

Returning to the model in (2), an interesting question is how accurately we can find its parameters, $\boldsymbol{\theta}$, and how the performance depends on the characteristics of the noise and even of the parameters characterizing the signal of interest. The Cramér-Rao lower bound (CRLB) (see, e.g., [43, 85]) is a very useful bound on the accuracy at which it is possible to estimate parameters from which much can often be learned about the problem at hand. There exists bounds that are able to predict more complex phenomena, such as the Barankin bound [86], which can predict threshold behavior, but those tend to be much more difficult to determine. The mean squared error (MSE) between an estimate $\hat{\theta}_i$ and the true value θ_i, which is the ith element of the vector $\boldsymbol{\theta} \in \mathbb{R}^K$, is given by

$$\begin{aligned} \text{MSE} &= \mathrm{E}\left\{ \left(\theta_i - \hat{\theta}_i\right)^2 \right\} && (18) \\ &= \left(\theta_i - \mathrm{E}\left\{\hat{\theta}_i\right\}\right)^2 + \mathrm{E}\left\{ \left(\hat{\theta}_i - \mathrm{E}\left\{\hat{\theta}_i\right\}\right)^2 \right\}. && (19) \end{aligned}$$

As can be seen, the MSE can be decomposed into a bias term and a variance term, where $\text{var}(\hat{\theta}_i) = \mathrm{E}\{\left(\hat{\theta}_i - \mathrm{E}\{\hat{\theta}_i\}\right)^2\}$ is then the variance. An estimate is unbiased if $\mathrm{E}\left\{\hat{\theta}_i\right\} = \theta_i \; \forall \theta_i$, and the difference, if such exists, is referred to as the bias. It should be noted that this condition applies to all possible values of the parameter, and it is thus a quite strong requirement. Then, it can be seen that for an unbiased estimator $\text{MSE} = \text{var}(\hat{\theta}_i)$ which is minimized by minimizing the variance. The CRLB for the parameter in question, which applies to all such unbiased estimators, is then given by

$$\text{var}(\hat{\theta}_i) \geq \left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{ii}, \tag{20}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher Information Matrix (FIM), which is defined as

$$\left[\mathbf{I}(\boldsymbol{\theta})\right]_{il} = -\mathrm{E}\left\{ \frac{\partial^2 \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_l} \right\}, \tag{21}$$

with $\ln p(\mathbf{x}; \boldsymbol{\theta})$ being the log-likelihood function for $\mathbf{x} \in \mathbb{R}^N$. The CRLB exists when the log-likelihood satisfies the regularity condition $\mathrm{E}\left\{ \frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\} = \mathbf{0}$

for all $\boldsymbol{\theta}$ (see [85]). An estimator that can be shown to attain this bound can then be claimed to be optimal, and such an estimator is said to be efficient. Interestingly, it can be shown that the CRLB for a parameter of a model can only get larger, not smaller, as more parameters are added to the model [87]. This is yet another argument for preferring simple models over more complex ones, particularly under adverse conditions with low SNRs. Consider the case of the model in (2) with Gaussian distributed noise with covariance matix $\mathbf{R}_{\mathbf{w}}$, the observed signal $\mathbf{x} \in \mathbb{R}^N$ is distributed as $\mathbf{x} \sim \mathcal{N}(\mathbf{Za}, \mathbf{R}_{\mathbf{w}})$. In that case, the likelihood function is given by

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^M \det(\mathbf{R}_{\mathbf{w}})}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{Za})^T \mathbf{R}_{\mathbf{w}}^{-1}(\mathbf{x}-\mathbf{Za})}, \tag{22}$$

where $\det(\cdot)$ denotes the determinant. The log-likelihood function is then given by

$$\ln p(\mathbf{x};\boldsymbol{\theta}) = -\frac{1}{2}\ln(2\pi)^M - \frac{1}{2}\ln\det(\mathbf{R}_{\mathbf{w}}) - \frac{1}{2}(\mathbf{x}-\mathbf{Za})^T \mathbf{R}_{\mathbf{w}}^{-1}(\mathbf{x}-\mathbf{Za}). \tag{23}$$

Then, it can be shown that the entries in the FIM are given by the following expression, which is known as Slepian-Bangs formula [88]:

$$[\mathbf{I}(\boldsymbol{\theta})]_{il} = \left[\frac{\partial \mathbf{Za}}{\partial \theta_i}\right]^T \mathbf{R}_{\mathbf{w}}^{-1} \left[\frac{\partial \mathbf{Za}}{\partial \theta_j}\right] + \frac{1}{2}\mathrm{trace}\left\{\mathbf{R}_{\mathbf{w}}^{-1}\frac{\partial \mathbf{R}_{\mathbf{w}}}{\partial \theta_i}\mathbf{R}_{\mathbf{w}}^{-1}\frac{\partial \mathbf{R}_{\mathbf{w}}}{\partial \theta_j}\right\}. \tag{24}$$

For the case that the noise covariance matrix, $\mathbf{R}_{\mathbf{w}}$ does not depend on any of the parameters in $\boldsymbol{\theta}$, which is the case for the way we defined the parameters of interest in connection with the model in (2), the second term is zero. It should be remarked, however, that while the assumption of known noise statistics is common, and practical, in some fields, it is generally not very useful in speech and audio due to the time-varying nature of such signals. For this reason, substantial research in speech and audio have over the past couple of decades been devoted to the problem of estimating noise statistics even in the presence of the signal of interest. The CRLB has been determined for many different signal models including the sinusoidal model, the harmonic model, and the auto-regressive model all based on the assumption of the noise being Gaussian (see, e.g., [43, 85]). Often, asymptotic approximations assuming a large number of samples, N, are exploited in deriving these bounds in which case they are only approximate. More accurate bounds can be determined numerically, but while such bounds are still useful as a benchmark in simulations, they tend to shed little or no light on how performance depends on various factors, and so it is difficult to analyze and understand the parameter estimation problem at hand.

3.2 Linear Least Squares Estimators

Returning to the problem of how to actually estimate parameters, the maximum likelihood estimator (MLE) is probably the most commonly used estimator in all of signal processing, and there are many good reasons for this. Firstly, its properties are well-understood [85] and, secondly, very often it proves to be tractable for estimation problems of interest and sometimes it even leads to simple estimators that have fast implementations. In MLE, the parameters that maximize the likelihood function, or rather the log-likelihood function, are chosen as the estimates, i.e.,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ln p(\mathbf{x}; \boldsymbol{\theta}). \tag{25}$$

The maximum likelihood estimator is asymptotically optimal (under some technical conditions [85]), meaning that for sufficiently large N, the estimates are unbiased and achieve the CRLB, and are distributed as

$$\boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\theta})). \tag{26}$$

For the case of the model in (2) with Gaussian distributed noise with covariance matix $\mathbf{R}_{\mathbf{w}}$, the observed signal $\mathbf{x} \in \mathbb{R}^N$ is, as before, distributed as $\mathbf{x} \sim \mathcal{N}(\mathbf{Za}, \mathbf{R}_{\mathbf{w}})$, i.e., the likelihood function is given by (22) and the log-likelihood is given by (23). When the noise covariance matrix, $\mathbf{R}_{\mathbf{w}}$, is known, and has full rank the maximum likelihood estimator can be seen to be

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} (\mathbf{x} - \mathbf{Za})^H \mathbf{R}_{\mathbf{w}}^{-1} (\mathbf{x} - \mathbf{Za}), \tag{27}$$

which is equivalent to weighted (nonlinear) least squares (WLS). While it may appear straightforward to use (27) to solve estimation problems, this is typically only the case for linear parameters, such as $\mathbf{a}$, while the nonlinear parameters that characterize the columns of $\mathbf{Z}$ are more difficult to estimate. For example, in the case of the sinusoidal model in (4), the estimation of the linear parameters reduces to linear (weighted) least squares, i.e.,

$$\hat{\mathbf{a}} = \left(\mathbf{Z}^H \mathbf{R}_{\mathbf{w}}^{-1} \mathbf{Z}\right)^{-1} \mathbf{Z}^H \mathbf{R}_{\mathbf{w}}^{-1} \mathbf{x}. \tag{28}$$

Concerning the specifics of estimating the complex amplitudes of the sinusoidal model in (4), we refer the interested reader to [89, 90]. The principles discussed therein also apply to the haromnic model in (6) (see [78]). Another problem that can be solved using (28) is that of estimating the coefficients of the linear prediction model in (17), or, equivalently the parameters of the auto-regressive model. In that model, the noise, $w(n)$, is called the excitation and is assumed to be white and Gaussian distributed, i.e., we have that $\mathbf{R}_{\mathbf{w}} = \sigma_w^2 \mathbf{I}_M$. In that case, (28) reduces to $\hat{\mathbf{a}} = \left(\mathbf{Z}^H \mathbf{Z}\right)^{-1} \mathbf{Z}^H \mathbf{x}$ and depending on the choice of N_1 and N_2 in (17) different methods from the literature

can be obtained [43]. For $N_1 = 1$ and $N_2 = N + L$ we obtain the autocorrelation method and for $N_1 = L + 1$ and $N_2 = N$ the covariance method is obtained. In the former case, the parameters can then be estimated using the Levinson-Durbin algorithm due to the Toeplitz structure of $\mathbf{Z}$. It should, though, be remarked that the so-obtained estimates are only conditional (or approximate) maximum likelihood estimates due to the memory effects of the recursive filters [91].

Interestingly, the optimal distribution in (26), and thus the CRLB, is actually useful for deriving estimators, including also estimators that can be claimed to be maximum likelihood estimators, for problems where a reparametrization of the problem is used, i.e., where the parameters of interest are expressed in terms of other parameters that are then estimated from the observed signal. Using the CRLB, the asymptotic variance of MLE estimates, and possibly also correlations between different parameter errors, are known and can the be weighted appropriately to obtain an optimal estimate [92]. As an example of how this can be used, consider the following. Given unconstrained parameter estimates $\boldsymbol{\theta} \in \mathbb{R}^K$ obtained using a MLE, we wish to map these estimates to a subset $\mathcal{C} \subset \mathbb{R}^K$. Then, the estimation errors are distributed according to (26), and the mapping can be performed as follows:

$$\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathcal{C}} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T \mathbf{I}(\boldsymbol{\theta}) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right). \tag{29}$$

A difficulty is then that the weighting in using this principle is that the FIM may depend on the true and unknown parameters in $\boldsymbol{\theta}$ but for sufficiently high N its estimate $\hat{\boldsymbol{\theta}}$ can be used instead based on continuity arguments [93]. In [94] it was demonstrated how this principle can be used for vector quantization based speech and audio processing in cases where parametrizations are commonly used, as in source separation and speech enhancement.

The principle can also be used for reparametrization of an estimation problem involving a linear transformation. As an example of this, consider the case where $\boldsymbol{\theta} = \mathbf{A}\boldsymbol{\theta}'$ with $\mathbf{A} \in \mathbb{R}^{K \times K'}$ and $\boldsymbol{\theta}' \in \mathbb{R}^{K'}$ for $K' \leq K$. Then, given $\boldsymbol{\theta} \in \mathbb{R}^K$, $\mathbf{I}(\boldsymbol{\theta})$, and $\mathbf{A}$ it is possible to solve for $\boldsymbol{\theta}'$. The principles involved in this are also known as the invariance principle and the extended invariance principles, depending on whether the transformation relating the parameters is invertible or not. In [95] it was shown how this principle can be used for fundamental frequency estimation, which was further explored in [C]. More specifically, define a vector containing the parameter estimates of the sinusoidal model as $\boldsymbol{\theta} = \begin{bmatrix} \omega_1 & A_1 & \phi_1 & \cdots & \omega_L & A_L & \phi_L \end{bmatrix}^T \in \mathbb{R}^{3L}$ and a vector containing the parameters of the harmonic model as $\boldsymbol{\theta}' = \begin{bmatrix} \omega_0 & A_1 & \phi_1 & \cdots & A_L & \phi_L \end{bmatrix}^T \in \mathbb{R}^{2L+1}$. Then, there exists a matrix $\mathbf{S} \in \mathbb{R}^{3L \times 2L+1}$ such that the two vectors can be related as $\boldsymbol{\theta} = \mathbf{S}\boldsymbol{\theta}'$ (see [C]). Then, $\boldsymbol{\theta}'$ can be estimated as $\hat{\boldsymbol{\theta}} = \left(\mathbf{S}^T \mathbf{I}(\boldsymbol{\theta}) \mathbf{S}\right)^{-1} \mathbf{S}^T \mathbf{I}(\boldsymbol{\theta}) \boldsymbol{\theta}$ provided that $\mathbf{S}$ has

full rank. This method is representative of a class of methods for fundamental frequency estimation, termed harmonic fitting, that employ similar two-stage procedures, only this is based on an optimal weighting of the estimates.

3.3 Nonlinear Least Squares Estimators

Unlike the problem of estimating amplitudes and linear prediction coefficients, the problem of estimating frequencies is nonlinear and is well-known to be very difficult to solve [43], particularly without intimate knowledge of the problem at hand. More specifically, the problem of estimating the frequencies $\{\omega_l\}_{l=1}^{L}$ of the model in (4) using (27) is nonlinear and the cost function is multi-modal and there exists no good way to solve it directly [43]. Instead, approximate solutions, based on for example the periodogram, are used, or iterative methods that solve for one component at the time are employed [96, 97]. It is then perhaps not surprising that many of the parameter estimation problems that the scientific community continues to work on are those involving nonlinear parameters, such as angles, frequencies, and damping factors. Concerning the estimation of such onlinear parameters, it should also be remarked that estimators tend to exhibit so-called threshold behavior for such problems. This refers to the phenomenon that below a certain SNR or number of samples, N, the estimator will essentially break down and produce meaningless estimates [85]. Hence, a big issue in finding good methods for estimating nonlinear parameters is at which point this behavior occurs. In fact, this may be a much bigger issue in practice than whether an estimator is efficient, or merely consistent, when dealing with speech and audio signals, where the SNRs are often very poor and the number of samples low! For example, it is sometimes argued that because subspace and optimal filtering methods are not statistically efficient for the analysis of time-series[4], they are not worthwhile investigating, but if they work at lower SNRs and N or are more robust to model-mismatch, interference, etc., then they may still be useful in adverse conditions.

The principle of maximum likelihood estimation can be, and has been, applied to the models considered in this thesis. For the specific problem of estimating the fundamental frequency of the harmonic model in (6), the estimator in (27) reduces to the following:

$$\hat{\omega}_0 = \arg\max_{\omega_0} \mathbf{x}^H \mathbf{R}_{\mathbf{w}}^{-1} \mathbf{Z} \left(\mathbf{Z}^H \mathbf{R}_{\mathbf{w}}^{-1} \mathbf{Z} \right)^{-1} \mathbf{Z}^H \mathbf{R}_{\mathbf{w}}^{-1} \mathbf{x}. \tag{30}$$

This estimator is exact, as no approximations have been used. Moreover, it explicitly takes the presence of colored noise into account. For the case of white noise, i.e., $\mathbf{R}_{\mathbf{w}} = \sigma_w^2 \mathbf{I}_M$, the estimator reduces to the following estimator,

[4] More specifically, they require that $M < N$ which in turn leads to a loss of optimality.

which is termed the nonlinear least squares (NLS) estimator:

$$\hat{\omega}_0 = \arg\max_{\omega_0} \mathbf{x}^H \mathbf{Z} \left(\mathbf{Z}^H \mathbf{Z} \right)^{-1} \mathbf{Z}^H \mathbf{x}. \tag{31}$$

This estimator is considered in [C] for the problem of estimating low fundamental frequencies using the harmonic model (6) where it is shown to have excellent performance. Moreover, an estimator based on this principle is also proposed in [D] for the harmonic chirp model, which is more complicated than for the harmonic model, as it comprises an additional nonlinear parameter.

By employing an asymptotic approximations in (13) and (14) in the estimator (30), we obtain the approximate nonlinear weighted least squares estimator for the colored noise case:

$$\hat{\omega}_0 = \arg\max_{\omega_0} \sum_{l=1}^{L} |X(\omega_0 l)|^2 / |W(\omega_0 l)|^2. \tag{32}$$

where $X(\omega) = \sum_{n=0}^{N-1} x(n) e^{-j\omega n}$, and similarly for $W(\omega)$, are periodogram estimates of the corresponding psds. When the noise is white, the estimator reduces to $\hat{\omega}_0 = \arg\max_{\omega_0} \sum_{l=1}^{L} |X(\omega_0 l)|^2$ which called the approximate nonlinear least squares (ANLS) method and is equivalent to the classical harmonic summation method first proposed in [98]. This simplified estimator is considered as the foundation for multi-pitch estimation in [38], where also an iterative method based on the expectation maximization algorithm [99] is proposed. This idea is similar to the one of [100], only that was based on a comb filter, which cannot be statistically optimal due to the memory of the filters.

The history of nonlinear least squares frequency estimation based on the harmonic model goes back at least to [101] and has also been investigated in [72, 102]. Despite having been known for a long time and having excellent performance, these estimators never found widespread use in the speech and audio community. The most likely explanation for this is that the computational complexity involved in solving (30) or (31) has been prohibitive for most applications. Moreover, the complications associated with also having to find the number of harmonics, a problem that does not occur directly in non-parametric methods, has probably also deterred many from using them. Recently, however, it has been shown that the exact estimator in (30) can be solved efficiently by exploiting the Hankel-plus-Toeplitz structure of this problem [103], which result in a complexity comparable to that of harmonic summation. When combined with a hidden Markov model (HMM) that exploits the smooth development of the fundamental frequency over time, state-of-the-art performance and robustness is obtained [104]. The remaining only remaining issue with maximum likelihood fundamental frequency estimation is how to best deal with colored noise. While solving for

the fundamental frequency using (30) appears straightforward in this case, the main problems are that the noise statistics are unknown $\mathbf{R_w}$ and that no fast implementation is known for this case, as the principle of [103] does not apply. To mitigate this, [D] proposes to use pre-whitening based on noise trackers such as [105–107]. Other options include iterative methods such as [108] and order-recursive decompositions [78, 109] and NMF-based adaptive prewhitening [110].

3.4 Subspace-based Estimators

The class of subspace methods is based on the following principles. We will here consider the form $\mathbf{x}(n) = \mathbf{Z}\mathbf{a}(n) + \mathbf{w}(n)$ of the model (9) with $\mathbf{x}(n) \in \mathbb{R}^M$ and $M < N$, for which the covariance matrix then is given by (11). For the case where the noise is white, i.e., $\mathbf{R_w} = \sigma_w^2 \mathbf{I}_M$, where the covariance matrix reduces to a scaled diagonal matrix, the covariance matrix model becomes

$$\mathbf{R_x} = \mathbf{Z}\mathbf{P}\mathbf{Z}^H + \sigma_w^2 \mathbf{I}_M, \tag{33}$$

where $\mathbf{Z}\mathbf{P}\mathbf{Z}^H$ has rank $L < M$ and $\mathbf{P} = \text{diag}\left(\left[\begin{array}{ccc} A_1^2 & \cdots & A_L^2 \end{array}\right]\right)$. When the noise is not white, pre-whitening can be applied, as for the case of the maximum likelihood estimator [111]. Let $\mathbf{R_x} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$ be the eigenvalue decomposition (EVD) of the $\mathbf{R_x}$, and let the matrices $\mathbf{S}_L$ and $\mathbf{G}_L$, whose columns span spaces which are termed the signal and noise subspaces, respectively, be formed as

$$\mathbf{S}_L = \left[\begin{array}{ccc} \mathbf{u}_1 & \cdots & \mathbf{u}_L \end{array}\right] \tag{34}$$

$$\mathbf{G}_L = \left[\begin{array}{ccc} \mathbf{u}_{L+1} & \cdots & \mathbf{u}_M \end{array}\right], \tag{35}$$

i.e., $\mathbf{S}_L$ is formed from the eigenvectors $\mathbf{u}_k$ corresponding to the L largest eigenvalues while $\mathbf{G}_L$ is formed from the eigenvectors $\mathbf{u}_k$ corresponding to the $M - L$ smallest ones. Since the covariance matrix is Hermitian, its eigenvectors are orthogonal, so we have that $\mathbf{S}_L^H \mathbf{G}_L = \mathbf{0}$. Then it can be shown that $\mathcal{R}(\mathbf{S}_L) = \mathcal{R}(\mathbf{Z})$ (where $\mathcal{R}(\cdot)$ denotes the range) and that $\mathcal{R}(\mathbf{Z}) \perp \mathcal{R}(\mathbf{G}_L)$, i.e., that columns of the matrix $\mathbf{Z}$ generated by the nonlinear parameters $\boldsymbol{\xi}$ are orthogonal to all the columns of $\mathbf{G}_L$. This can be exploited to obtain estimates. More specifically, by measuring the angles between subspaces [112], we can obtain an estimate of the nonlinear parameters as

$$\hat{\boldsymbol{\xi}} = \arg\min_{\boldsymbol{\xi}} \|\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{G}_L\mathbf{G}_L^H\|_F^2. \tag{36}$$

This maximizes the sum of cosine to the angles between the subspaces $\mathcal{R}(\mathbf{Z})$ and $\mathcal{R}(\mathbf{G}_L)$ squared [A]. For the case where the columns of $\mathbf{Z}$ are orthogonal, or approximately so, and are generated by individual parameters in $\boldsymbol{\xi}$, the cost function in (36) decouples into L independent minimizations [78]. This is the case, for example, for the sinusoidal model in (4) but not for the

harmonic model. For both the sinusoidal model (4) and the harmonic model (6), the expression in (36) can be simplified by applying the asymptotic approximation $\mathbf{Z}^H\mathbf{Z} \approx M\mathbf{I}_L$ and noticing that the product $\mathbf{Z}^H\mathbf{G}_L$ for different parameters can be computed using fast Fourier transforms (FFTs). The principle of subspace orthogonality can be used for estimating the fundamental frequency, as first shown in [113]. This work was later extended to the multi-pitch case in [38] and refined further in paper [C]. Concerning the estimation of the eigenvectors, which is often the dominant source of computational complexity for these methods, subspace trackers can be used for efficient, time-recursive computation of these [114–119].

As shown in [113] and [A], it is also possible to use the principle for estimating the model order, L, using the principle of angles between subspaces. In that case, however, the criterion in (36) must be modified to as follows. Since the dimensions of both $\mathbf{Z}$ and $\mathbf{G}_L$ depend on the model order, L, it is nontrivial to measure and compare the angles between two subspaces. However, averaging instead over the cosine to the principal angles squared, we obtain the following measure, which was first proposed in [A] and used for fundamental frequency estimation in [78]:

$$\{\hat{\boldsymbol{\xi}}, \hat{L}\} = \arg\min_{\boldsymbol{\xi},L} \frac{1}{MT} \|\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{G}_L\mathbf{G}_L^H\|_F^2, \tag{37}$$

where $T = \min\{L, M-L\}$. As previously mentioned, there exists quite a few different approaches to model order selection. Classical examples of this are the minimum description length [4], Akaike's information criterion [5, 120], the Bayesian information criterion [121], and hybrid methods based on the g-prior [122]. The difficulty in several of these is that it might be quite difficult, depending on the model, to determine the appropriate penalty-term. For sinusoidal and polynomial models, dedicated works include [123–126]. In the context of subspace methods, it is worth noting that [6] expressed the order estimation problem for certain models in terms of the eigenvalues of the covariance matrix (see also [127]), and that in [128] a criterion for estimating the model order based on the shift-invariance principle was derived.

Another related subspace method is the ESPRIT method [129], which exploits the shift-invariance property of the matrix $\mathbf{Z}$ that some models, such as the sinusoidal model, exhibit. It was shown [C] (and initially in [130]) that it is possible to exploit this property for the estimation of the fundamental frequency. Define the matrices $\underline{\mathbf{Z}}$ and $\overline{\mathbf{Z}}$ from $\mathbf{Z}$ as

$$\underline{\mathbf{Z}} = [\ \mathbf{I}_{M-1}\ \mathbf{0}\ ]\,\mathbf{Z} \quad \text{and} \quad \overline{\mathbf{Z}} = [\ \mathbf{0}\ \mathbf{I}_{M-1}\ ]\,\mathbf{Z}. \tag{38}$$

Similarly, we obtain $\underline{\mathbf{S}}_L$ and $\overline{\mathbf{S}}_L$ from $\mathbf{S}_L$. For the two matrices, $\overline{\mathbf{Z}}$ from $\mathbf{Z}$, we have that $\overline{\mathbf{Z}} = \underline{\mathbf{Z}}\mathbf{D}$ with

$$\mathbf{D} = \operatorname{diag}\left(\left[\ e^{j\omega_0}\ \cdots\ e^{j\omega_0 L}\ \right]\right), \tag{39}$$

due the the shift-invariance of the model. Also, since the columns of the matrices $\mathbf{S}_L$ and $\mathbf{Z}$ span the same space, it follows that they can be related as $\mathbf{S}_L = \mathbf{Z}\mathbf{Q}$ where $\mathbf{Q} \in \mathbb{C}^{L\times L}$ is an invertible matrix. Then we have that $\overline{\mathbf{S}}_L = \underline{\mathbf{S}}_L\mathbf{\Sigma}$ with $\mathbf{\Sigma} = \mathbf{Q}^{-1}\mathbf{D}\mathbf{Q}$. This can be used for estimating frequencies by first finding the eigenvectors in $\mathbf{S}_L$ from an estimate of the covariance matrix, from which $\overline{\mathbf{S}}_L$ and $\underline{\mathbf{S}}_L$ can be computed, and then finding an estimate $\widehat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$ using least squares or total least squares. Then, the matrix $\mathbf{D}$ contains the eigenvalues of $\widehat{\mathbf{\Sigma}}$ from which the frequencies can be found [43, 129]. To use the principle for estimation of the fundamental frequency is, however, more complicated but can be achieved as follows: Let $\widehat{\mathbf{\Sigma}} = \mathbf{C}\widehat{\mathbf{D}}\mathbf{C}^{-1}$ denote the empirical EVD of $\widehat{\mathbf{\Sigma}}$ and let the entries in $\widehat{\mathbf{D}}$ and the corresponding eigenvectors be sorted such that $\hat{\omega}_1 < \ldots < \hat{\omega}_L$ and $\widehat{\mathbf{D}} = \text{diag}\left(\begin{bmatrix} e^{j\hat{\omega}_1} & \cdots & e^{j\hat{\omega}_L} \end{bmatrix}\right)$. Then the fundamental frequency can be estimated as

$$\hat{\omega}_0 = \arg\min_{\omega_0} \|\overline{\mathbf{S}}_L\mathbf{C} - \underline{\mathbf{S}}_L\mathbf{C}\mathbf{D}\|_F^2, \tag{40}$$

where the matrix $\mathbf{D}$ is constructed from the candidate fundamental frequency using (39). Once the EVDs have been performed, this results in a very simple estimator with a very smooth cost function [130]. As we have seen, thus both the shift-invariance property and the subspace orthogonality principle can be used for estimating not only the fundamental frequency but also the model order without resorting to statistical principles. It is interesting to note that while subspace methods, such as MUSIC [127] (see also [131]), ESPRIT [129], unitary ESPRIT [132], and weighted subspace fitting [133] are well-known in communications and other fields where they have been applied to many problems, they are much less frequently used in speech and audio. One possible explanation for this may be that, as mentioned earlier, background noise in speech and audio signals are rarely white and most often non-stationary, which makes it difficult to apply the principles of subspace methods in practice. Another possible explanation is that when these subspace methods were first invented, they were applied to array signal processing problems. Compared to speech and audio signals, such problems tend to have a much lower dimensionality, so applying these principles to speech and audio processing may simply have been impractical at the time.

3.5 Sparsity-based Estimators

Much progress has been made in signal processing in the past couple of decades based on the ideas of sparse approximations, which can be traced back to early work such as matching pursuit [134] and basis pursuit [135]. These actually made an early impact in speech and audio processing, with pioneering work such as [136] and [137] being early examples, and were only picked up much later by others in the general signal processing community. Curiously, the ideas of exploiting sparsity can be traced quite far

back in speech and audio processing, and some of the algorithms for sparse approximations are identical to algorithms that had been known for a long time [138]. These ideas can be applied to the models and estimation problems considered herein, as will be described next. As before, consider the general complex model of the signal $\mathbf{x} \in \mathbb{C}^N$ given by

$$\mathbf{x} = \mathbf{Z}\mathbf{a} + \mathbf{w}. \tag{41}$$

In the terminology of sparse approximations, $\mathbf{Z} \in \mathbb{C}^{N \times L}$ is referred to as the dictionary with each column being a so-called atom and $\mathbf{a} \in \mathbb{C}^L$ as the coefficients. Unlike the cases considered so far, the dictionary, $\mathbf{Z}$, is typically a fat matrix with $L \geq N$ in sparse approximations, and very often $L \gg N$. The idea is then to model $\mathbf{x}$ using only as few columns of $\mathbf{Z}$ as possible, meaning that $\mathbf{a}$ should contain only a few non-zero coefficients. Basis pursuit [135] achieves this by casting the problem as the following convex optimization problem:

$$\underset{\mathbf{a}}{\text{minimize}} \, \|\mathbf{a}\|_1 \quad \text{s.t.} \quad \mathbf{Z}\mathbf{a} = \mathbf{x}, \tag{42}$$

where $\|\cdot\|_1$ denotes the 1-norm. For the case of real quantities, this problem can be shown to be a linear programming problem while for the complex case considered here, it is a second-order cone programming problem. The relation to the problems considered herein is then that the dictionary, $\mathbf{Z}$, may be populated with all the possible models, for example the complex exponential functions of the sinusoidal model, with $\mathbf{Z}$ being defined as (4) only with $\omega_l = 2\pi\frac{l-1}{L}$ for $l = 1, \ldots, L$ and $L \gg N$, and then a sparse coefficient vector, $\mathbf{a}$, corresponds to selecting a few of those. As can be seen, the application of these principles to spectral analysis is relative straightforward.

Although conceptually simple and intuitive, there are multiple problems with using (42) for sparse approximations. First, the 1-norm is not a good measure of sparsity (although it does produce results that are optimal also in the sense of sparsity in some cases), and, second, the problem in (42) ignores the presence of noise. Much of the early work on sparse approximation was based on convex optimization and considered the problem as a deterministic one, which made the connection to estimation theory and statistical interpretations unclear. Yet such a connection clearly exists. For example, matching pursuit [134] is based on an iterative minimization of the 2-norm of the residual, which means that it can be thought of as an approximate nonlinear least squares method, and it is thus an approximate maximum likelihood method for white Gaussian noise, as was noted in [139]. Many early methods for sparse approximations rely on oracle information, such as noise variance, to implement, for example, a threshold on the coefficients on $\mathbf{a}$ to achieve sparsity and to take the presence of noise into account [140], and this complicates their application to speech and audio signals. An alternative definition of the

sparse approximation problem that explicitly takes the presence of noise into account is the following, which is referred to as basis pursuit denoising:

$$\underset{\mathbf{a}}{\text{minimize}} \|\mathbf{a}\|_1 \quad \text{s.t.} \quad \|\mathbf{x} - \mathbf{Z}\mathbf{a}\|_2 \leq \epsilon, \tag{43}$$

where $\|\cdot\|_2$ is the 2-norm. As can be seen, this also relies on information concerning the noise, $\mathbf{w}$, via the constraint ϵ. A method closely related to this way of posing the sparse approximation problem is the Lasso, which was originally proposed in [141] and later developed further in [142–144], wherein a fitting criterion is combined with a penalty that promotes sparsity, i.e.,

$$\underset{\mathbf{a}}{\text{minimize}} \|\mathbf{x} - \mathbf{Z}\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_1, \tag{44}$$

where $\lambda \geq 0$ is then used to control the level of sparsity. It can be interpreted in several ways. In relation to (43), it can be seen as a Lagrange multiplier related to the constraint. Then, (44) is simply the Lagragian associated with (43). Another possible interpretation is that the term $\lambda\|\mathbf{a}\|_1$ implements a prior, in this case a Laplacian prior, on the entries in the vector $\mathbf{a}$ while $\|\mathbf{x} - \mathbf{Z}\mathbf{a}\|_2^2$ is the likelihood, and the estimator in (44) is then a maximum a posteriori estimator. This relation also makes it clear how colored noise should be taken into account in the estimation process, as the relation to maximum likelihood estimation is now clear.

While the application of these principles to the sinusoidal model is relatively straightforward, their application to the harmonic model is more difficult, as the dictionary would need to contain a multitude of possible harmonic models, corresponding to different fundamental frequencies. In that case, simply promoting sparsity in the associated coefficient vector is insufficient. Instead, it is necessary to introduce the principle of block-sparsity to solve this problem [145]. This way of posing the problem naturally extends to the case of multi-pitch estimation [146–148], and it has also been applied to the harmonic chirp model [149]. However, it also limits us to a finite set of possible fundamental frequencies, but this has also been addressed in the context of the harmonic model [150]. There was an early attempt at using sparse approximations with the harmonic model in the harmonic matching pursuit [151]. Interestingly, this algorithm is mathematically quite closely related to the EM-based algorithm proposed in [38], only no re-estimation is used and different signal estimates are used in the process. However, the algorithm in [151] does not take an unknown number of harmonics into account and thus only works for very simple signals.

The principles of sparse approximation can also be applied to the linear prediction model in a number of ways, as first proposed in [F]. Let $\mathbf{x}$, $\mathbf{Z}$, $\mathbf{a}$, and $\mathbf{w}$ now be defined as in (17). Recall that in the traditional linear prediction model, the excitation, $\mathbf{w}$ is assumed to be Gaussian distributed, which causes

many problems, particularly for voiced speech where the excitation is better modeled as a pulse train. Over the years, many different approaches and heuristics have been proposed solve these problems, e.g., [59, 152]. Suppose instead that $\mathbf{w}$ is instead Laplace distributed, then the problem of estimating the linear prediction coefficients is given by

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{Z}\mathbf{a}\|_1. \tag{45}$$

Then, this leads to an estimate of $\mathbf{w}$ given by $\hat{\mathbf{w}} = \mathbf{x} - \mathbf{Z}\hat{\mathbf{a}}$ that is closer to being sparse than that of the traditional linear prediction wherein the 2-norm is minimized. Therefore, this method is called sparse linear prediction. This idea can be traced back to [153]. Indeed, as shown in [F], this approach solves several of the problems associated with the traditional linear prediction, such as the dependency of the estimates $\hat{\mathbf{a}}$ on pitch and time-shifts. It should be stressed that the formulation in (45) is different from those of traditional sparse approximation methods wherein the coefficient vector, $\mathbf{a}$, is sparse. Rather it is here in the fitting criterion that sparsity is promoted and the dictionary is generated from the signal itself. It is, though, also possible to exploit sparsity in the coefficient vector in connection with linear prediction, as we will see next. Speech coders based on linear prediction frequently employ two predictors: a short-term predictor that captures the effect of the vocal tract and a long-term predictor that captures the effect of the pitch (see, e.g., [154]). However, it is also possible combine these two predictors which results in a high-order, but sparse, linear predictor. This idea, combined with the sparse linear prediction in (45), leads to the following generalized sparse linear prediction problem:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{Z}\mathbf{a}\|_p^p + \lambda \|\mathbf{a}\|_q^q. \tag{46}$$

As before, the parameter $\lambda \geq 0$ controls the tradeoff between the fitting term and the penalty term, while p and q determine in which sense these should be minimized. For $p = 2$ and $q = 1$ we obtain the Lasso while for $p = 1$ and $\lambda = 0$ we obtain the sparse linear prediction of (45). To obtain the aforementioned high-order sparse linear predictor, we can select $p = 2$, $q = 1$, and $\lambda > 0$. It is interesting to note the similarities between (46) and statistical methods for model selection (see, e.g., [7]), as both feature a fitting term and a penalty term. More sophisticated predictors, such as those obtained with (45) or (46), often come at the cost of added computational complexity, and this is also the case here. In [60,61], dedicated real-time solvers exploiting the matrix structures for the different sparse linear prediction problems were proposed, and in [64] it was shown how to modify the optimization problems to ensure stable filters, e.g., via additional constraints on $\mathbf{a}$. So, the two reasons for not using the sparse linear prediction, namely the stability and complexity issues, have been solved. Interestingly, sparse linear prediction has found

some uses in speech dereverberation [155, 156], an unsolved problem that continues cause problems in hands-free communication and distant speech recognition.

In connection with coding, a different interpretation of the sparse linear prediction than the estimation theoretical one exists. Given that the speech in some speech coders is reconstructed using sparse patterns in the excitation, such as regular-pulse excitation [157] and multi-pulse excitation [158], the optimal predictor for this problem is then, arguably, not one that minimizes the 2-norm, as in traditional linear prediction, but one that takes the sparse nature of the excitation in the decoder into account. For an overview of such speech coders and their history, we refer the interested reader to [65].

Finally, it should be remarked that over the past couple of decades an abundance of different methods for sparse approximations in different contexts have been proposed, aside from those already mentioned here, including orthogonal matching pursuit [159], sparse Bayesian learning [160], subspace pursuit [161], and the re-weighted 1-norm and 2-norm methods [162, 163]. New applications of these principles to speech and audio continue to emerge, including beamforming [164, 165]. Important advances have also been made that put sparse approximations on firmer mathematical ground. Principles such as sparse Bayesian learning [160] and the horseshoe prior [166] appear to be particularly promising and capable of overcoming the problems of earlier methods, such as the selection of the appropriate regularization constant and the crude measures of sparsity employed in some methods.

4 Model-based Filtering

Linear filtering is one of the most used tools of signal processing practitioners and academics alike. In relation to the problems considered in this thesis, it is also the predominant way of enhancing and separating signals. Optimal filtering, as understood in the context of stochastic signals, is the foundation on which echo/noise cancellation, noise reduction, source separation, and beamforming are built (see, e.g., [167]). In these problems, optimal filters, which can be either adaptive or static (e.g., [168]), are derived as solutions to optimization problems stated in terms of signal statistics, such as the minimization of the MSE or the maximization of the output power of a filter. In most of these applications of linear filters, the involved signals are characterized in terms of their second-order statistics, such as covariance matrices, power spectral densities, or coherence matrices. As such, they employ what may be characterized as vague models of the involved signals. In contrast, we will here show how the parametric signal models considered so far can be incorporated into optimal filtering.

The Wiener filter and its extensions remain the de facto standard for signal enhancement, despite it being a certain special case of a more general class of filters (see [169]). In source separation, for example, the frequency-domain Wiener filter (or some variation thereof) is typically used [13]. Multi-channel enhancement methods often employ a multi-channel Wiener filter [170–172], or features some combination of a beamformer and a postfilter, a construction which is equivalent to the multi-channel Wiener when a minimum variance distortionless response (MVDR) beamformer and a Wiener filter are used [173]. It is interesting to note that much of the research in speech enhancement over the past couple of decades has focused on the problem of determining the statistics required by the optimal filters, something that is typically achieved via noise trackers [105–107, 174], although model-based methods, which tend to perform better for non-stationary noise types and in the presence of multiple speakers, also have been proposed [175–179].

In the preceding sections, we have discussed models of speech and audio signals and how to find the parameters of these models in adverse conditions, and we will now turn our attention to how these models, and the knowledge gained from them, can be used for processing of signals. By incorporating signal models in the processing of signals, we are essentially infusing our domain knowledge about the problem at hand into the problem definition and its solution. In doing this, we will focus on how this can be done in linear filtering for noise reduction and beamforming, although the principle, of course, holds more generally, and can be applied to a wealth of problems (e.g., [180, 181]). Aside from the models discussed in this thesis, it should be remarked that another type of model-based optimal filtering that, although interesting, will not be discussed further here, is the Kalman filter, which via its state equation enables the incorporation of certain models in optimal filtering (see, e.g., [72, 177, 182–184] for some examples of this).

4.1 Classical Optimal Filtering

Before going into details about more sophisticated methods and model-based filtering, we will first review classical optimal filtering methodology for noise reduction on which the majority of contemporary methods is still based. Consider the general model of the observed signal $\mathbf{x}(n) \in \mathbb{R}^M$ defined as

$$\mathbf{x}(n) = \mathbf{s}(n) + \mathbf{w}(n), \tag{47}$$

where $\mathbf{s}(n)$ is the signal of interest and $\mathbf{w}(n)$ the noise, which are assumed to be statistically independent. Moreover, all signals are assumed to be zero-mean. Then, the purpose in noise reduction is to design a filter $\mathbf{h} \in \mathbb{R}^M$ that when applied to the input signal produces as its output, $y(n)$, an estimate of

$s(n)$, i.e.,

$$y(n) = \mathbf{h}^H \mathbf{x}(n) \tag{48}$$

$$= \mathbf{h}^H \mathbf{s}(n) + \mathbf{h}^H \mathbf{w}(n). \tag{49}$$

In estimation theoretical terms, linear filtering belongs to the class of linear estimators, where it is the special case concerned with estimation from an ordered set of samples using a linear combination of those samples (see, e.g., [85]). As can be seen from (49), the linear filter affects the signal of interest, $\mathbf{s}(n)$, and the noise, $\mathbf{w}(n)$. Ideally, we would like for the filter to attenuate the noise while leaving the signal of interest unchanged, but these two requirements are often at odds with each other, and noise reduction comes at the cost of signal distortion. Noise reduction is then the art of finding a compromise between these two conflicting demands. For example, a filter can be designed by minimizing the MSE between the output of the filter, $y(n)$, and the signal of interest $s(n)$, i.e.,

$$\text{MSE} = \mathrm{E}\left\{ (s(n) - y(n))^2 \right\} \tag{50}$$

$$= \mathrm{var}(s(n)) + \mathbf{h}^H \mathbf{R}_\mathbf{x} \mathbf{h} - 2\mathbf{h}^H \boldsymbol{\rho}, \tag{51}$$

where $\mathbf{R}_\mathbf{x}$ is the covariance matrix of $\mathbf{x}(n)$ and $\boldsymbol{\rho} = \mathrm{E}\{s(n)\mathbf{x}(n)\}$ is the cross-correlation between $s(n)$ and $\mathbf{x}(n)$. The solution that minimizes this MSE is the well-known Wiener filter [167] which is given by

$$\mathbf{h}^\star = \mathbf{R}_\mathbf{x}^{-1} \boldsymbol{\rho}. \tag{52}$$

Observe that $s(n) = \mathbf{s}^H(n)\mathbf{i}_1$ where $\mathbf{i}_m$ is the mth column of the identity matrix $\mathbf{I}_M \in \mathbb{R}^{M \times M}$ and that $\boldsymbol{\rho} = \mathrm{E}\{s(n)\mathbf{s}(n)\}$ due to independence. From this it can easily be seen that the cross-correlation in (52), $\boldsymbol{\rho}$, can be expressed as $\boldsymbol{\rho} = \mathbf{R}_\mathbf{s}\mathbf{i}_1$ where $\mathbf{R}_\mathbf{s} = \mathrm{E}\left\{\mathbf{s}(n)\mathbf{s}^H(n)\right\}$. To obtain the ubiquitous frequency-domain formulation of the Wiener filter, the asymptotic approximation in (14), can be used, i.e.,

$$\mathbf{F}^H \mathbf{h}^\star = \mathbf{F}^H \mathbf{R}_\mathbf{x}^{-1} \mathbf{R}_\mathbf{s} \mathbf{i}_1 \tag{53}$$

$$\approx \boldsymbol{\Gamma}_\mathbf{x}^{-1} \boldsymbol{\Gamma}_\mathbf{s} \mathbf{1}_M \tag{54}$$

$$= \left[\frac{P_s(\omega_1)}{P_x(\omega_1)} \; \frac{P_s(\omega_2)}{P_x(\omega_2)} \; \cdots \frac{P_s(\omega_M)}{P_x(\omega_M)} \right]^T. \tag{55}$$

It should be stressed that there is only generally equivalence between the frequency- and time-domain Wiener filters in the case of either infinitely long filters or periodic signals having certain periods. Due to the Wiener filter being based on the MSE between the signal of interest and the output of the filter, it offers what is essentially a pre-determined tradeoff between signal distortion and noise reduction [169].

Since we have that $\mathbf{R}_\mathbf{x} = \mathbf{R}_\mathbf{s} + \mathbf{R}_\mathbf{w}$, the time- and frequency-domain Wiener filters in (52) and (55) can be computed in a number of equivalent ways, depending on what quantities can most easily be estimated in a given context. An estimate $\widehat{\mathbf{R}}_\mathbf{x}$ of $\mathbf{R}_\mathbf{x}$ can be obtained from the observed signal but then either $\mathbf{R}_\mathbf{s}$ or $\mathbf{R}_\mathbf{w}$ has to be estimated in some other way, and typically $\mathbf{R}_\mathbf{w}$ is estimated. Indeed, much of the research on noise reduction in the past two decades has essentially focused on finding the involved signal statistics, most often the statistics of the noise [11, 105, 106, 185], but there are also methods that estimate both speech and noise statistics jointly [175]. The model-based approach can also be used for solving this problem, which is particularly useful for non-stationary noise [186]. For example, the covariance matrix model in (11) can be used to express $\mathbf{R}_\mathbf{s}$ in terms of the model and thus parametrizes the unknown covariance matrix of the signal of interest, i.e.,

$$\begin{aligned} \mathbf{R}_\mathbf{x} &= \mathbf{R}_\mathbf{s} + \mathbf{R}_\mathbf{w} && (56) \\ &= \mathbf{Z}\mathbf{P}\mathbf{Z}^H + \mathbf{R}_\mathbf{w}. && (57) \end{aligned}$$

Then, the estimate $\widehat{\mathbf{R}}_\mathbf{s} = \mathbf{Z}\mathbf{P}\mathbf{Z}^H$ of $\mathbf{R}_\mathbf{s}$ can then either be used directly to compute an optimal filter based on estimates of the nonlinear parameters $\boldsymbol{\xi}$ characterizing the columns of $\mathbf{Z}$ and the amplitudes $\mathbf{a}$, such as the Wiener filter in (52), or it can be used to estimate the noise covariance matrix as $\widehat{\mathbf{R}}_\mathbf{w} = \widehat{\mathbf{R}}_\mathbf{x} - \mathbf{Z}\mathbf{P}\mathbf{Z}^H$. However, there are many more ways in which the signal models and their estimators can be exploited in connection with linear filtering for noise reduction (see, e.g., [187]), some of which will be explored next.

4.2 Model-based Optimal Filtering

Distortionless optimal filter designs have a rich history in both beamforming and spectral analysis [188–190] but are rare in speech and audio processing, except for in beamformers. A way to exploit the model of the observed signal directly in the design of optimal filters will be explored next. As it turns out, such filters can be used for both noise reduction and parameter estimation when combined with signal models. Recall that the filtered observed signal in (49) can be expressed also in terms of the model as

$$y(n) = \mathbf{h}^H\mathbf{Z}(n)\mathbf{a} + \mathbf{h}^H\mathbf{w}(n). \tag{58}$$

To obtain a distortionless estimate that minimizes the detrimental effect of the filter on the signal of interest, we can impose the constraint $\mathbf{h}^H\mathbf{Z}(n) = \mathbf{i}_1^T\mathbf{Z}(n)$ [B]. More generally, however, we can obtain an estimate of an arbitrary sample of $s(n)$ by appropriately selecting the corresponding constraint as $\mathbf{h}^H\mathbf{Z}(n) = \mathbf{i}_m^T\mathbf{Z}(n)$ where $\mathbf{i}_m$ is the mth column of the identity matrix $\mathbf{I}_M \in \mathbb{R}^{M\times M}$, as this then leads to $y(n)$ being an estimate of sample $s(n+m-1)$. Aside from this constraint, we must choose an objective function to be

minimized (or maximized) to find an optimal filter. Assuming independence between $\mathbf{s}(n)$ and $\mathbf{w}(n)$, the output power of the filter can be expressed as

$$\mathbf{h}^H \mathbf{R}_\mathbf{x} \mathbf{h} = \mathbf{h}^H \mathbf{R}_\mathbf{s} \mathbf{h} + \mathbf{h}^H \mathbf{R}_\mathbf{w} \mathbf{h}. \tag{59}$$

Clearly, from (59) we observe that to attenuate the noise, we must minimize the power of the term $\mathbf{h}^H \mathbf{w}(n)$, which is equal to $\mathbf{h}^H \mathbf{R}_\mathbf{w} \mathbf{h}$. Yet, at the same time, it makes sense to instead minimize the output power of the filters, i.e., maximize $\mathbf{h}^H \mathbf{R}_\mathbf{x} \mathbf{h}$, as the filters are already distortionless. Indeed, as shown in [190] for a related problem, the solutions (i.e., the resulting filters) are identical for the two problems. This means that in finding the optimal filter for noise reduction, we can simply use $\mathbf{R}_\mathbf{x}$, which is directly available, and do not have to solve the somewhat complicated problem of estimating $\mathbf{R}_\mathbf{w}$. However, for clarity we will proceed to use $\mathbf{R}_\mathbf{w}$ in what follows. We can pose the optimal filter design problem as the following quadratic optimization problem with linear constraints [B]:

$$\underset{\mathbf{h}}{\text{minimize}}\, \mathbf{h}^H \mathbf{R}_\mathbf{w} \mathbf{h} \quad \text{s.t.} \quad \mathbf{Z}^H(n)\mathbf{h} = \mathbf{Z}^H(n)\mathbf{i}_m, \tag{60}$$

which has the well-known solution [191]

$$\mathbf{h}^\star = \mathbf{R}_\mathbf{w}^{-1} \mathbf{Z}(n) \left(\mathbf{Z}^H(n) \mathbf{R}_\mathbf{w}^{-1} \mathbf{Z}(n) \right)^{-1} \mathbf{Z}^H(n) \mathbf{i}_m. \tag{61}$$

The above filter design is an example of what can be achieved with a model-based approach. We can now use this filter for different purposes. For noise reduction the filter in (61) is not only distortionless–it is also optimal, in the sense of maximizing the output SNR, as $\mathbf{h}^H \mathbf{R}_\mathbf{w} \mathbf{h}$ is minimized. It should be noted that the filter is only truly distortionless given perfect knowledge of the matrix $\mathbf{Z}(n)$. In the presence of noise, estimation errors will lead to $\mathbf{Z}(n)$ not being known perfectly, which, in turn, will lead to distortion of the signal of interest. In that case, the equivalence between using $\mathbf{R}_\mathbf{x}$ and $\mathbf{R}_\mathbf{s}$ no longer holds. Also, it should be noted that for the case that $L = M$, the solution to the filter design problem is trivial. The constraint above can be implemented using the principle of the generalized sidelobe canceller [192], which has found many uses in speech and audio processing (e.g., [193, 194]), whereby the solution can be found using standard adaptive filtering algorithms.

Meanwhile, to estimate the nonlinear parameters of the matrix $\mathbf{Z}(n)$ using such optimal filters, we insert the solution (61), with $\mathbf{R}_\mathbf{w}$ replaced by $\mathbf{R}_\mathbf{x}$, into $\mathbf{h}^H \mathbf{R}_\mathbf{x} \mathbf{h}$ and then maximize the output power, i.e.,

$$\hat{\boldsymbol{\xi}} = \arg \max_{\boldsymbol{\xi}} \mathbf{i}_m^H \mathbf{Z}(n) \left(\mathbf{Z}^H(n) \mathbf{R}_\mathbf{x}^{-1} \mathbf{Z}(n) \right)^{-1} \mathbf{Z}^H(n) \mathbf{i}_m. \tag{62}$$

For the case where $m = 1$ and $n = 0$ we have that for the models in (4) and (6) the constraint reduces to $\mathbf{Z}^H \mathbf{h} = \mathbf{1}_L$ with $\mathbf{1}_L = [\, 1 \; \cdots \; 1 \,]^T \in \mathbb{R}^L$. In that

case, the estimator above reduces to

$$\hat{\xi} = \arg\max_{\xi} \mathbf{1}_L^H \left(\mathbf{Z}^H \mathbf{R}_{\mathbf{x}}^{-1} \mathbf{Z} \right)^{-1} \mathbf{1}_L. \tag{63}$$

This principle has been used for a number of estimation problems, such as spectral analysis, amplitude estimation, direction-of-arrival estimation, problems which all have similar models (see, e.g., [43]). Several contributions have been made in this thesis based on the principle of optimal filtering for parameter estimation. An estimator of the fundamental frequency based on optimal filtering was first proposed in [38]. This was later explored in more detail in [78, 94] and [C]. Moreover, in [195] it was shown how the structure of the model and the involved matrices can be exploited to obtain a fast implementation.

The filter design above can also be modified to obtain an estimate of the entire vector $\mathbf{s}(n) \in \mathbb{R}^M$ using a filter matrix $\mathbf{H} \in \mathbb{R}^{M \times M}$ as

$$\mathbf{y}(n) = \mathbf{H}^H \mathbf{Z}(n) \mathbf{a} + \mathbf{H}^H \mathbf{w}(n). \tag{64}$$

In that case, the output power is given by $\text{Tr}\left\{ \mathbf{H}^H \mathbf{R}_{\mathbf{w}} \mathbf{H} \right\}$. It then follows that for the filter to be distortionless, it must satisfy $\mathbf{H}^H \mathbf{Z}(n) = \mathbf{Z}(n)$ which leads to the following optimization problem:

$$\underset{\mathbf{H}}{\text{minimize}}\, \text{Tr}\left\{ \mathbf{H}^H \mathbf{R}_{\mathbf{w}} \mathbf{H} \right\} \quad \text{s.t.} \quad \mathbf{Z}^H(n) \mathbf{H} = \mathbf{Z}^H(n). \tag{65}$$

The optimal filter matrix is then given by

$$\mathbf{H}^\star = \mathbf{R}_{\mathbf{w}}^{-1} \mathbf{Z}(n) \left(\mathbf{Z}^H(n) \mathbf{R}_{\mathbf{w}}^{-1} \mathbf{Z}(n) \right)^{-1} \mathbf{Z}^H(n). \tag{66}$$

As we can see, we can, given perfect knowledge of the matrix $\mathbf{Z}(n)$, extract a distortionless estimate of the entire speech vector, $\mathbf{s}(n)$ based on the model $\mathbf{Z}(n)$. At this point it should be stressed that in the derivation of this optimal filter, we did not exploit any knowledge of the structure of the model in $\mathbf{Z}(n)$, which means that the principle is quite general and holds for many models.

In speech and audio applications, we, as already mentioned, have the problem that the noise statistics, $\mathbf{R}_{\mathbf{w}}$, are unknown. While various methods have been proposed over the years to estimate these, an alternative is to integrate it into the design of the filer based on the sinusoidal model in (4), which then also applies to the harmonic model in (6). This idea is based on the APES filter, which originated in spectral estimation. Simply put, the idea is that we would like to design the filter $\mathbf{h}$ for the signal model in (4) for $n = 0$ under the constraint that $\mathbf{h}^H \mathbf{Z} = \mathbf{i}_1^T \mathbf{Z}$ where the output of the filter should then resemble a sum of sinusoids. This is achieved by minimizing the

MSE between the output of the filter and a sum of sinusoids having unknown parameters, i.e.,

$$\text{MSE} = \frac{1}{N-M+1} \sum_{n=M-1}^{N-1} \left| y(n) - \sum_{l=1}^{L} a_l e^{j\omega_l n} \right|^2 \tag{67}$$

$$= \frac{1}{N-M+1} \sum_{n=M-1}^{N-1} |\mathbf{h}^H \mathbf{x}(n) - \mathbf{a}^H \boldsymbol{\psi}(n)|^2, \tag{68}$$

with $\boldsymbol{\psi}(n) = \mathbf{Z}^T(n)\mathbf{i}_1 = \left[\begin{array}{ccc} e^{j\omega_1 n} & \cdots & e^{j\omega_L n} \end{array} \right]^T$. Solving for the amplitudes that minimize the MSE yields $\hat{\mathbf{a}} = \boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}_{\mathbf{x}}\mathbf{h}$, where

$$\boldsymbol{\Psi}_{\mathbf{x}} = \frac{1}{N-M+1} \sum_{n=M-1}^{N-1} \boldsymbol{\psi}(n)\mathbf{x}^H(n) \tag{69}$$

$$\boldsymbol{\Omega} = \frac{1}{N-M+1} \sum_{n=M-1}^{N-1} \boldsymbol{\psi}(n)\boldsymbol{\psi}^H(n). \tag{70}$$

Inserting these into (68) along with the amplitude estimates, and observing that $\widehat{\mathbf{R}}_{\mathbf{x}} = 1/(N-M+1)\sum_{n=M-1}^{N-1} \mathbf{x}(n)\mathbf{x}^H(n)$, we obtain the following expression for the MSE:

$$\text{MSE} = \mathbf{h}^H \left(\widehat{\mathbf{R}}_{\text{x}} - \boldsymbol{\Psi}_{\mathbf{x}}^H \boldsymbol{\Omega}^{-1} \boldsymbol{\Psi}_{\text{x}} \right) \mathbf{h}. \tag{71}$$

From this we see that with this approach, we estimate the noise statistics, $\mathbf{R}_{\mathbf{w}}$, implicitly as

$$\widehat{\mathbf{R}}_{\text{w}} = \widehat{\mathbf{R}}_{\mathbf{x}} - \boldsymbol{\Psi}_{\mathbf{x}}^H \boldsymbol{\Omega}^{-1} \boldsymbol{\Psi}_{\mathbf{x}}. \tag{72}$$

An optimal filter can now be obtained by minimizing the MSE in (71) subject to the aforementioned constraint $\mathbf{h}^H\mathbf{Z}(n) = \mathbf{i}_1^T\mathbf{Z}(n)$, i.e.,

$$\underset{\mathbf{h}}{\text{minimize}}\, \mathbf{h}^H \widehat{\mathbf{R}}_{\mathbf{w}} \mathbf{h} \quad \text{s.t.} \quad \mathbf{Z}^H(n)\mathbf{h} = \mathbf{Z}^H(n)\mathbf{i}_1, \tag{73}$$

where $\widehat{\mathbf{R}}_{\mathbf{w}}$ then is the particular estimate of the noise covariance matrix in (72). As before, the solution is given by $\mathbf{h}^\star = \mathbf{R}_{\mathbf{w}}^{-1}\mathbf{Z}(n)(\mathbf{Z}^H(n)\mathbf{R}_{\mathbf{w}}^{-1}\mathbf{Z}(n))^{-1}\mathbf{Z}^H(n)\mathbf{i}_1$. There are a number of interesting special cases of the optimal filters obtained with this approach, as shown in [B]. Approximating $\boldsymbol{\Omega}$ with $\mathbf{I}$, for example, yields the usual noise covariance matrix estimate based on the noise covariance matrix model (11). Capon-like filters can also be obtained by using $\widehat{\mathbf{R}}_{\mathbf{x}}$ instead, which results in the optimal filter $\mathbf{h}^\star = \widehat{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{Z}(n)(\mathbf{Z}^H(n)\widehat{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{Z}(n))^{-1}\mathbf{Z}^H(n)\mathbf{i}_1$. Meanwhile assuming that the noise is white, i.e., $\widehat{\mathbf{R}}_{\mathbf{x}} = \sigma_w^2\mathbf{I}_M$, yields $\mathbf{h}^\star = \mathbf{Z}(n)(\mathbf{Z}^H(n)\mathbf{Z}(n))^{-1}(n)\mathbf{Z}^H(n)\mathbf{i}_1$. Noting that $\lim_{M\to\infty} M\mathbf{Z}(n)(\mathbf{Z}^H(n)\mathbf{Z}(n))^{-1} = \mathbf{Z}(n)$, we get $\mathbf{h}^\star = 1/M\mathbf{Z}(n)\mathbf{Z}^H(n)\mathbf{i}_1$, in which case the filters reduce to simply the Fourier basis. When $n = 0$

the filters are furthers simplified as $\mathbf{Z}^H(0)\mathbf{i}_1 = \mathbf{1}_L$. The filters discussed here also apply to the harmonic model in (6), as we then have that $\omega_l = \omega_0 l$. In that case, the resulting filters are essentially comb filters, a classical type of filter for processing speech and audio signals in many different applications [100, 196–200]. However, unlike their classical precursors, these filters are adaptive, meaning that they adapt to the observed signal and, thereby, the background noise and any interferences, and they are optimal. These filters have many possible uses in in speech and audio processing, as shown in papers [B], [D], and [E]. While it is not surprising that the filter designs considered here apply to the harmonic model, it is perhaps more surprising that, as those papers show, they apply to the harmonic chirp model (16) as well [E].

A concern in these methods is the robustness to model mismatch. It is well-known from the literature that, for example, the Capon beamformer exhibits problems for coherent signals, a situation that occurs frequently in speech and audio signals whenever reverberation is present, and when there are errors in the steering vector [201, 202]. To solve this, a number of robust beamformer designs have been proposed over the years, e.g. [203]. In [204, 205], it was shown that these methods also work well for speech and audio signals and increase the robustness of frequency-domain adaptive beamformers. It remains to be investigated, however, whether similar approaches can be taken in the time-domain described here, where problems may also occur for model mismatch, for example, when the harmonics are not exact integral multiples of the fundamental [206], but the transferal of the aforementioned principles to these methods is non-trivial, not least due to the much higher dimensionality of time-domain filtering problems.

4.3 Subspace-based Optimal Filtering

As we have seen, subspace methods and optimal filtering methods can be used for parameter estimation and both can also be used for noise reduction with early pioneering work on subspace methods for enhancement being [207, 208]. The two methodologies have their origins in different communities and are based on different methodologies. They were for a long time, at least by some, considered unrelated, competing methodologies (see, e.g., the description in [209]), although an early attempt at explaining subspace methods in terms of filtering was given in [210]. The variable span filtering framework introduced in [211] unifies the two methodologies based on ideas dating back at least to [212, 213]. This leads to a framework wherein it is possible to trade off signal distortion for noise reduction, and vice versa, while making use of the knowledge gained via the model in (2). Classical subspace methods for noise reduction rely on models similar to those of the covariance matrix in (11) wherein the noise is assumed to be white. Colored

noise can then be accounted for using pre-whitening or using joint diagonalization [208,214]. The variable span filter framework uses the latter approach. Recall the model of $\mathbf{x}(n) \in \mathbb{R}^M$

$$\mathbf{x}(n) = \mathbf{s}(n) + \mathbf{w}(n) \tag{74}$$
$$= \mathbf{Z}\mathbf{a}(n) + \mathbf{w}(n), \tag{75}$$

where $\mathbf{Z} \in \mathbb{C}^{M \times L}$ and $\mathbf{a}(n) \in \mathbb{C}^L$. Assuming that the signal of interest and the noise are uncorrelated, we have that $\mathbf{R_x} = \mathbf{R_s} + \mathbf{R_w}$. The covariance matrix of the signal of interest, $\mathbf{R_s} = \mathbf{Z}\mathbf{P}\mathbf{Z}^H$, is here assumed to have rank $L \leq M$, which holds, e.g., for the models in (4) and (6) for a distinct set of frequencies and a non-zero fundamental frequency, resptively, while the noise covariance matrix $\mathbf{R_w}$ is assumed to have rank M. The case where the noise covariance matrix is rank deficient is treated in [111]. We now seek to design an optimal filter to estimate $s(n)$. The variable span filters are based on the principle of joint diagonalization, which will now briefly be presented. The covariance matrices $\mathbf{R_s}$ and $\mathbf{R_w}$ can be jointly diagonalized by the full-rank matrix $\mathbf{B} \in \mathbb{C}^{M \times M}$, i.e.,

$$\mathbf{B}^H \mathbf{R_s} \mathbf{B} = \mathbf{\Lambda}, \tag{76}$$
$$\mathbf{B}^H \mathbf{R_w} \mathbf{B} = \mathbf{I}_M, \tag{77}$$

where $\mathbf{\Lambda}$ is a diagonal matrix containing the real, non-negative elements $\{\lambda_i\}_{i=1}^M$ on the diagonal. These are assumed to be sorted in descending order, i.e., as $\lambda_1 \geq \cdots \geq \lambda_L > \lambda_{L+1} = \cdots = \lambda_M = 0$. The matrix $\mathbf{B}$ contains the corresponding eigenvectors $\mathbf{b}_1, \ldots, \mathbf{b}_M$. These can also be seen to be the eigenvectors of $\mathbf{R_w}^{-1}\mathbf{R_s}$, which, in other words, means that we also have that $\mathbf{R_w}^{-1}\mathbf{R_s}\mathbf{B} = \mathbf{B}\mathbf{\Lambda}$. Furthermore, the covariance matrix of the observed signal, $\mathbf{R_x}$, can also be diagonalized by $\mathbf{B}$, i.e., $\mathbf{B}^H \mathbf{R_x} \mathbf{B} = \mathbf{\Lambda} + \mathbf{I}_M$. Let the matrix $\mathbf{B}$ be partitioned as $\mathbf{B} = \begin{bmatrix} \mathbf{B}'_Q & \mathbf{B}''_Q \end{bmatrix}$ where

$$\mathbf{B}'_Q = \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_Q \end{bmatrix} \tag{78}$$
$$\mathbf{B}''_Q = \begin{bmatrix} \mathbf{b}_{Q+1} & \mathbf{b}_{Q+2} & \cdots & \mathbf{b}_M \end{bmatrix}. \tag{79}$$

We remark that when we select $Q = L$, we have that $\mathbf{B}''^H_L \mathbf{s}(n) = \mathbf{0}$. The idea of the variable span filters is to express a filter as a linear combination of the Q first columns of $\mathbf{B}$, i.e.,

$$\mathbf{h} = \mathbf{B}'_Q \mathbf{a}'_Q, \tag{80}$$

where $\mathbf{a}'_Q \in \mathbb{C}^Q$ then can be found in different ways to yield different solutions in terms of output SNR and signal distortion. More specifically, the variable span linear filters are of the following form:

$$\mathbf{h} = \sum_{q=1}^{Q} \frac{\mathbf{b}_q \mathbf{b}_q^H}{\mu + \lambda_q} \mathbf{R_s} \mathbf{i}_1, \tag{81}$$

where $\mu \geq 0$ is a Lagrange multiplier, often treated as a user parameter to be chosen by itself, that controls the tradeoff between noise reduction and signal distortion. Interestingly, the variable span filters can be related to the model-based filter designs considered in Subsection 4.2. Observe that for the sinusoidal and harmonic models in (4) and (6), respectively, the variable span filters can also be expressed as

$$\mathbf{h} = \sum_{q=1}^{Q} \frac{\mathbf{b}_q \mathbf{b}_q^H}{\mu + \lambda_q} \mathbf{Z}\mathbf{P}\mathbf{Z}\mathbf{i}_1, \tag{82}$$

by using the covariance matrix model (11). Recall that distortionless filter for the signal model in (75) is given by (61) for which we consider the case where we wish to estimate the first sample $s(n)$ of the vector $\mathbf{s}(n)$. The inverse noise covariance matrix, $\mathbf{R}_\mathbf{w}^{-1}$, can be expressed using $\mathbf{B}$ as $\mathbf{R}_\mathbf{w}^{-1} = \mathbf{B}\mathbf{B}^H$. In that case, the model-based optimal filter in (61) can be written as

$$\mathbf{h} = \mathbf{B}\mathbf{B}^H\mathbf{Z}\left(\mathbf{Z}^H\mathbf{B}\mathbf{B}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{i}_1. \tag{83}$$

Next, observe that this optimal filter can be expressed in terms of $\mathbf{B}'_Q$ with $Q = L$ due to $\mathbf{B}_L''^H\mathbf{Z} = \mathbf{0}$. Moreover, since the matrix $\mathbf{Z}^H\mathbf{B}'_L$ is both square and invertible, the above can be rewritten as

$$\mathbf{B}\mathbf{B}^H\mathbf{Z}\left(\mathbf{Z}^H\mathbf{B}\mathbf{B}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{i}_1 = \mathbf{B}'_L\mathbf{B}'^H_L\mathbf{Z}\left(\mathbf{B}'^H_L\mathbf{Z}\right)^{-1}\left(\mathbf{Z}^H\mathbf{B}'_L\right)^{-1}\mathbf{Z}^H\mathbf{i}_1 \tag{84}$$

$$= \mathbf{B}'_L\left(\mathbf{Z}^H\mathbf{B}'_L\right)^{-1}\mathbf{Z}^H\mathbf{i}_1. \tag{85}$$

We can simplify this further by using the covariance matrix model, i.e., $\mathbf{R}_\mathbf{s} = \mathbf{Z}\mathbf{P}\mathbf{Z}^H$, due to which we have that $\mathbf{B}'^H_L\mathbf{Z}\mathbf{P}\mathbf{Z}^H\mathbf{B}'_L = \mathbf{\Lambda}'$ where $\mathbf{\Lambda}' = \operatorname{diag}\left([\lambda_1 \ \cdots \ \lambda_L]\right)$ contains the L nonzero eigenvalues in $\mathbf{\Lambda}$. From this we can conclude that $\left(\mathbf{Z}^H\mathbf{B}'_L\right)^{-1} = \mathbf{\Lambda}'^{-1}\mathbf{B}'^H_L\mathbf{Z}\mathbf{P}$ and thus

$$\mathbf{B}'_L\left(\mathbf{Z}^H\mathbf{B}'_L\right)^{-1}\mathbf{Z}^H\mathbf{i}_1 = \mathbf{B}'_L\mathbf{\Lambda}'^{-1}\mathbf{B}'^H_L\mathbf{Z}\mathbf{P}\mathbf{Z}^H\mathbf{i}_1 \tag{86}$$

$$= \sum_{q=1}^{L} \frac{\mathbf{b}_q\mathbf{b}_q^H}{\lambda_q}\mathbf{Z}\mathbf{P}\mathbf{Z}\mathbf{i}_1, \tag{87}$$

which demonstrates that the filter design in (61) is equivalent to that of (82) for $Q = L$ and $\mu = 0$, a filter which is sometimes referred to as the MVDR filter. Moreover, the aforementioned equivalence between using $\mathbf{R}_\mathbf{x}$ and $\mathbf{R}_\mathbf{w}$ in (61) can also easily be proven in a similar fashion, only $\mathbf{B}\left(\mathbf{\Lambda} + \mathbf{I}_M\right)^{-1}\mathbf{B}^H$ should be inserted into (61) instead of $\mathbf{B}\mathbf{B}^H$. As has been shown, a clear connection exists between the model-based and the subspace-based optimal filtering approaches. However, while the variable span linear filters require

that the joint diagonalization of $\mathbf{R_s}$ and $\mathbf{R_w}$ be computed, which in turn requires that either of those matrices be estimated somehow, the model-based optimal filtering only requires that the parameters characterizing $\mathbf{Z}$, which for the case of the harmonic model (6) is only the fundamental frequency, be found.

Besides from this case, which is of particular importance to this thesis, a number of other classical filter designs are special cases of the variable span filters as shown in [211, 215]: for example, with $Q = M$ and $\mu = 1$ we obtain the Wiener filter, while for $Q = 1$ we obtain the maximum SNR filter. For $Q \leq L$ and $\mu = 0$ we obtain minimum distortion filters, and for $Q \geq L$ and $\mu \geq 0$ we get tradeoff filters. As can be seen, the variable span filters thus not only contain a number of classical filter designs as special cases, but a continuum of solutions that have different properties in terms of signal distortion and noise reduction capabilities. The variable span filters can be applied equally well to spatial filtering problems and in the frequency domain [215] (see also [216]). More importantly, however, it is possible to bound and relate the performance of the different filter designs in the variable span filter framework in terms of output SNR and signal distortion. For example, the maximum SNR filter achieves the highest possible output SNR but also incurs the most distortion on the signal of interest. It should, though, be remarked that there are multiple ways to arrive at these filter. For example, in [217] similar filters are obtained via a low-rank approximation of the covariance matrix of the observed signal.

4.4 Multi-Channel Model-based Filtering

Much of the progress within speech and audio processing in recent years, both in terms of research and the development of new technology, has been achieved via the use of microphone arrays, which with the availability of cheap, small microphones, such as MEMS microphones, can now been found in many devices. It is interesting to note that in array signal processing for speech and audio signals, propagation models are commonplace but signal models, such as those considered in this thesis, are rare (see, e.g., [13]). Indeed, signals are most often characterized simply via second-order statistics [171,182,218–222], and often only spatial statistics are used [202,204,223,224]. As we shall see, however, using signal models not only makes it possible to exploit knowledge about the signal of interest in, e.g., localization or beamforming, the models considered herein are actually well-suited for this. There are many ways in which these models can be used in filtering in a multi-channel setting either for the purposes of reducing noise or separating sources or for estimating parameters. We will now present some of these. Let subscript $k = 1, \ldots, K$ denote the kth channel, and consider the following

multi-channel model:

$$x_k(n) = h_k(n) * s(n) + w_k(n). \tag{88}$$

Here, $s(n)$ is the signal of interest, $h_k(n)$ is the impulse response describing the propagation of the signal $s(n)$ from its origin to microphone k, and $w_k(n)$ is channel-specific noise. A simple yet effective model that can account for several effects is the following, which is based on (8) with $\mathbf{x}_k(n) \in \mathbb{C}^M$:

$$\mathbf{x}_k(n) = \mathbf{Z}\mathbf{D}(n - \tau_k)\mathbf{a}_k + \mathbf{w}_k(n). \tag{89}$$

In this model, the effect of the impulse response, $h_k(n)$, on the signal of interest, $s(n)$ is modeled as a channel-dependent delay, τ_k, that explains the propagation time from the source to microphone k, and channel-dependent complex amplitudes, $\mathbf{a}_k$, which account for the filtering effect of the air on the signal of interest. Recall that for the sinusoidal model the matrix $\mathbf{D}(n)$ is given by $\mathbf{D}(n) = \text{diag}\left(\begin{bmatrix} e^{j\omega_1 n} & e^{j\omega_2 n} & \cdots e^{j\omega_L n} \end{bmatrix}\right)$ and thus

$$\mathbf{D}(n - \tau_k) = \begin{bmatrix} e^{j\omega_1(n-\tau_k)} & 0 & \cdots & 0 \\ 0 & e^{j\omega_2(n-\tau_k)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{j\omega_L(n-\tau_k)} \end{bmatrix}. \tag{90}$$

The model in (89) is accurate for large N and when $s(n)$ is stationary over the N samples, as the effect of the convolution in (88) is then simply that the complex amplitudes of the model are changed for each channel. In this model, the channel-dependency is then explained purely in the matrix $\mathbf{D}(n)$ and the complex amplitudes, $\mathbf{a}_k$, meaning that the nonlinear part of the model, $\mathbf{Z}$, is shared across channels. Since the noise in the model can be different for different channels, it can easily account for different SNRs, etc. For the harmonic model in the form of the model in (89), it is possible to derive the MLE to estimate the ω_0 jointly across all channels, while estimating the delay, τ_k, and complex amplitudes, $\mathbf{a}_k$ for each channel k. It is also easy to constrain the model to share complex amplitudes across channels, i.e., $\mathbf{a}_k = \mathbf{a} \forall k$, and to incorporate different array structures in the delays such that these are determined by a single DOA. Moreover, as shown in [225], it is also possible to incorporate near-field models into the model and to estimate its parameters and to account for reverberation, to some extent [226]. The principles can also be used to design a directional binaural fundamental frequency estimator for hearing aids that estimates the fundamental frequency of a source coming from the nose direction of the user [177].

Next, we will exemplify what can achieved using the model-based approach based on [G], wherein the problem of joint fundamental frequency and localization is considered and its potential explored. In other words, we

will now consider the case where $\omega_l = \omega_0 l$. A special case of the model in (89) is when the signal of interest is purely delayed across channels, i.e., $x_k(n) = s(n - \tau_k) + w_k(n)$, as is the case in far-field conditions and anechoic environments. Moreover, for a uniform linear array with microphones $k = 1, \ldots, K$ placed Δ apart, the delays between microphone k and 1 of the observed signal (when microphone 1 is used as reference) is given by $\tau_k = \frac{\Delta \sin \varphi}{c}(k-1)$ with φ being the DOA and c the propagation speed. From this, we can define the spatial frequency as $\omega_\mathrm{s} = \omega_0 f_\mathrm{s} \frac{\Delta \sin \varphi}{c}$ with f_s being the sampling frequency. The observed signal for channel k can now be expressed as $x_k(n) = \sum_{l=1}^{L} a_l e^{j\omega_0 l n} e^{-j\omega_\mathrm{s} l(k-1)} + w_k(n)$. The observed signal collected in snapshots of M samples and aggregated across channels can be organized in a matrix $\mathbf{X}(n) \in \mathbb{C}^{K \times M}$ defined as

$$\mathbf{X}(n) = \begin{bmatrix} x_1(n) & \cdots & x_1(n-M+1) \\ \vdots & \ddots & \vdots \\ x_K(n) & \cdots & x_K(n-M+1) \end{bmatrix}. \tag{91}$$

Defining $\mathbf{i}_k$ as the kth column of $\mathbf{I}_K$, the observed signal in the matrix in (91) can be related to the model in (89) as $\mathbf{i}_k^T \mathbf{X}(n) = \mathbf{x}_k^T(n)$ with $\mathbf{a}_k = \mathbf{a} \forall k$, i.e.,

$$\mathbf{X}^T(n)\mathbf{i}_k = \mathbf{Z}\mathbf{D}(n - \tau_k)\mathbf{a} + \mathbf{w}_k(n). \tag{92}$$

Introducing $\gamma_l(n) = a_l e^{j\omega_0 l n}$, which can be identified as the individual elements of the vector resulting from the matrix-vector product $\mathbf{D}(n)\mathbf{a}$, the matrix in (91) can be modeled as

$$\mathbf{X}(n) = \sum_{l=1}^{L} \gamma_l(n) \mathbf{z}_\mathrm{s}(\omega_\mathrm{s} l) \mathbf{z}_\mathrm{t}^T(\omega_0 l) + \mathbf{W}(n), \tag{93}$$

where the noise term $\mathbf{W}(n) \in \mathbb{C}^{K \times M}$ is defined similarly as $\mathbf{X}(n)$ in (91) and $\mathbf{z}_\mathrm{t}(\omega_0 l)$ and $\mathbf{z}_\mathrm{s}(\omega_\mathrm{s} l)$ are temporal and spatial model vectors defined as

$$\mathbf{z}_\mathrm{t}(\omega_0 l) = \begin{bmatrix} 1 & e^{-j\omega_0 l} & \cdots & e^{-j\omega_0 l(M-1)} \end{bmatrix}^T \tag{94}$$

and

$$\mathbf{z}_\mathrm{s}(\omega_\mathrm{s} l) = \begin{bmatrix} 1 & e^{-j\omega_\mathrm{s} l} & \cdots & e^{-j\omega_\mathrm{s} l(P-1)} \end{bmatrix}^T, \tag{95}$$

respectively. Defining $\bar{\mathbf{x}}(n) = \mathrm{vec}\{\mathbf{X}(n)\}$ where $\mathrm{vec}\{\cdot\}$ is the vectorization operator that stacks the columns of the argument, and similarly for $\bar{\mathbf{w}}(n)$, the model can be written as

$$\bar{\mathbf{x}}(n) = \sum_{l=1}^{L} \gamma_l(n) \bar{\mathbf{z}}_l + \bar{\mathbf{w}}(n), \tag{96}$$

where $\bar{\mathbf{z}}_l$ is the vectorized version of the spatio-temporal model, i.e.,

$$\bar{\mathbf{z}}_l = \text{vec}\{\mathbf{z}_s(\omega_s l)\mathbf{z}_t^T(\omega_0 l)\} \tag{97}$$

$$= \mathbf{z}_s(\omega_s l) \otimes \mathbf{z}_t(\omega_0 l), \tag{98}$$

which, as can be seen, can also be expressed using the Kronecker product, $\otimes$. Next, consider the problem of designing an optimal vectorized spatio-temporal filter $\bar{\mathbf{h}} \in \mathbb{C}^{KM}$ to be applied to $\bar{\mathbf{x}}(n)$. From the optimal filter designs considered earlier, we see that by imposing constraints on an optimal filter, we can extract the signal of interest in a distortionless manner while attenuating noise and interference as much as possible. To do this for the multi-channel model, we observe from the model in (96) that if we impose the constraint $\bar{\mathbf{h}}^H \bar{\mathbf{z}}_l = 1 \forall l$, the harmonics having fundamental frequency ω_0 impinging on the array from an angle of φ will pass undistorted as $\bar{\mathbf{h}}^H \left(\sum_{l=1}^{L} \gamma_l(n) \bar{\mathbf{z}}_l \right) = s(n)$. Define $\bar{\mathbf{R}}_{\mathbf{x}}$ as the covariance matrix of $\bar{\mathbf{x}}(n)$, then associated optimal filter design problem can be cast as the following:

$$\underset{\bar{\mathbf{h}}}{\text{minimize}}\, \bar{\mathbf{h}}^H \bar{\mathbf{R}}_{\mathbf{x}} \bar{\mathbf{h}} \quad \text{s.t.} \quad \bar{\mathbf{z}}_l^H(n) \bar{\mathbf{h}} = 1 \tag{99}$$

$$\text{for} \quad l = 1, \ldots, L. \tag{100}$$

Defining the matrix $\bar{\mathbf{Z}} = \begin{bmatrix} \bar{\mathbf{z}}_1 & \cdots & \bar{\mathbf{z}}_L \end{bmatrix}$, comprising the vectorized harmonic components, the constraint can be expressed compactly as $\bar{\mathbf{Z}}^H \bar{\mathbf{h}} = \mathbf{1}_L$, and the solution to the above optimization problem is given by:

$$\bar{\mathbf{h}}^\star = \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \bar{\mathbf{Z}} \left(\mathbf{Z}^H \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \bar{\mathbf{Z}} \right)^{-1} \mathbf{1}_L. \tag{101}$$

This requires that the inverse of the matrix $\bar{\mathbf{R}}_{\mathbf{x}}$ exists, which is satisfied for $N + 1 \geq M(K + 1)$ when the sample covariance matrix estimate is computed from the vectorized observed signal $\bar{\mathbf{x}}_k(n)$ [G]. The filter in (101), which operates jointly in time and space, can then be used for joint estimation of the fundamental frequency, ω_0, and the angle, φ, by applying the optimal filter to the observed signal and then maximize the output power. This yields the following estimator:

$$\{\hat{\omega}_0, \hat{\varphi}\} = \arg\max_{\omega_0, \varphi} \mathbf{1}_L^T \left(\mathbf{Z}^H \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \bar{\mathbf{Z}} \right)^T \mathbf{1}_L. \tag{102}$$

As shown in [G], joint estimation of the fundamental frequency and angle lead to an estimator that is more robust than estimating the parameter sequentially, as is often done. Moreover, estimators built on this principle are typically also robust to noise and interference. Other works that have explored similar ideas, but in different ways and contexts, include [227–231] which involve joint pitch estimation and localization while [232] explores the

related idea of performing localization while modeling the signal of interest using an auto-regressive model. General multi-channel pitch estimators that do not exploit array geometry are a rarity, and only a few examples can be found in the literature [35, 233, 234].

The optimal filter in (101) can also be used for beamforming, in which case the covariance matrix of the observed signal, $\bar{\mathbf{R}}_{\mathbf{x}}$, should be replaced by the covariance matrix of the vectorized noise, i.e., $\bar{\mathbf{R}}_{\mathbf{w}}$. Various simplifications and special cases of the optimal filter design can also be obtained, similarly to the single-channel case. For example, certain noise distributions can be assumed, or asymptotic approximations can be used to simplify the filters. Interestingly, it is also possible to derive a filterbank version of the optimal filter and to derive a filter that outputs a vector.

5 Conclusion

In what follows, the specific contributions of the individual papers included in this thesis will be stated after which they will be discussed and directions for future research will be given.

5.1 Contributions

[A] The paper considers the problem of determining the number of sinusoidal components from a noisy signal using a subspace method that exploits the orthogonality between the signal and noise subspaces. In [113], it was first proposed to do this by scaling a subspace orthogonality criterion based on the Frobenius norm to account for the unknown and thus varying number of dimensions of the involved subspaces. This was introduced based on the Cauchy-Schwarz inequality and while the notion of estimating the model order based on subspace orthogonality is sound, the criterion itself lacked a proper mathematical foundation. This paper formalizes these ideas in the context of the sinusoidal model, a problem also considered in [235]. It does so by connecting the underlying problem of measuring orthogonality of subspaces having variable dimensions (due to the unknown number of sinusoids) to the concept of angles between subspaces. Defining such angles in multi-dimensional spaces is a non-trivial problem, but based on this connection, a new criterion is proposed. This new criterion is compared to other methods for model and order selection, including well-known statistical methods and other subspace-based criteria, and the results clearly demonstrate that the principle works well. Interestingly, it is typically simpler to apply this new principle to nonlinear models than criteria such as the BIC, which require complicated statistical analysis be performed to de-

termine the penalty. This principle was adopted for joint fundamental frequency and order estimation in [78].

[B] This paper explores the ideas behind the optimal filtering method originally proposed for fundamental frequency estimation in [38] further and explores their novel application to enhancement and separation of periodic signals, problems often encountered in speech and audio processing, building on the preliminary results first presented in [236]. This results in a set of filter designs that can be thought of as adaptive, optimal finite impulse response comb filters with a wide range of applications. A number of different filter designs are proposed and analyzed, including filterbank formulations and filters that implicitly estimate the noise statistics in the computation of the optimal filters. Moreover, a number of simplifications and special cases are investigated, and it is shown that the filters reduce to well-known filters and concepts in particular special cases. Experiments demonstrate how these filters can be applied to speech and audio signals and investigate their performance when enhancing and separating such signals. Particularly, the ability of these filters to attenuate interference without explicit knowledge about their presence is remarkable.

[C] Most of the fundamental frequency estimators encountered in the literature employ asymptotic approximations, very often implicitly. When the fundamental frequency is not close to zero and when the number of samples is high, these approximations are accurate and good estimates can be obtained, even in the presence of noise. The harmonic summation method is an example of this. However, when this is not the case, such estimators may fail completely, as they ignore the interaction between the harmonics. This paper thus considers the problem of estimating low fundamental frequencies from noisy observations. The problem is further complicated by the real-valued nature of audio signals which creates interactions between the positive and negative sides of the spectrum, something that is not normally an issue. The issues surrounding this are poorly understood and have not been investigated in detail in the literature. The paper analyzes the problem in depth using exact CRLBs computed numerically and a collection of exact estimators, expanding on the preliminary work [237, 238], that do not make use of the aforementioned asymptotic approximations are derived and tested in simulations. The results show that it is possible to achieve better performance with the exact estimators with the nonlinear least squares method clearly performing the best.

[D] An often heard criticism of the harmonic model is that it assumes that the signal is stationary over the segments that are being analyzes or pro-

cessed, as it is well-known that voiced speech is continuously changing. To investigate whether this is important, the harmonic chirp model was proposed in [56]. In this paper, this model and the associated estimator are explored further and its use for analysis of voiced speech is investigated. Moreover, optimal segmentation in combination with a MAP criterion i used to find the combination of the optimal model and segmentation of the observed signal. To account for the background noise rarely being white in speech and audio signals, adaptive pre-whitening based on an auto-regressive noise model, which serves to smooth the estimated noise spectrum, is used. The results demonstrate that the harmonic chirp model is a better model of voiced speech, particularly for longer segments, than the harmonic model, and that the model-based fundamental frequency estimators achieve competitive performance in terms of traditionally used metrics on real speech data while having the additional benefits of robustness and statistically optimal continuous pitch estimates.

[E] This paper continues the work of papers [B], [D], and the preliminary results reported in [239], by considering model-based enhancement for non-stationary periodic signals, such as voiced speech. It is shown in the paper that it possible to account for the non-stationary nature of speech in the design of optimal filter for noise reduction based on the harmonic chirp model. Considering that most existing methods assume that speech and noise are stationary within segments, this is a notable feature. Based on the harmonic chirp model, optimal filters that let the signal of interest pass undistorted while attenuating the background noise optimally by maximizing the output SNR. Moreover, it is also demonstrated that in the process, the principle behind the APES filters can be used to estimate the noise statistics required by various optimal filtering methods, such as the Wiener filter. Hence, a number of problems in speech enhancement are solved simultaneously this way.

[F] Linear prediction, which is closely related to the auto-regressive model, is a ubiquitous tools in speech processing. Traditional linear prediction is performed by minimizing the 2-norm of the prediction error. Its properties are well-understood and its problems when applied to speech signals have been well-known since the 70s. Inspired by advances in convex optimization and compressed sensing, sparse linear prediction is proposed and its applications to speech processing explored, building on the preliminary work first reported in [240–247]. Since voiced speech can be modeled (crudely) as a pulse-train passed through an all-pole filter, it is better to use the 1-norm than the 2-norm of traditional linear prediction when analyzing and processing voiced speech, as this renders the residual sparse. This way, several of the well-known problems of

traditional linear prediction, such as the estimates' dependency on time-shifts and pitch, are solved. Moreover, by using high-order but sparse predictors, the combined effects of long- and short-term predictors can be captured. The paper considers different algorithms for implementing the idea of sparse linear prediction and explores several possible speech applications.

[G] In this paper, the idea of model-based microphone array processing is explored, based on the preliminary ideas and results presented in [248, 249]. More specifically, it is proposed to incorporate the harmonic model in an optimal spatio-temporal filtering framework which is then used for localization by jointly estimating the fundamental frequency and the direction-of-arrival of speech signals impinging on a uniform array. Hereby, the piece-wise smooth evolution of the fundamental frequency of voiced speech along with movements in space can be exploited for better localization of speakers in complex acoustic environments. Similarly to its single-channel counterparts explored in, e.g., [B], different filter designs can be derived and a number of interesting special cases and simplifications exist. As part of the work, a recently proposed way of estimating signal statistics is also investigated. The good performance and robust nature of the proposed method is demonstrated in the experiments.

5.2 Discussion

As highlighted in the previous section, the papers in this thesis make a number of specific contributions to signal modeling, parameter estimation, and model-based filtering within speech and audio processing. These contributions each provide a small piece to a bigger puzzle. The papers have made significant advances to model-based fundamental frequency estimation by introducing and analyzing a number of new methods for solving this problem, but also by proposing new ways of solving some of the associated problems, problems that occur in many different contexts in speech and audio processing. These include the problem of order selection, which is a general problem in science and engineering. The papers also introduce different ways in which the presence of noise, and, in particular, colored noise can be handled and its statistics estimated. The papers explore how other interfering, periodic sources can be dealt with via optimal filtering for both the single-channel and multi-channel cases. The multi-channel models and methods also demonstrate the potential of the model-based approach. It is also demonstrated how more advanced models, such as the harmonic chirp model, can be be used to help answer questions whether the harmonic model, for example, is robust to non-stationarity. Similarly, the exact estimators developed for estimating low fundamental frequencies help answer the ques-

tion of how accurate the commonly used asymptotic approximations are and whether they are already sufficient. Sparse linear prediction demonstrates how the inherent problems associated with the ubiquitous linear prediction based on 2-norm minimization can be solved, not via heuristics, but by examining the underlying statistical assumptions and by casting the problem as a statistically meaningful convex optimization problem. More recent work has shown how this can be taken even further using a fully Bayesian framework [62, 63].

Seen as a whole, a number of important and more general lessons can also be learned from the papers included in this thesis. The thesis demonstrates that signal models can be used for solving quite a number of different important problems in speech and audio processing, such as fundamental frequency estimation, localization, noise statistics estimation, separation, enhancement, and beamforming, and that there are inherent advantages of doing so, such as treating the fundamental frequency as the continuous parameter it is and obtaining methods that are robust by explicitly accounting for noise. These usages go beyond the typical uses of such models, which have in the past mostly been limited to signal modification and analysis, although the harmonic model was, at one point, also considered for both speech and audio coding (e.g., [16]) but it never gained widespread use. The thesis also demonstrates that the problems associated with the use of these models can, quite often, be solved, one by one. For example, the papers demonstrate that the presence of colored noise can be handled in a number of ways, for example by explicitly taking its presence into account in the estimation process or by pre-processing. Similarly, the often heard criticism of the models not taking the non-stationarity of speech into account can be addressed, as demonstrated by the harmonic chirp model. Moreover, it is only because of the methodological approach to solving the estimation and modeling problems that the weaknesses of the obtained methods are revealed. For example, the absence of explicit assumptions of non-parametric methods may seem like an advantage at first sight, but it is really a disadvantage down the road, as there is no systematic approach to resolving any issues the methods may have, which are then only also discovered experimentally. With the model-based approach, for example, it is quite clear that model and order selection are inherent parts of the solution to many problems, as multiple models are considered in the process of solving an estimation or modeling problem while is less obvious in other approaches. In noise trackers [12, 105, 174], for example, a hypothesis test selecting between noise and speech-plus-noise models is typically performed, and the compared models are of different complexity, which causes problems, and similar observations hold for many source separation methods. The papers in this thesis also show that there are several ways in which signals models can be used in speech and audio processing. It is well-known that the harmonic model can be used for

fundamental frequency estimation [101], although questions concerning the robustness of methods based on this model have remained unanswered until recently [104], but it is less obvious that the models can be used directly in filtering for noise reduction and beamforming, including also the non-stationary harmonic chirp model. It is also important to note that another common criticism of the models considered here and their estimators is that, due to their nonlinear nature, they lead to nonlinear, complicated estimators that are so much more complex than their non-parametric counterparts that it defeats any advantage they may have in accuracy or robustness. However, as shown in [58,103,195] it is possible to derive fast estimators for these models, and, in the case of the method of [103], the fast implementation is of a complexity comparable to harmonic summation while the resulting method is much more robust than non-parametric methods. There thus appears to be very little reason to not use these methods anymore, as compared to the alternatives.

The models discussed in this thesis are mostly Gaussian, and it is sometimes argued that the scientific community should turn its attention to non-Gaussian models and nonlinear signal processing. While it is of course true these would be more general and can describe more complex phenomena and do things that linear signal processing cannot, such models and methods are also much more difficult to deal with, meaning that there is a tractability to the models considered herein that may be lost for more complex models. More specifically, closed-form solutions may not exist, and it may be very difficult to analyze and understand the behaviour of the algorithms, and thus it can also be very difficult to improve them, other than by trial-and-error. Deep learning is an example of this. Many interesting and promising results and applications within speech and audio, and in other fields too, have been reported recently (e.g., [250–253]), but, as pointed out in [254,255], that type of research suffers from some issues that we are now also seeing in signal processing research. The incorporation of more structure (i.e., models) in deep learning is a possible solution to at least some of these issues [256]. More complex models also generally lead to poorer estimates [87], so as simple models as possible should be preferred. Concerning the usage of the Gaussian distribution, it should also be stressed that it is actually the worst case distribution in the sense that it leads to the largest CRLB [257], and it is the distribution that maximizes the entropy. It should also be noted, that in the models considered herein, the part that is not Gaussian is explained by the possibly nonlinear model, which is thought of as being deterministic. However, if the distribution of the parameters that generates this model is taken into account, then the results would in many cases be a non-Gaussian model. Moreover, as sparse Bayesian learning shows, the Gaussian distribution can be utilized to solve complicated problems in clever ways by looking at the problems differently [160]. Interestingly, NMF methods for speech

and audio processing are sometimes argued to be non-Gaussian by nature, but Gaussian signals are parametrized by additive power spectra, on which NMF are typically based, and the NMF model is thus closely related to the Gaussian assumption. Furthermore, the Kullbach-Leibler divergence reduces to the often-used Itakura-Saito distance for infinitely long, Gaussian signals, which means that the usage of spectral measures, which are common for NMF methods, is well-founded for Gaussian signals, but not necessarily for others (see [84]). Concerning the relation between the models and methods considered in this thesis and NMF, it is possible to use a parametrization of the involved spectra using, for example, an auto-regressive model [258]. In that case, the NMF model corresponds to a stochastic model comprising a sum of auto-regressive processes [259]. The NMF-based methods, which in speech and audio originated in source separation, have spread to other related problems over the past decade, including also some of the problems considered in this thesis [260, 261], which has led to both new methods and valuable new insights. This can be seen as part of a larger trend, where a convergence of source separation, signal enhancement, and array processing is going on, as argued in [13], and this is likely to continue in the future.

5.3 Future Research

The present thesis has made a number of contributions to making parameter estimation in speech and audio processing robust to the presence of additive, background noise, and to the filtering of such signals with the objective of suppressing said noise. However, additive noise is not the only kind of signal degradation that speech and audio signals are subjected to. Except for situations when the microphone is very close to the source, as is the case in many headsets, the recorded signals to be analysed and processed will suffer from reverberation. For the problems considered here, very few estimators have been derived that explicitly take the presence of noise and reverberation into account, and this is a major source of error, and future search should be conducted to address this problem. Dereverberation, i.e., the problem of enhancing signals that have been subjected to reverberation, has been an open problem for a long time (see, e.g., [262]), and much research has been devoted to it in recent years [20, 156, 171, 218, 220, 221, 263]. More recently, the combination of additive and convolutive noise has also been considered in so-called comprehensive speech enhancement [171, 182, 222], two problems that should be addressed simultaneously to obtain optimal solutions. Very little work has, though, been devoted to taking revereberation into account or studying its effect in parameter estimation in speech and audio, however, with a few notable exceptions [13,264], and future research should be devoted to finding principled ways of taking unknown reverberation into account in parameter estimation, for example in fundamental frequency estimators, to

make such estimator more robust. One possibility for doing this is to use the principle of [218, 219] wherein reverberation is taken into account by splitting the impulse responses into early, sparse parts and late, stochastic parts that can be treated as noise. Aside from reverberation, the models and estimators discussed in this thesis also do not take the presence of nonlinear degradations into account. Such nonlinearities, however, occur frequently in many situations. For example, signals are often clipped during recording, distorted by loudspeakers during playback, or compressed by dynamic range control. There exists methods for dealing with very specific cases of such nonlinearities in isolation and in specific contexts [30, 265–267], but no general approach exists. It is interesting to note that in control theory and system identification, methods exist for taking the presence of nonlinearities into account [268, 269], and it is possible that a similar methodology can be adopted in speech and audio processing. Furthermore, taking the ideas of comprehensive enhancement a step further, the presence of nonlinear signal degradations aside from background noise and reverberation would be desirable, even if very difficult. As it stands, no method capable of handling all these phenomena simultaneously currently exist. That these problems can be handled is particularly important in, for example, forensics and for studying historically important recordings [270].

Another interesting transformation in speech and audio processing, is the development from acoustic arrays having fixed, predetermined geometries to wireless acoustic sensor networks (WASNs) comprising a number of independent devices in an acoustic environment [271]. Such networks have the potential to form the foundation for the solutions to a number of long-standing problems of both scientific and practical interest in speech and audio processing, such as the cocktail party problem. While this is not a new idea, it is not until recently that it has become practical. The traditional methodology in beamforming, localization, etc., is, however, rooted in the telephony system where all parts of the system were known and well-understood, but for WASNs this is no longer the case. Moreover, WASNs introduce a number of problems that are not present in traditional arrays, such as synchronization, widely different SNRs, different and unknown microphone responses, etc. In other words, to realize the potential of such WASN a number of scientific problems have to be solved, and this requires a more general way of looking at problems such as parameter estimation and signal enhancement. For example, rather than using a beamformer that exploits the array geometry followed by a post-filter, a more general approach, such as a multi-channel Wiener filter, could be taken [170, 272], wherein the problem then is to estimate the required statistics. In relation to this, much research has, as has already been discussed, been devoted to noise trackers of the past couple of decades, starting with [106], and those have enjoyed much success in single-channel enhancement, but the generalization of these

principles to WASNs is highly non-trivial. Moreover, they have problems handling non-stationary noise and multiple speakers where more sophisticated methods are needed [18,177,258]. In the presence of multiple speakers as well as noise sources, even defining the signal of interest, and thus the problem to be solved, is difficult. The model-based approach has the potential to be a possible solution to these problems, as demonstrated in [179,273], as models of the signal of interest may be the only reliable information that can be exploited, as the geometry of the array is unknown, far-field models do not apply, etc. Combined with distributed signal processing for solving the involved optimization problems, such models could thus form the foundation for signal processing in WASNs. Moreover, generalizing this further and combining this notion with the aforementioned comprehensive enhancement, an even more general way of looking at these problems would be as a collection of independent recordings, each recorded by different devices with possibly different bandwidths, sampling frequencies, jitter, and SNRs in different places, suffering from different types and amounts of degradations (background noise, reverberation, and distortions), that combine to form an ensemble of recordings of an acoustic phenomenon.

In terms of methodology, the future of signal processing is, in this author's opinion, Bayesian, and the continued development of models for speech and audio signals that are amenable to efficient computation or have closed-form solutions to integrals, are thus important. Similarly, it can be learned from the dereverberation and array processing literature that, as previously mentioned, there are many ways of looking at a problem, and once the right way is found, much scientific progress can be gained. The adoption of relative transfer functions [274] in localization, beamforming, and dereverberation is an example of this, and so is sparse linear prediction. Hence, the continued search for good models of speech and audio signals and good ways of posing the estimation problems is extremely important. Furthermore, the potential of the Bayesian methodology can also be seen from the advances in sparse approximations, which was predominantly based on deterministic convex optimization but is now based on probabilistic methodology [160]. In a sense, the ideas of sparse approximations have enabled us to linearize the nonlinear models of localization, frequency estimation, and many other problems via the dictionaries, which are then highly structured, something that can be exploited both for interpretability but also in achieving computationally efficient implementations. Exploiting such structure is thus extremely important in casting the involved learning and estimation problems in ways that have as few degrees of freedom as possible, examples of which are shift-invariance [275] and the Kronecker decomposition [276]. Moreover, the Bayesian methodology also allows for combining different models and using prior knowledge in hierarchies, exploiting spatio-spectral patterns, for example, something that is otherwise difficult to do in a consistent manner.

Aside from the aforementioned reasons for finding good ways of posing parameter estimation and linear filtering problems, it is also important that the search for good models continues in that better models lead to better results in terms of better characterizing the involved signals, something that is important in many contexts (e.g., in biomedical applications), but also that models and methods that allow for model mismatch are found, an example of which is [204]. Similarly, it is equally important that the importance and effect of model mismatch be understood. Important new insights concerning this has been recently reported in [277] based on the concept of the model misspecification and the miss-specified CRLB [278], which enables the analysis of the effects of using the wrong model, such as the harmonic model for a case where inharmonicity is known to be present. There is another aspect of the models that has not been mentioned so far, namely that of the spaces of the parameters of the models (or priors as one might call them). In this thesis, those parameter spaces have not been modeled in a data-driven way or specified analytically or probabilistically in much detail. NMF-based methods for speech enhancement and source separation basically work by exploiting that different sources lie in different parts of the parameter space, which is also how the method of [258] works[5]. However, the methods for first training and later manipulating parameters of different models in these spaces is comparably underdeveloped in speech and audio. For example, methods based on the auto-regressive model basically still rely on methods from speech processing from decades ago [279] which employ reparametrization to ensure stability. Hence, it is possible that much progress and more refined models and methods can be achieved by leveraging advances in, for example, manifold learning [280–283] which have already been used for, e.g., localization and tracking [284]. It should be noted that this way of integrating signal processing and machine learning retains both the tractable nature of traditional signal processing and leads to interpretable, meaningful representations, yet builds on the advances in and advantages of machine learning and data-driven methods. In relation to this, it seems clear that signal processing is currently undergoing an transformation wherein machine learning is integrated into the core of its methodology, something that will lead to many new insights and advances of both academic and practical value, and these directions for future research ideas are thus very much aligned with this development.

[5]The method of [258] can be shown to be mathematically equivalent to NMF.

References

[1] G. E. Box, "Science and statistics," *Journal of the American Statistical Association*, vol. 71, no. 356, pp. 791–799, 1976.

[2] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[3] T. F. Quatieri, *Discrete-Time Speech Signal Processing*. Pearson Professional Education, 2001.

[4] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 468–478, 1978.

[5] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, pp. 716–723, 1974.

[6] M. Wax and T. Kailath, "Detection of the number of signals by information theoretic criterion," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, pp. 387–392, Apr. 1985.

[7] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[8] S. Wisdom, "Improving and unfolding statistical models of nonstationary signals," Ph.D. dissertation, University of Washington, 2017.

[9] K. Wilson, M. Chinen, J. Thorpe, B. Patton, J. Hershey, R. A. Saurous, J. Skoglund, and R. F. Lyon, "Exploring tradeoffs in models for low-latency speech enhancement," in *Proc. Int. Workshop on Acoustic Signal Enhancement*, 2018, pp. 366–370.

[10] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2159–2169, Sep. 2011.

[11] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, 2002.

[12] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[13] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and

source separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[14] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 744–754, 1986.

[15] R. J. McAulay and T. F. Quatieri, "Shape Invariant Time-Scale and Pitch Modification of Speech," in *IEEE Trans. Signal Processing*, vol. 40, no. 3, 1992, pp. 497–510.

[16] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 4, pp. 121–174.

[17] N. Mohammadiha, J. Taghia, and A. Leijon, "Single channel speech enhancement using Bayesian NMF with recursive temporal updates of prior distributions," in *IEEE Trans. Acoust., Speech, Signal Process.* IEEE, 2012, pp. 4561–4564.

[18] N. Mohammadiha and A. Leijon, "Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 5, pp. 998–1011, 2013.

[19] N. Mohammadiha and P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, Jun. 2013.

[20] N. Mohammadiha and S. Doclo, "Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 276–289, Feb. 2016.

[21] W. B. Kleijn and K. K. Paliwal, *Speech coding and synthesis*. Elsevier Science Inc., 1995.

[22] E. B. George and M. J. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 5, pp. 389–406, 1997.

[23] R. Heusdens, J. Jensen, W. Kleijn, V. Kot, O. Niamut, S. van de Par, N. van Schjindel, and R. Vafin, "Bit-rate scalable intraframe sinusoidal audio coding based on rate-distortion optimization," *J. Audio Eng. Soc.*, vol. 54, no. 3, pp. 167–188, 2006.

[24] N. H. van Schijndel *et al.*, "Adaptive RD optimized hybrid sound coding," *J. Audio Eng. Soc.*, vol. 56(10), pp. 787–809, 2008.

[25] M. H. Larsen, M. G. Christensen and S. H. Jensen, "Multiple description trellis-coded quantization of sinusoidal parameters," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 5287–5291, 2008.

[26] J. O. Smith III and X. Serra, "PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation," in *Proc. of the Int. Computer Music Conf.*, 1987.

[27] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis-synthesis of musical tones," *J. Audio Eng. Soc.*, vol. 40, no. 6, pp. 497–516, 1992.

[28] C. A. Rødbro, "Speech processing methods for the packet loss problem," Ph.D. dissertation, Aalborg University, 2004.

[29] J. Lindblom, "A sinusoidal voice over packet coder tailored for the frame-erasure channel," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 787–798, 2005.

[30] J. K. Nielsen, M. G. Christensen, A. T. Cemgil, S. J. Godsill, and S. H. Jensen, "Bayesian interpolation and parameter estimation in a dynamic sinusoidal model," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 1986–1998, 2011.

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

[32] P. Mowlaee, R. Saeidi, M. G. Christensen, Z.-H. Tan, T. Kinnunen, P. Fränti, and S. H. Jensen, "A joint approach for single-channel speaker identification and speech separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 9, pp. 2586–2601, 2012.

[33] J. R. Jensen, J. K. Nielsen, M. G. Christensen and S. H. Jensen, "On frequency domain models for TDOA estimation," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2015, pp. 11–15.

[34] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear least squares methods for joint DOA and pitch estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 5, pp. 923–933, 2013.

[35] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2012, pp. 409–412.

[36] M. W. Hansen, J. R. Jensen, and M. G. Christensen, "Estimation of fundamental frequencies in stereophonic music mixtures," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 2, pp. 296–310, 2019.

[37] M. W. Hansen, J. M. Hjerrild, M. G. Christensen, and J. Kjeldskov, "Parametric multi-channel separation and re-panning of harmonics sources„" in *Proc. Int. Conf. Digital Audio Effects*, 2018.

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

[39] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramér-Rao bound," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 5, pp. 720–741, May 1989.

[40] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramer-Rao bound," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 5, pp. 720–741, 1989.

[41] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramer-Rao bound; further results and comparisons," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 12, pp. 2140–2150, 1990.

[42] R. M. Gray *et al.*, "Toeplitz and circulant matrices: A review," *Foundations and Trends® in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[43] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Prentice Hall, 2005.

[44] H. Fletcher, "Normal vibration frequencies of a stiff piano string," in *J. Acoust. Soc. Amer.*, vol. 36, no. 1, 1962.

[45] S. Godsill and M. Davy, "Bayesian harmonic models for musical pitch estimation and analysis," in *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 2, 2002, pp. 1769–1772.

[46] G. Li, L. Qiu, and L. K. Ng, "Signal representation based on instantaneous amplitude models with application to speech synthesis," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 353–357, 2000.

[47] M. Zivanovic and J. Schoukens, "Single and piecewise polynomials for modeling of pitched sounds," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 4, pp. 1270–1281, 2012.

[48] Y. Pantazis, O. Rosec, and Y. Stylianou, "Chirp rate estimation of speech based on a time-varying quasi-harmonic model," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2009, pp. 3985–3988.

[49] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis," *IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3024–3051, 1993.

[50] B. Santhanam and P. Maragos, "Demodulation of discrete multicomponent AM-FM signals using periodic algebraic separation and energy demodulation," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, vol. 3, 1997, pp. 2409–2412.

[51] B. Santhanam and P. Maragos, "Multicomponent AM-FM demodulation via periodicity-based algebraic separation and energy-based demodulation," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 473–490, 2000.

[52] M. Képesi and L. Weruaga, "Adaptive chirp-based time-frequency analysis of speech signals," *Speech Communication*, vol. 48, no. 5, pp. 474–492, 2006.

[53] L. Weruaga and M. Képesi, "The fan-chirp transform for non-stationary harmonic signals," *Signal Processing*, vol. 87, no. 6, pp. 1504–1522, 2007.

[54] B. Resch, M. Nilsson, A. Ekman, and W. Kleijn, "Estimation of the instantaneous pitch of speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 813–822, 2007.

[55] S. Gonzalez and M. Brookes, "PEFAC-a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 2, pp. 518–530, 2014.

[56] M. G. Christensen and J. R. Jensen, "Pitch estimation for non-stationary speech," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2014, pp. 1400–1404.

[57] P. Djuric and S. M. Kay, "Parameter estimation of chirp signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 12, pp. 2118–2126, 1990.

[58] T. L. Jensen, J. K. Nielsen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "A fast algorithm for maximum-likelihood estimation of harmonic chirp parameters," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 5137–5152, 2017.

[59] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," *IEEE Trans. Signal Process.*, vol. 39, no. 2, pp. 411–423, 1991.

[60] T. L. Jensen, D. Giacobello, T. van Waterschoot, and M. G. Christensen, "Fast algorithms for high-order sparse linear prediction with applications to speech processing," *Speech Communication*, vol. 76, pp. 143–156, 2016.

[61] T. L. Jensen, D. Giacobello, M. G. Christensen, S. H. Jensen, and M. Moonen, "Real-time implementations of sparse linear prediction for speech processing," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2013, pp. 8184–8188.

[62] L. Shi, J. K. Nielsen, J. R. Jensen, and M. G. Christensen, "A variational em method for pole-zero modeling of speech with mixed block sparse and gaussian excitation," in *Proc. European Signal Process. Conf.*, 2017.

[63] L. Shi, J. R. Jensen, and M. G. Christensen, "Least 1-norm pole-zero modeling with sparse deconvolution for speech analysis." in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2017.

[64] D. Giacobello, M. G. Christensen, T. L. Jensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Stable 1-norm error minimization based linear predictors for speech modeling," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 5, pp. 912–922, 2014.

[65] W. C. Chu, *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*. Wiley, 2003.

[66] P. Kroon and W. B. Kleijn, "Linear-prediction based analysis-by-synthesis coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 3, pp. 79–119.

[67] J. Valin and J. Skoglund, "LPCNET: Improving neural speech synthesis through linear prediction," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2019, pp. 5891–5895.

[68] F. S. C. Lim, W. Bastiaan Kleijn, M. Chinen, and J. Skoglund, "Robust low rate speech coding based on cloned networks and Wavenet," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2020, pp. 6769–6773.

[69] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, ser. Signals and Communication Technology. Springer, 2005.

[70] B. G. Quinn and E. J. Hannan, *The Estimation and Tracking of Frequency*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2001.

[71] A. T. Cemgil, H. J. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 2, pp. 679–694, 2006.

[72] L. Parra and U. Jain, "Approximate kalman filtering for the harmonic plus noise model," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 2001, pp. 75–78.

[73] M. Davy, S. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *J. Acoust. Soc. Am.*, vol. 119, no. 4, pp. 2498–2517, 2006.

[74] D. Clark, A. T. Cemgil, P. Peeling, and S. Godsill, "Multi-object tracking of sinusoidal components in audio with the Gaussian mixture probability hypothesis density filter," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 2007, pp. 339–342.

[75] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[76] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 5, pp. 495–518.

[77] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–1652, 2008.

[78] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, ser. Synthesis Lectures on Speech & Audio Processing. Morgan & Claypool Publishers, 2009, vol. 5, 160 pages.

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

[80] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, May 2000.

[81] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays - Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Springer-Verlag, 2001, ch. 8, pp. 157–180.

[82] 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, 2003.

[83] P. Smaragdis, C. Fevotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 66–75, 2014.

[84] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[85] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, 1993.

[86] R. J. McAulay and E. M. Hofstetter, "Barankin bounds on parameter estimation," *IEEE Trans. Inf. Theory*, vol. 17, no. 6, pp. 669–676, 1971.

[87] D. Zachariah and P. Stoica, "Cramér-Rao Bound Analog of Bayes' Rule," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 164–168, 2015.

[88] P. Stoica, A. Jakobsson, and J. Li, "Cisoid parameter estimation in the colored noise case: asymptotic Cramér-Rao bound, maximum likelihood, and nonlinear least-squares," *IEEE Trans. Signal Process.*, vol. 45, no. 8, pp. 2048–2059, 1997.

[89] A. Jakobsson and P. Stoica, "Combining Capon and APES for Estimation of both Amplitude and Frequency of Spectral Lines," *Circuits, Systems, and Signal Process.*, vol. 19, no. 2, pp. 159–169, 2000.

[90] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: Survey, new results and an application," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 338–352, 2000.

[91] S. M. Kay, *Modern Spectral Estimation: Theory and Application*. Prentice-Hall, 1988.

[92] P. Stoica and T. Söderström, "On reparameterization of loss functions used in estimation and the invariance principle," *Signal Processing*, vol. 17, pp. 383–387, 1989.

[93] A. Swindlehurst and P. Stoica, "Maximum likelihood methods in radar array signal processing," *Proc. IEEE*, vol. 86, no. 2, pp. 421–441, 1998.

[94] M. G. Christensen, "Metrics for vector quantization-based parametric speech enhancement and separation," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 3062–3071, 2013.

[95] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Signal Processing*, vol. 80, pp. 1937–1944, 2000.

[96] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 4, pp. 477–489, 1988.

[97] J. Li and P. Stoica, "Efficient mixed-spectrum estimation with application to target feature extraction," *IEEE Trans. Signal Process.*, vol. 44, no. 2, pp. 281–295, 1996.

[98] M. Noll, "Pitch determination of human speech by harmonic product spectrum, the harmonic sum, and a maximum likelihood estimate," in *Proc. Symp. Comput. Process. Commun.*, 1969, pp. 779–797.

[99] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 4, pp. 477–489, 1988.

[100] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation," in *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 2, 1993, pp. 728–731.

[101] B. Quinn and P. Thomson, "Estimating the frequency of a periodic function," *Biometrika*, vol. 78, no. 1, pp. 65–74, 1991.

[102] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 76–87, 2004.

[103] J. K. Nielsen, T. L Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient," *Signal Processing*, vol. 135, pp. 188–197, 2017.

[104] L. Shi, J. K Nielsen, J. R. Jensen, M. A. Little, and M. G. Christensen, "Robust Bayesian pitch tracking based on the harmonic model," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 11, pp. 1737–1751, 2019.

[105] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[106] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[107] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[108] A. E. Jaramillo, A. Jakobsson, J. K. Nielsen, and M. G. Christensen, "Robust fundamental frequency estimation in coloured noise," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2020.

[109] M. G. Christensen and S. H. Jensen, "Variable order harmonic sinusoidal parameter estimation for speech and audio signals," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2006, pp. 1126–1130.

[110] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "Adaptive prewhitening based on parametric NMF," in *Proc. European Signal Process. Conf.*, 2019.

[111] P. C. Hansen and S. H. Jensen, "Prewhitening for rank-deficient noise in subspace methods for noise reduction," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3718–3726, 2005.

[112] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. The Johns Hopkins University Press, 1996.

[113] M. G. Christensen, A. Jakobsson and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 5, pp. 1635–1644, 2007.

[114] B. Yang, "Projection approximation subspace tracking," *IEEE Trans. Signal Process.*, vol. 43, no. 1, pp. 95–107, Jan. 1995.

[115] D. J. Rabideau, "Fast, rank adaptive subspace tracking and applications," *IEEE Trans. Signal Process.*, vol. 44, no. 9, pp. 2229–2244, 1996.

[116] P. Strobach, "Low-rank adaptive filters," *IEEE Trans. Signal Process.*, vol. 44, no. 12, pp. 2932–2947, 1996.

[117] R. Badeau, G. Richard, and B. David, "Sliding window adaptive SVD algorithms," *IEEE Trans. Signal Process.*, vol. 52, no. 1, pp. 1–10, Jan 2004.

[118] R. Badeau, B. David, and G. Richard, "Fast approximated power iteration subspace tracking," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2931–2941, Aug. 2005.

[119] P. Strobach, "The fast recursive row-householder subspace tracking algorithm," *Signal Processing*, vol. 89, no. 12, pp. 2514 – 2528, 2009.

[120] H. Bozdogan, "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions," *Psychometrika*, vol. 52, no. 3, pp. 345–370, 1987.

[121] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726–2735, 1998.

[122] J. K. Nielsen, M. G. Christensen, A. T. Cemgil, and S. H. Jensen, "Bayesian model comparison with the g-prior," *IEEE Trans. Signal Process.*, vol. 62, no. 1, pp. 225–238, 2014.

[123] L. Kavalieris and E. J. Hannan, "Determining the number of terms in a trigonometric regression," *J. on Time Series Analysis*, vol. 15, no. 6, pp. 613–625, 1994.

[124] B. G. Quinn, "Estimating the number of terms in a sinusoidal regression," *J. on Time Series Analysis*, vol. 10, no. 1, pp. 71–75, 1989.

[125] X. Wang, "An AIC type estimator for the number of cosinusoids," *J. on Time Series Analysis*, vol. 14, no. 4, pp. 434–440, 1993.

[126] E. J. Hannan, *Developments in Time Series Analysis*. Chapman and Hall, 1993, ch. Determining the number of jumps in a spectrum, pp. 127–138.

[127] G. Bienvenu and L. Kopp, "Optimality of high resolution array processing using the eigensystem approach," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 5, pp. 1235–1248, 1983.

[128] R. Badeau, B. David, and G. Richard, "A new perturbation analysis for signal enumeration in rotational invariance techniques," *IEEE Trans. Signal Process.*, vol. 54, no. 2, pp. 450–458, 2006.

[129] R. Roy and T. Kailath, "ESPRIT – estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, 1989.

[130] M. G. Christensen, A. Jakobsson and S. H. Jensen, "Fundamental frequency estimation using the shift-invariance property," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2007, pp. 631–635.

[131] G. Bienvenu, "Influence of the spatial coherence of the background noise on high resolution passive methods," in *IEEE Trans. Acoust., Speech, Signal Process.*, 1979, pp. 306–309.

[132] M. Haardt and J. A. Nossek, "Unitary esprit: how to obtain increased estimation accuracy with a reduced computational burden," *IEEE Trans. Signal Process.*, vol. 43, pp. 1232–1242, 1995.

[133] M. Viberg, B. Ottersten, and T. Kailath, "Detection and estimation in sensor arrays using weighted subspace fitting," *IEEE Trans. Signal Process.*, vol. 39, no. 11, pp. 2436–2449, 1991.

[134] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[135] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[136] M. M. Goodwin, "Adaptive Signal Models: Theory, Algorithms, and Audio Applications," Ph.D. dissertation, University of California, Berkeley, 1997.

[137] R. Gribonval, "Approximations non-linéaires pour l'analyse des signaux sonores," Ph.D. dissertation, Université de Paris IX Daupine, Paris, France, Sep. 1999.

[138] M. G. Christensen, "Estimation and modeling problems in parametric audio coding," Ph.D. dissertation, Aalborg University, Jul. 2005.

[139] M. G. Christensen and S. H. Jensen, "On perceptual distortion minimization and nonlinear least-squares frequency estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 1, pp. 99–109, 2006.

[140] T. Blumensath and M. Davies, "Iterative thresholding for sparse approximations," *J. Fourier Analysis Applications*, vol. 14, no. 5, pp. 629–654, 2008.

[141] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[142] A. Maleki, L. Anitori, Z. Yang, and R. G. Baraniuk, "Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP)," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4290–4308, jul 2013.

[143] H. Zou, "The adaptive Lasso and its oracle properties," *J. Amer. Stat. Assoc.*, vol. 101, no. 476, pp. 1418–1429, dec 2006.

[144] R. J. Tibshirani and J. Taylor, "The solution path of the generalized lasso," *Ann. Stat.*, vol. 39, no. 3, pp. 1335–1371, 2011.

[145] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Multi-pitch estimation exploiting block sparsity," *Signal Processing*, vol. 109, pp. 236–247, 2015.

[146] L. Shi, J. R. Jensen, J. K. Nielsen, and M. G. Christensen, "Multipitch estimation using block sparse Baysian learning and intra-block clustering," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2018, pp. 666–670.

[147] F. Elvander, J. Swärd, and A. Jakobsson, "Online estimation of multiple harmonic signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 273–284, 2017.

[148] F. Elvander, T. Kronvall, S. Adalbjörnsson, and A. Jakobsson, "An adaptive penalty multi-pitch estimator with self-regularization," *Signal Processing*, vol. 127, pp. 56 – 70, 2016.

[149] J. Swärd, J. Brynolfsson, A. Jakobsson, and M. Hansson-Sandsten, "Sparse semi-parametric estimation of harmonic chirp signals," *IEEE Trans. Signal Process.*, vol. 64, no. 7, pp. 1798–1807, 2016.

[150] J. Swärd, H. Li, and A. Jakobsson, "Off-grid fundamental frequency estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 2, pp. 296–303, 2018.

[151] R. Gribonval and E. Bacry, "Harmonic Decomposition of Audio Signals with Matching Pursuit," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 101–111, Jan. 2003.

[152] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 221–239, 2000.

[153] E. Denoel and J.-P. Solvay, "Linear prediction of speech with a least absolute error criterion," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 6, pp. 1397–1403, 1985.

[154] W. B. Kleijn and K. K. Paliwal, Eds., *Speech Coding and Synthesis*. Elsevier Science B.V., 1995.

[155] A. Jukić, T. van Waterschoot, and S. Doclo, "Adaptive speech dereverberation using constrained sparse multichannel linear prediction," *IEEE Signal Process. Lett.*, vol. 24, no. 1, pp. 101–105, 2017.

[156] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multichannel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 9, pp. 1509–1520, 2015.

[157] P. Kroon, E. D. F. Deprettere, and R. J. Sluyter, "Regular-pulse excitation - a novel approach to effective multipulse coding of speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 5, pp. 1054–1063, 1986.

[158] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, vol. 7, 1982.

[159] J. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[160] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, no. Jun, pp. 211–244, 2001.

[161] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2230–2249, 2009.

[162] D. Wipf and S. Nagarajan, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 317–329, 2010.

[163] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ 1 minimization," *J. Fourier Analysis and Appl.*, vol. 14, no. 5-6, pp. 877–905, 2008.

[164] A. Xenaki, J. B. Boldt, and M. G. Christensen, "Sound source localization and speech enhancement with sparse Bayesian learning beamforming," *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3912–3921, 2018.

[165] A. Xenaki, "High-resolution imaging methods in array signal processing," Ph.D. dissertation, Technical University of Denmark, 2015.

[166] C. M. Carvalho, N. G. Polson, and J. G. Scott, "Handling sparsity via the horseshoe," in *Proc. Int. Conf. on Artificial Intelligence and Stat.*, 2009.

[167] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer Science & Business Media, 2007.

[168] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays - Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer-Verlag, 2001, ch. 2, pp. 19–38.

[169] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218–1234, 2006.

[170] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, 2002.

[171] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multi-channel equalization," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 4, pp. 680–693, 2016.

[172] A. Spriet, M. Moonen, and J. Wouters, "Robustness analysis of multichannel wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 487–503, 2005.

[173] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays - Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer-Verlag, 2001, ch. 3, pp. 39–60.

[174] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 260–276, 2010.

[175] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 441–452, 2007.

[176] M. S. Kavalekalam, J. K. Nielsen, M. G. Christensen, and J. B. Boldt, "A study of noise PSD estimators for single channel speech enhancement," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2018, pp. 5464–5468.

[177] M. S. Kavalekalam, J. K. Nielsen, J. B. Boldt, and M. G. Christensen, "Model-based speech enhancement for intelligibility improvement in binaural hearing aids," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 1, pp. 99–113, 2019.

[178] J. K. Nielsen, M. S. Kavalekalam, M. G. Christensen, and J. B. Boldt, "Model-based noise PSD estimation from speech in non-stationary noise," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2018, pp. 5424–5428.

[179] Y. Zhao, J. K. Nielsen, J. Chen, and M. G. Christensen, "Model-based distributed node clustering and multi-speaker speech presence probability estimation in wireless acoustic sensor networks," *J. Acoust. Soc. Am.*, vol. 147, no. 6, pp. 4189–4201, 2020.

[180] K. Ngo, T. van Waterschoot, M. G. Christensen, M. Moonen, and S. H. Jensen, "Improved prediction error filters for adaptive feedback cancellation in hearing aids," *Signal Processing*, vol. 91, no. 11, pp. 3062–3075, 2013.

[181] C Sørensen, J. B. Boldt, and M. G. Christensen, "Harmonic beamformers for non-intrusive speech intelligibility prediction," in *Proc. Interspeech*, 2019.

[182] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Integrated sidelobe cancellation and linear prediction Kalman filter for joint multi-microphone speech dereverberation, interfering speech cancellation, and noise reduction," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 28, pp. 740–754, 2020.

[183] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 12, 1987, pp. 177–180.

[184] Z. Goh, K. C. Tan, and B. T. G. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 5, pp. 510–524, 1999.

[185] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 6, pp. 1741–1752, 2007.

[186] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 7, pp. 1948–1963, 2012.

[187] J. R. Jensen, "Enhancement of periodic signals: with application to speech signals," Ph.D. dissertation, Aalborg University, 2012.

[188] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[189] M. A. Lagunas, M. E. Santamaria, A. Gasull, and A. Moreno, "Maximum likelihood filters in spectral estimation problems," *Signal Processing*, vol. 10, no. 1, pp. 19–34, 1986.

[190] P. Stoica, A. Jakobsson, and J. Li, "Matched-filterbank interpretation of some spectral estimators," *Signal Processing*, vol. 66, pp. pp. 45–59, May 2005.

[191] A. Antoniou and W.-S. Lu, *Practical Optimization: Algorithms and Engineering Applications*. Springer Verlag, 2007.

[192] K. Buckley, "Broad-band beamforming and the generalized sidelobe canceller," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 5, pp. 1322–1323, 1986.

[193] S. Markovich-Golan, S. Gannot, and I. Cohen, "Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 2, pp. 343–356, 2013.

[194] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 7, pp. 1420–1434, 2009.

[195] J. R. Jensen, G.-O. Glentis, M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Fast LCMV-based methods for fundamental frequency estimation," *IEEE Trans. Signal Process.*, vol. 61, no. 12, pp. 3159–3172, 2013.

[196] V. C. Shields, Jr., "Separation of added speech signals by digital comb filtering," Master's thesis, Massachusetts Institute of Technology, Sep. 1970.

[197] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 5, pp. 1124–1138, 1986.

[198] J. Moorer, "The optimum comb method of pitch period analysis of continuous digitized speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 22, no. 5, pp. 330–338, 1974.

[199] J. Lim, A. Oppenheim, and L. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 4, pp. 354–358, 1978.

[200] V. Valimaki, J. D. Parker, J. O. S. L. Savioja, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 20, no. 5, pp. 1421–1448, 2012.

[201] J. Li, P. Stoica, and Z. Wang, "On robust Capon beamforming and diagonal loading," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1702–1715, Jul. 2003.

[202] S. Doclo and M. Moonen, "Superdirective beamforming robust against microphone mismatch," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 2, pp. 617–631, 2007.

[203] J. Li, P. Stoica, and Z. Wang, "Doubly constrained robust capon beamformer," *IEEE Trans. Signal Process.*, vol. 52, no. 9, pp. 2407–2423, 2004.

[204] Y. Zhao, J. R. Jensen, M. G. Christensen, S. Doclo, and J. Chen, "Experimental study of robust beamforming techniques for acoustic applications," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 2017, pp. 86–90.

[205] Y. Zhao, J. R. Jensen, T. L. Jensen J. Chen, and M. G. Christensen, "Experimental study of robust acoustic beamforming for speech acquisition in reverberant and noisy environments," *Applied Acoustics*, vol. 170, p. 107531, 2020.

[206] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "On the influence of inharmonicities in model-based speech enhancement," in *Proc. European Signal Process. Conf.*, 2013, pp. 1–5.

[207] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, 1995.

[208] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sørensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 6, pp. 439–448, 1995.

[209] P. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[210] P. Hansen and S. Jensen, "FIR filter representations of reduced-rank noise reduction," *IEEE Trans. Signal Process.*, vol. 46, no. 6, pp. 1737–1741, 1998.

[211] J. R. Jensen, J. Benesty, and M. G. Christensen, "Noise reduction with optimal variable span linear filter," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 631–644, 2016.

[212] L. L. Scharf, "The SVD and reduced rank signal processing," *Signal Processing*, vol. 25, pp. 113–133, 1991.

[213] P. Strobach, "Low-rank adaptive filters," *IEEE Trans. Signal Process.*, vol. 44, no. 12, pp. 2932–2947, Dec. 1996.

[214] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP J. on Advances in Signal Process.*, vol. 2007, no. 1, p. 092953, 2007.

[215] J. Benesty, M. G. Christensen, and J. R. Jensen, *Signal Enhancement with Variable Span Linear Filters*. Springer-Verlag, 2016.

[216] J. Benesty, J. R. Jensen, M. G. Christensen, and J. Chen, *Speech enhancement: A signal subspace perspective*. Elsevier, 2014.

[217] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 4, pp. 785–799, 2014.

[218] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Square root-based multi-source early PSD estimation and recursive RETF update in reverberant environments by means of the orthogonal procrustes problem," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 28, pp. 755–769, 2020.

[219] I. Kodrasi and S. Doclo, "Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 26, no. 6, pp. 1106–1118, 2018.

[220] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 9, pp. 1879–1890, 2013.

[221] O. Schwartz, S. Gannot, and E. A. P. Habets, "An expectation-maximization algorithm for multimicrophone speech dereverberation and noise reduction with coherence matrix estimation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 9, pp. 1495–1510, Sep. 2016.

[222] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 7, pp. 1492–1501, 2017.

[223] A. I. Koutrouvelis, T. W. Sherson, R. Heusdens, and R. C. Hendriks, "A low-cost robust distributed linearly constrained beamformer for wireless acoustic sensor networks with arbitrary topology," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1434–1448, 2018.

[224] J. Benesty, Y. Huang, and J. Chen, *Microphone Array Signal Processing*. Springer-Verlag, 2008, vol. 1.

[225] J. R. Jensen and M. G. Christensen, "Near-field localization of audio: A maximum likelihood approach," in *Proc. European Signal Process. Conf.*, 2014, pp. 895–899.

[226] J. R. Jensen, J. K. Nielsen, R. Heusdens, and M. G. Christensen, "DOA estimation of audio sources in reverberant environments," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2016, pp. 176–180.

[227] L. Y. Ngan, Y. Wu, H. C. So, P. C. Ching, and S. W. Lee, "Joint time delay and pitch estimation for speaker localization," in *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 3, May 2003, pp. 722–725.

[228] M. Képesi, L. Ottowitz, and T. Habib, "Joint position-pitch estimation for multiple speaker scenarios," in *Proc. Hands-Free Speech Commun. Microphone Arrays*, 2008, pp. 85–88.

[229] M. Wohlmayr and M. Képesi, "Joint position-pitch extraction from multichannel audio," in *Proc. Interspeech*, 2007, pp. 1629–1632.

[230] 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 Process.*, vol. 2012, no. 1, pp. 1–11, 2012.

[231] X. Qian and R. Kumaresan, "Joint estimation of time delay and pitch of voiced speech signals," *Rec. Asilomar Conf. Signals, Systems, and Computers*, vol. 1, pp. 735–739, 1995.

[232] J. Dmochowski, J. Benesty, and S. Affes, "Linearly constrained minimum variance source localization and spectral estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1490–1502, 2008.

[233] F. Flego and M. Omologo, "Robust f0 estimation based on a multi-microphone periodicity function for distant-talking speech," in *Proc. European Signal Process. Conf.*, Sep. 2006, pp. 1–4.

[234] T. Gerkmann, R. Martin, and D. Dalga, "Multi-microphone maximum a posteriori fundamental frequency estimation in the cepstral domain," in *IEEE Trans. Acoust., Speech, Signal Process.*, 2009, pp. 4505–4508.

[235] M. G. Christensen, A. Jakobsson and S. H. Jensen, "Sinusoidal order estimation using the subspace orthogonality and shift-invariance properties," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2007, pp. 651–655.

[236] M. G. Christensen and A. Jakobsson, "Optimal filters for extraction and separation of periodic sources," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2009, pp. 376–379.

[237] M. G. Christensen and A. Jakobsson, "Improved subspace-based frequency estimation for real-valued data using angles between subspaces," in *Proc. European Signal Process. Conf.*, 2010.

[238] M. G. Christensen, "On the estimation of low fundamental frequencies," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 2011, pp. 169–172.

[239] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Enhancement of non-stationary speech using harmonic chirp filters," in *Proc. Interspeech*, 2015.

[240] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Sparse linear predictors for speech processing," in *Proc. Interspeech*, 2008.

[241] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Joint estimation of short-term and long-term predictors in speech coders," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2009, pp. 4109–4112.

[242] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Speech coding based on sparse linear prediction," in *Proc. European Signal Process. Conf.*, 2009, pp. 2524–2528.

[243] D. Giacobello, M. N. Murthi, M. G. Christensen, S. H. Jensen, and M. Moonen, "Re-estimation of linear predictive parameters in sparse linear prediction," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2009, pp. 1770–1773.

[244] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Enhancing sparsity in linear prediction of speech by iteratively reweighted 1-norm minimization," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2010, pp. 4650–4653.

[245] D. Giacobello, M. N. Murthi, M. G. Christensen, S. H. Jensen, and M. Moonen, "Estimation of frame independent and enhancement components for speech communication over packet networks," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2010, pp. 4682–4685.

[246] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Retrieving sparse patterns using a compressed sensing framework: Applications to speech coding based on sparse linear prediction," *IEEE Signal Process. Lett.*, vol. 17, no. 1, pp. 103–106, 2010.

[247] D. Giacobello, T. van Waterschoot, M. G. Christensen, S. H. Jensen, and M. Moonen, "High-order sparse linear predictors for audio processing," in *Proc. European Signal Process. Conf.*, 2010.

[248] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Joint DOA and Fundamental Frequency Estimation Methods based on 2-D Filtering," in *Proc. European Signal Process. Conf.*, 2010.

[249] J. R. Jensen and M. G. Christensen, "DOA and pitch estimation of audio sources using IAA-based filtering," in *Proc. European Signal Process. Conf.*, 2014, pp. 900–904.

[250] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, 2013.

[251] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks,"

IEEE Trans. Audio, Speech, Language Process., vol. 25, no. 1, pp. 102–111, 2017.

[252] M. Kolbæk, D. Yu, Z. H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[253] E. M. Grais, G. Roma, A. J. R. Simpson, and M. D. Plumbley, "Two-stage single-channel audio source separation using deep neural networks," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 9, pp. 1773–1783, Sep. 2017.

[254] D. Sculley, J. Snoek, A. B. Wiltschko, and A. Rahimi, "Winner's curse? on pace, progress, and empirical rigor," in *Int. Conf. Learning Representations*, 2018.

[255] Z. C. Lipton and J. Steinhardt, "Troubling trends in machine learning scholarship," *Queue*, vol. 17, no. 1, Feb. 2019.

[256] P. Velickovic, "The resurgence of structure in deep neural networks," Ph.D. dissertation, University of Cambridge, 2019.

[257] P. Stoica and P. Babu, "The Gaussian Data Assumption Leads to the Largest Cramér-Rao Bound," *IEEE Signal Process. Mag.*, vol. 28, no. 3, pp. 132–133, 2011.

[258] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 1, pp. 163–176, 2005.

[259] M. S. Kavalekalam, J. K. Nielsen, L. Shi, M. G. Christensen, and J. B. Boldt, "Online parametric nmf for speech enhancement," in *Proc. European Signal Process. Conf.*, 2018, pp. 2320–2324.

[260] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *IEEE Trans. Acoust., Speech, Signal Process.*, 2008, pp. 4029–4032.

[261] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2140–2151, 2013.

[262] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation*. Springer Science & Business Media, 2010.

[263] A. Jukić and S. Doclo, "Speech dereverberation using weighted prediction error with Laplacian model of the desired signal," in *IEEE Trans. Acoust., Speech, Signal Process.*, 2014, pp. 5172–5176.

[264] N. D. Gaubitch, D. B. Ward, and P. A. Naylor, "Statistical analysis of the autoregressive modeling of reverberant speech," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 4031–4039, 2006.

[265] B. Defraene, N. Mansour, S. D. Hertogh, T. van Waterschoot, M. Diehl, and M. Moonen, "Declipping of audio signals using perceptual compressed sensing," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 12, pp. 2627–2637, 2013.

[266] S. Godsill, P. Rayner, and O. Cappé, "Digital audio restoration," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds. Boston, MA: Springer, 2002, pp. 133–194.

[267] S. Godsill, "The restoration of degraded audio signals," Ph.D. dissertation, University of Cambridge, 1993.

[268] J. Paduart, L. Lauwers, J. Swevers, K. Smolders, J. Schoukens, and R. Pintelon, "Identification of nonlinear systems using polynomial nonlinear state space models," *Automatica*, vol. 46, no. 4, pp. 647 – 656, 2010.

[269] J. Schoukens, R. Pintelon, T. Dobrowiecki, and Y. Rolain, "Identification of linear systems with nonlinear distortions," *Automatica*, vol. 41, no. 3, pp. 491 – 504, 2005.

[270] A. Sangwan, L. Kaushik, C. Yu, J. H. L. Hansen, and D. W. Oard, "'Houston, we have a solution': using NASA apollo program to advance speech and language processing technology," in *Proc. Interspeech*, 2013.

[271] J. Plata-Chaves, A. Bertrand, M. Moonen, S. Theodoridis, and A. M. Zoubir, "Heterogeneous and multitask wireless sensor networks–algorithms, applications, and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 3, pp. 450–465, 2017.

[272] A. Hassani, A. Bertrand, and M. Moonen, "GEVD-based low-rank approximation for distributed adaptive node-specific signal estimation in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2557–2572, May 2016.

[273] Y. Zhao, J. K. Nielsen, M. G. Christensen, and J. Chen, "Model-based voice activity detection in wireless acoustic sensor networks," in *Proc. European Signal Process. Conf.*, 2018, pp. 425–429.

[274] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 4, pp. 546–555, 2009.

[275] F. Barzideh, K. Skretting, and K. Engan, "Imposing shift-invariance using flexible structure dictionary learning (FSDL)," *Digital Signal Processing*, vol. 69, pp. 162 – 173, 2017.

[276] C. Paleologu, J. Benesty, and S. Ciochină, "Linear system identification based on a Kronecker product decomposition," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 26, no. 10, pp. 1793–1808, 2018.

[277] F. Elvander, J. Ding, and A. Jakobsson, "On harmonic approximations of inharmonic signals," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2020, pp. 5360–5364.

[278] C. D. Richmond and L. L. Horowitz, "Parameter bounds on estimation accuracy under model misspecification," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2263–2278, 2015.

[279] W. B. Kleijn and K. K. Paliwal, "Quantization of LPC Parameters," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 12.

[280] T. Lin and H. Zha, "Riemannian manifold learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 796–809, 2008.

[281] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[282] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[283] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems*. MIT Press, 2002, vol. 14, pp. 585–591.

[284] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "A hybrid approach for speaker tracking based on TDOA and data-driven models," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 26, no. 4, pp. 725–735, 2018.

Paper A

Sinusoidal Order Estimation using Angles between Subspaces

M. G. Christensen, A. Jakobsson, and S. H. Jensen

The paper has been published in the
EURASIP Journal on Advances in Signal Processing, Article ID 948756, pp. 1—11, 2009.

Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 948756, 11 pages
doi:10.1155/2009/948756

Research Article

Sinusoidal Order Estimation Using Angles between Subspaces

Mads Græsbøll Christensen,[1] Andreas Jakobsson (EURASIP Member),[2] and Søren Holdt Jensen (EURASIP Member)[3]

[1] *Department of Media Technology, Aalborg University, Niels Jernes Vej 14, 9220 Aalborg, Denmark*
[2] *Department of Mathematical Statistics, Lund University, 221 00 Lund, Sweden*
[3] *Department of Electronic Systems, Aalborg University, Niels Jernes Vej 12, 9220 Aalborg, Denmark*

Correspondence should be addressed to Mads Græsbøll Christensen, mgc@imi.aau.dk

Received 12 June 2009; Revised 2 September 2009; Accepted 16 September 2009

Recommended by Walter Kellermann

We consider the problem of determining the order of a parametric model from a noisy signal based on the geometry of the space. More specifically, we do this using the nontrivial angles between the candidate signal subspace model and the noise subspace. The proposed principle is closely related to the subspace orthogonality property known from the MUSIC algorithm, and we study its properties and compare it to other related measures. For the problem of estimating the number of complex sinusoids in white noise, a computationally efficient implementation exists, and this problem is therefore considered in detail. In computer simulations, we compare the proposed method to various well-known methods for order estimation. These show that the proposed method outperforms the other previously published subspace methods and that it is more robust to the noise being colored than the previously published methods.

1. Introduction

Estimating the order of a model is a central, yet commonly overlooked, problem in parameter estimation, with the majority of literature assuming prior knowledge of the model order. In many cases, however, the order cannot be known a priori and may change over time. This is the case, for example, in speech and audio signals. Many parameter estimation methods, like the maximum likelihood and subspace methods, require that the order is known to work properly. The consequence of choosing an erroneous order, aside from the size of the parameter set being wrong, is that the found parameters may be biased or suffer from a huge variance. The most commonly used methods for estimating the model order are perhaps the minimum description length (MDL) [1, 2], the Akaike information criterion (AIC) [3], and the maximum a posteriori (MAP) rule of [4]. These methods are based on certain asymptotic approximations and on statistical models of the observed signal, like the noise being white and Gaussian distributed. We refer the interested reader to [4, 5] for an overview of such statistical methods. A notable feature of the MAP rule of [4] is that it shows that linear and nonlinear parameters should be penalized differently, something that not recognized by many prior methods (on this topic, see also [6]). In this paper, we are concerned with a more specific, yet important, case, namely, that of finding the number of complex sinusoids buried in noise. This problem is treated in great detail from a statistical point of view in [4] and is also exemplified in [5] and other notable approaches include those of [7–13]. A different class of methods is subspace methods, which is also the topic of interest here. In subspace methods, the eigenvectors of the covariance matrix are divided into a set that spans the space of the signal of interest, called the signal subspace, and its orthogonal complement, the noise subspace. These subspaces and their properties can then be used for various estimation and identification tasks. Subspace methods have a rich history in parameter estimation and signal enhancement. Especially for the estimation of sinusoidal frequencies and finding the direction of arrival of sources in array processing, these methods have proven successful during the past three decades. The most common subspace methods for parameter estimation are perhaps the MUSIC (MUltiple SIgnal Classification) method

[14, 15] and the ESPRIT (Estimation of Signal Parameters via Rotational Invariance Techniques) method of [16] while the earliest example of such methods is perhaps Pisarenko's method [17]. In the context of subspace methods, the typical way of finding the dimensions of the signal and noise subspaces is based on statistical principles where the likelihood function of the observation vector is combined with one of the aforementioned order selection rules with the likelihood function depending on the ratio between the arithmetic and geometric means of the eigenvalues [18, 19]. Recently, the underlying principles of ESPRIT and MUSIC have been extended to the problem of order estimation by exploiting the properties of the eigenvectors rather than the eigenvalues. Compared to the order estimation techniques based on the eigenvalues, one can interpret these methods as being based on the geometry of the space rather than the distribution of energy. Specifically, two related subspace methods based on ESPRIT have been proposed, namely, to the ESTimation ERror (ESTER) method [20] and the Subspace-based Automatic Model Order Selection (SAMOS) method [21]. Similarly, it was shown in [22] that the orthogonality principle of MUSIC can be used for finding the number of harmonics for a set of harmonically related sinusoids when normalized appropriately. See also [23] for a comparison of this method with the ESTER and SAMOS methods. An attractive property of the subspace-based order estimation criteria is that they do not require prior knowledge of the probability density function (pdf) of the observation noise but only a consistent covariance matrix estimate. This means that the subspace methods will work in situations where the statistical methods may fail due to the assumed pdf not being a good approximation of the observed data. Furthermore, it can be quite difficult to derive a method like the MAP rule of [4] for complicated signal models.

Mathematically, the specific problem considered herein can be stated as follows. A signal consisting of complex sinusoids having frequencies $\{\omega_l\}$ is corrupted by additive noise, $\epsilon(n)$, for $n = 0, \ldots, N-1$,

$$x(n) = \sum_{l=1}^{L} A_l e^{j(\omega_l n + \phi_l)} + \epsilon(n), \tag{1}$$

where $A_l > 0$ and ϕ_l are the amplitude and the phase of the lth sinusoid. Here, $\epsilon(n)$ is assumed to be white complex symmetric zero-mean noise. The problem considered is then how to estimate the model order L. The model in (1) may seem a bit restrictive, but the proposed method can in fact be used for more general problems. Firstly, the proposed method is valid for a large class of signal models; however, for the case of complex exponentials a computationally efficient implementation of our method exists. This is also the case for damped sinusoids where the principles of unitary ESPRIT may be applied [24]. Secondly, for the case of colored noise, the proposed method is also applicable by the use of prewhitening.

In this paper, we study the problem of finding the model order using the angles between a candidate signal subspace and the signal subspace in depth. In the process of finding the model order, nonlinear model parameters are also found. The concept of angles between subspaces has previously been applied within the field of signal processing to, among other things, analysis of subspace-based enhancement algorithms, for example, [25, 26], and multipitch estimation [27]. For complex sinusoids, the measure based on angles between subspaces reduces to a normalization of the well-known cost function first proposed for frequency and direction-of-arrival estimation in [14] for a high number of observations. We analyze, discuss, and compare the measure and its properties to other commonly used measures of the angles between subspaces and show that the proposed measure provides an upper bound for some other more complicated measures. These other measures turn out to be less useful for our application, and, in simulations, we compare the proposed method to other methods for finding the number of complex sinusoids. Our results show that the method has comparable performance to commonly used methods and is generally best among the subspace-based methods. It is also demonstrated, however, that the method is more robust to model violations, like colored noise. As an aside, our results also establish the MUSIC criterion for parameter estimation [14] as an approximation to the angles between the noise and candidate model subspaces.

The remaining part of this paper is organized as follows. First, we recapitulate the covariance matrix model that forms the basis for the subspace methods and briefly describe the MUSIC method in Section 2. In Section 3, we then move on to derive the new measure based on angles between subspaces. We relate this measure to other similar measures and proceed to discuss its properties and application to the problem interest. The statistical performance of the method is then evaluated in simulations studies in Section 4 and compared to a number of related parametric and nonparametric methods and, in Section 5, the results are discussed. Finally, we conclude on our work in Section 6.

2. Fundamentals

We start out this section by presenting some fundamental definitions, relations, and results. First, we define $\mathbf{x}(n)$ as a signal vector, referred to as a subvector, containing $M < N$ samples of the observed signal, that is,

$$\mathbf{x}(n) = \begin{bmatrix} x(n) & x(n+1) & \cdots & x(n+M-1) \end{bmatrix}^T \tag{2}$$

with $(\cdot)^T$ denoting the transpose. Assuming that the phases of the sinusoids are independent and uniformly distributed on the interval $(-\pi, \pi]$, the covariance matrix $\mathbf{R} \in \mathbb{C}^{M \times M}$ of the signal in (1) can be written as [5]

$$\mathbf{R} = \mathrm{E}\left\{\mathbf{x}(n)\mathbf{x}^H(n)\right\} = \mathbf{A}\mathbf{P}\mathbf{A}^H + \sigma^2\mathbf{I}_M, \tag{3}$$

where $\mathrm{E}\{\cdot\}$ and $(\cdot)^H$ denote the statistical expectation and the conjugate transpose, respectively. We here require that $L < M$. Moreover, we note that for the above to hold, the noise need not be Gaussian. The matrix $\mathbf{P}$ is diagonal and contains the squared amplitudes, that is,

$\mathbf{P} = \text{diag}([\ A_1^2\ \cdots\ A_L^2\ ])$, and $\mathbf{A} \in \mathbb{C}^{M\times L}$ is a Vandermonde matrix defined as

$$\mathbf{A} = \left[\mathbf{a}(\omega_1)\ \cdots\ \mathbf{a}(\omega_L)\right], \tag{4}$$

where $\mathbf{a}(\omega) = [\ 1 e^{j\omega}\ \cdots\ e^{j\omega(M-1)}\ ]^T$. Also, σ^2 denotes the variance of the additive noise, $\epsilon(n)$, and $\mathbf{I}_M$ is the $M \times M$ identity matrix. Assuming that the frequencies $\{\omega_l\}$ are distinct, the columns of $\mathbf{A}$ are linearly independent and $\mathbf{A}$ and $\mathbf{APA}^H$ have rank L. Let

$$\mathbf{R} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^H \tag{5}$$

be the eigenvalue decomposition (EVD) of the covariance matrix. Then, $\mathbf{Q}$ contains the M orthonormal eigenvectors of $\mathbf{R}$, that is, $\mathbf{Q} = [\ \mathbf{q}_1\ \cdots\ \mathbf{q}_M\ ]$ and $\mathbf{\Lambda}$ is a diagonal matrix containing the corresponding eigenvalues, λ_k, with

$$\lambda_1 \geq \cdots \geq \lambda_L \geq \lambda_{L+1} = \cdots = \lambda_M = \sigma^2. \tag{6}$$

The subspace-based methods are based on a partitioning of the eigenvectors into a set belonging to the signal subspace spanned by the columns of $\mathbf{A}$ and its orthogonal complement known as the noise subspace. Let $\mathbf{S}$ be formed from the eigenvectors corresponding to the L most significant eigenvalues, that is,

$$\mathbf{S} = \left[\mathbf{q}_1\ \cdots\ \mathbf{q}_L\right]. \tag{7}$$

We denote the space spanned by the columns of $\mathbf{S}$ as $\mathcal{R}(\mathbf{S})$ and henceforth refer to it as the signal subspace. Similarly, let $\mathbf{G}$ be formed from the eigenvectors corresponding to the $M - L$ least significant eigenvalues, that is,

$$\mathbf{G} = \left[\mathbf{q}_{L+1}\ \cdots\ \mathbf{q}_M\right], \tag{8}$$

where $\mathcal{R}(\mathbf{G})$ is referred to as the noise subspace. Using the EVD in (5), the covariance matrix model in (3) can now be written as $\mathbf{Q}(\mathbf{\Lambda} - \sigma^2\mathbf{I}_M)\mathbf{Q}^H = \mathbf{APA}^H$. Introducing $\mathbf{\Lambda}_S = \text{diag}([\ \lambda_1 - \sigma^2\ \cdots\ \lambda_L - \sigma^2\ ])$, we can write this as

$$\mathbf{S}\mathbf{\Lambda}_S\mathbf{S}^H = \mathbf{APA}^H. \tag{9}$$

From the last equation, it can be seen that the columns of $\mathbf{A}$ span the same space as the columns of $\mathbf{S}$ and that $\mathbf{A}$ therefore also must be orthogonal to $\mathbf{G}$, that is,

$$\mathbf{A}^H\mathbf{G} = \mathbf{0}. \tag{10}$$

In practice, the eigenvectors are of course unknown and are replaced by estimates. Here, we will estimate the covariance matrix as

$$\hat{\mathbf{R}} = \frac{1}{N-M+1}\sum_{n=0}^{N-M}\mathbf{x}(n)\mathbf{x}^H(n), \tag{11}$$

which is a consistent estimate for ergodic processes and the maximum likelihood estimate for Gaussian noise. The eigenvector estimates obtained from this matrix are then also consistent and the covariance matrix model (3) and the orthogonality property (10) therefore hold asymptotically.

Since the covariance matrix and eigenvectors are estimated from a finite set of vectors, the orthogonality property in (10) only holds approximately. In the MUSIC algorithm [14, 15], the set of distinct frequencies $\{\omega_l\}$ are found by minimizing the Frobenius norm, denoted $\|\cdot\|_F$, of (10), that is,

$$\{\hat{\omega}_l\} = \arg\min_{\{\omega_l\}}\left\|\mathbf{A}^H\mathbf{G}\right\|_F^2. \tag{12}$$

Since the squared Frobenius norm is additive over the columns of $\mathbf{A}$, we can find the individual sinusoidal frequencies for $l = 1, \ldots, L$ as

$$\hat{\omega}_l = \arg\min_{\omega_l}\left\|\mathbf{a}^H(\omega_l)\mathbf{G}\right\|_F^2 \tag{13}$$

with the requirements that the frequencies are distinct and fulfill the two following conditions:

$$\frac{\partial\|\mathbf{a}^H(\omega_l)\mathbf{G}\|_F^2}{\partial\omega_l} = 0, \qquad \frac{\partial^2\|\mathbf{a}^H(\omega_l)\mathbf{G}\|_F^2}{\partial\omega_l^2} > 0. \tag{14}$$

The reciprocal form of the cost function in (13) is sometimes referred to as spectral MUSIC and $1/\|\mathbf{a}^H(\omega_l)\mathbf{G}\|_F^2$ as the pseudospectrum from which the L frequencies are obtained as the peaks. We mention in passing that it is possible to solve (13) using numeric rooting methods [28] or FFTs. Regarding the statistical properties of MUSIC, the effects of order estimation errors, that is, the effect of choosing an erroneous $\mathbf{G}$ in (13), on the parameter estimates obtained using MUSIC have been studied in [29] in a slightly different context and it was concluded that the MUSIC estimator is more sensitive to underestimation of L than overestimation. The more common case of L being known has been treated in great detail, with the statistical properties of MUSIC having been studied in [30–34].

3. Angles between Subspaces

3.1. Definition and Basic Results. The orthogonality property states that for the true parameters, the matrix $\mathbf{A}$ is orthogonal to the noise subspace eigenvectors in $\mathbf{G}$. For estimation purposes, we need a measure of this. The concept of orthogonality is of course closely related to the concept of angles, and how to define angles in multidimensional spaces is what we will now investigate further.

The principal (nontrivial) angles $\{\theta_k\}$ between the two subspaces $\mathcal{A} = \mathcal{R}(\mathbf{A})$ and $\mathcal{G} = \mathcal{R}(\mathbf{G})$ are defined recursively for $k = 1, \ldots, K$ as (see, e.g., [35])

$$\cos(\theta_k) = \max_{\mathbf{u}\in\mathcal{A}}\max_{\mathbf{v}\in\mathcal{G}}\frac{\mathbf{u}^H\mathbf{v}}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2} \triangleq \mathbf{u}_k^H\mathbf{v}_k. \tag{15}$$

The quantity K is the minimal dimension of the two subspaces, that is, $K = \min\{L, M - L\}$, which is the number of nontrivial angles between the two subspaces. Moreover, the directions along which the angles are defined

are orthogonal, that is, $\mathbf{u}^H\mathbf{u}_i = 0$ and $\mathbf{v}^H\mathbf{v}_i = 0$ for $i = 1,\ldots,k-1$.

We will now rewrite (15) into something more useful, and in doing this, we will make extensive use of projection matrices. The (orthogonal) projection matrix for a subspace $\mathcal{X}$ spanned by the columns of a matrix $\mathbf{X}$ is defined as $\mathbf{\Pi}_X = \mathbf{X}(\mathbf{X}^H\mathbf{X})^{-1}\mathbf{X}^H$. Such projections matrices are Hermitian, that is, $\mathbf{\Pi}_X^H = \mathbf{\Pi}_X$ and have the properties $\mathbf{\Pi}_X^m = \mathbf{\Pi}_X$ for $m = 1,2,\ldots$ and $\|\mathbf{\Pi}_X\|_F^2 = \dim(\mathcal{X})$ where $\dim(\cdot)$ is the dimension of the subspace. Let $\mathbf{\Pi}_G$ be the projection matrix for subspace $\mathcal{G}$, and the $\mathbf{\Pi}_A$ the projection matrix for subspace $\mathcal{A}$. Using the two projection matrices, we can write the vector $\mathbf{u} \in \mathcal{A}$ as $\mathbf{\Pi}_A\mathbf{y}$ and $\mathbf{v} \in \mathcal{G}$ as $\mathbf{\Pi}_G\mathbf{z}$ with $\mathbf{y}, z \in \mathbb{C}^M$. This allows us to express (15) as

$$\begin{aligned}\cos(\theta_k) &= \max_{\mathbf{y}\in\mathbb{C}^M}\max_{\mathbf{z}\in\mathbb{C}^M}\frac{\mathbf{y}^H\mathbf{\Pi}_A\mathbf{\Pi}_G\mathbf{z}}{\|\mathbf{y}\|_2\|\mathbf{z}\|_2}\\ &\triangleq \mathbf{y}_k^H\mathbf{\Pi}_A\mathbf{\Pi}_G\mathbf{z}_k = \sigma_k\end{aligned} \tag{16}$$

for $k = 1,\ldots,K$. Again, we require that $\mathbf{y}^H\mathbf{y}_i = 0$ and $\mathbf{z}^H\mathbf{z}_i = 0$ for $i = 1,\ldots,k-1$, that is, that the vectors are orthogonal. Futhermore, the denominator ensures that the vectors have unit norm. It then follows that $\{\sigma_k\}$ are the singular values of the matrix product $\mathbf{\Pi}_A\mathbf{\Pi}_G$, and that the two sets of vectors $\{\mathbf{y}\}$ and $\{\mathbf{z}\}$ are the left and right singular vectors, respectively. Regarding the mapping of the singular values to actual angles, a difficult problem, we refer the interested reader to [36] for a numerically stable algorithm.

The set of principal angles obey the following inequality:

$$0 \le \theta_1 \le \cdots \le \theta_K \le \frac{\pi}{2}. \tag{17}$$

Next, the singular values are related to the Frobenius norm of the product $\mathbf{\Pi}_A\mathbf{\Pi}_G$ as

$$\|\mathbf{\Pi}_A\mathbf{\Pi}_G\|_F^2 = \mathrm{Tr}\{\mathbf{\Pi}_A\mathbf{\Pi}_G\} = \sum_{k=1}^{K}\sigma_k^2, \tag{18}$$

and therefore also to the angles between the subspaces, that is,

$$\sum_{k=1}^{K}\cos^2(\theta_k) = \|\mathbf{\Pi}_A\mathbf{\Pi}_G\|_F^2. \tag{19}$$

3.2. A Simplified Measure. We will now show how the concepts introduced in the previous section can be simplified for use in estimation. The Frobenius norm of the product $\mathbf{\Pi}_A\mathbf{\Pi}_G$ can be expressed as

$$\|\mathbf{\Pi}_A\mathbf{\Pi}_G\|_F^2 = \mathrm{Tr}\left\{\mathbf{\Pi}_A\mathbf{\Pi}_G\mathbf{\Pi}_G^H\mathbf{\Pi}_A^H\right\} = \mathrm{Tr}\left\{\mathbf{\Pi}_A\mathbf{\Pi}_G^H\right\} \tag{20}$$

$$= \mathrm{Tr}\left\{\mathbf{A}\left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{A}^H\mathbf{G}\mathbf{G}^H\right\}. \tag{21}$$

This expression can be seen to be complicated since it involves matrix inversion and it does not decouple the problem of estimating the parameters of the column of $\mathbf{A}$. Additionally, it is not related to the MUSIC cost function in a simple way. It can, though, be simplified in the following way. The columns of $\mathbf{A}$ consist of complex sinusoids, and for any distinct set of frequencies these are asymptotically orthogonal, meaning that

$$\begin{aligned}\lim_{M\to\infty} M\mathbf{\Pi}_A &= \lim_{M\to\infty} M\mathbf{A}\left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{A}^H\\ &= \mathbf{A}\mathbf{A}^H.\end{aligned} \tag{22}$$

We can now simplify (21) and manipulate it into a familiar form, that is,

$$\begin{aligned}\|\mathbf{\Pi}_A\mathbf{\Pi}_G\|_F^2 &= \mathrm{Tr}\left\{\mathbf{A}\left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{A}^H\mathbf{G}\mathbf{G}^H\right\}\\ &\approx \frac{1}{M}\mathrm{Tr}\left\{\mathbf{A}^H\mathbf{G}\mathbf{G}^H\mathbf{A}\right\} = \frac{1}{M}\left\|\mathbf{A}^H\mathbf{G}\right\|_F^2,\end{aligned} \tag{23}$$

which, except for the scaling $1/M$, is the reciprocal of the original MUSIC cost function as introduced in [14]. From (19) and (23), we get

$$\frac{1}{M}\left\|\mathbf{A}^H\mathbf{G}\right\|_F^2 \approx \sum_{k=1}^{K}\cos^2(\theta_k). \tag{24}$$

This shows that the original MUSIC cost function can be explained and understood in the context of angles between subspaces. At this point, it must be emphasized that this interpretation only holds for signal models consisting of vectors that are orthogonal or asymptotically orthogonal. Consequently, it holds for sinusoids, for example, but not for damped sinusoids.

We now arrive at a convenient measure of the extent to which the orthogonality property in (10) holds, which is the average over all the principal (nontrivial) angles between $\mathcal{A}$ and $\mathcal{G}$:

$$\frac{1}{K}\sum_{k=1}^{K}\cos^2(\theta_k) = \frac{1}{K}\sum_{k=1}^{K}\sigma_k^2 \approx \frac{1}{MK}\left\|\mathbf{A}^H\mathbf{G}\right\|_F^2 \triangleq J \tag{25}$$

with $K = \min\{L, M-L\}$. This measure is only zero when all angles are $\pi/2$, that is, when the subspaces $\mathcal{A}$ and $\mathcal{B}$ are orthogonal in all directions. Additionally, the intersection of the subspaces is the range of the set of principal vectors for which $\cos(\theta_k) = 1$. Due to the normalization $1/K$, the measure can be seen to be bounded as

$$0 \le \frac{1}{K}\sum_{k=1}^{K}\cos^2(\theta_k) \le 1. \tag{26}$$

This bound is also asymptotically valid for the rightmost expression in (25) and is otherwise an approximation for finite lengths. To put the derived measure a bit into perspective, it can, in fact, be brought into a form similar as the aforementioned and well-known statistical methods (MDL, AIC, etc.) by taking the logarithm of (25), that is,

$$\ln J = \ln\left\|\mathbf{A}^H\mathbf{G}\right\|_F^2 - \ln(MK), \tag{27}$$

which consists of two familiar terms: a "goodness of fit" measure and an order-dependent penalty function, which in this case is a nonlinear function of the model order, unlike, for example, MDL and AIC.

3.3. Relation to Other Measures. We will now proceed to relate the derived measure to some other measures. Interestingly, the Frobenius norm of the difference between the two projection matrices can be expressed as

$$\begin{aligned}\|\mathbf{\Pi}_A - \mathbf{\Pi}_G\|_F^2 &= \mathrm{Tr}\{\mathbf{\Pi}_A + \mathbf{\Pi}_G - 2\mathbf{\Pi}_A\mathbf{\Pi}_G\} \\ &= M - 2\|\mathbf{\Pi}_A\mathbf{\Pi}_G\|_F^2,\end{aligned} \tag{28}$$

which shows that minimizing (18) is the same as maximizing the Frobenius norm of the difference between the two projection matrices. This puts the original MUSIC cost function into perspective, as it was originally motivated in [14] as the distance between the subspaces.

In [22], it was proposed to measure the orthogonality using the following normalized Frobenius norm of the matrix product $\mathbf{A}^H\mathbf{G}$:

$$\frac{||\mathbf{A}^H\mathbf{G}||_F^2}{LM(M-L)}, \tag{29}$$

which was derived from the Cauchy-Schwarz inequality. A new derivation of the measure in (29) is provided in the appendix in which it is shown that this too can be interpreted as an average over cosine to angles, more specifically, between each vector pair. However, the definition of the angles differs from that of the angles between subspaces, and, as a result, the normalizations differ as well. Clearly, we have that

$$ML(M-L) \geq M\min\{L, M-L\} \tag{30}$$

and thus

$$\frac{||\mathbf{A}^H\mathbf{G}||_F^2}{ML(M-L)} \leq \frac{||\mathbf{A}^H\mathbf{G}||_F^2}{M\min\{L, M-L\}}. \tag{31}$$

That the two approaches lead to different normalizations may seem like a minor detail, but this is in fact also the fundamental difference between the AIC, MDL, MAP, and so forth, order selection rules. These all provide a different order dependent scaling of the likelihood function. At the very least, the new normalization is mathematically more tractable than the old one. In Figure 1, the two normalizations, namely, $ML(M-L)$ and $M\min\{L, M-L\}$, are shown as a function of L for $M = 50$. Note that the curves have been scaled by their respective maximum values. Interestingly, both the measures defined in (29) and (25), respectively, are consistent with finding the frequencies using (13) in the sense that the frequencies that minimize (13) also minimize either of these measures for a given order L.

The measure in (25) can also be related to some other measures that have been defined in relation to angles between subspaces, like the projection 2-norm [37]. The distance or gap between subspaces, is defined for $L = M/2$ as [35–37]

$$\mathrm{dist}(\mathcal{A}, \mathcal{G}) = \|\mathbf{\Pi}_A - \mathbf{\Pi}_G\|_2 \tag{32}$$

and is related to the concept of angles between subspaces in the sense that (see, e.g., [35])

$$\mathrm{dist}(\mathcal{A}, \mathcal{G}) = \sin(\theta_K) = \sqrt{1 - \cos(\theta_K)}, \tag{33}$$

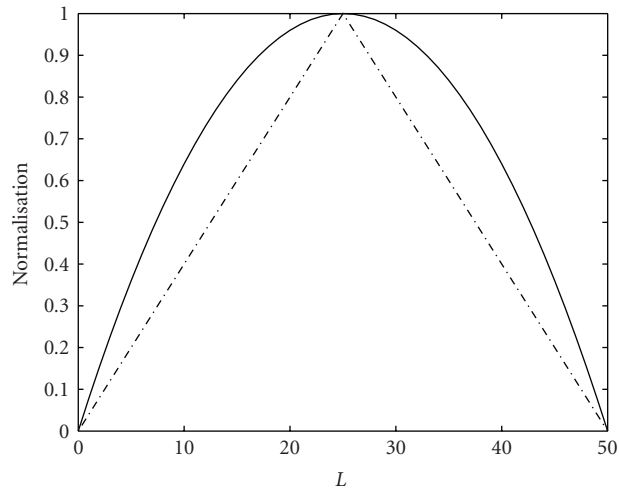


Figure 1: Normalization factors (scaled for convenience) as a function of L for the measure in [22] (solid) and based on the theory of angles between subspaces (dash-dotted).

which is given by the Kth singular value of the matrix product $\mathbf{\Pi}_A\mathbf{\Pi}_G$ as $\theta_K = \arccos(\sigma_K)$. Another measure of interest is the minimum principal angle which by the definition in (15) is a function of the maximum singular value as $\theta_1 = \arccos(\sigma_1)$ and is given by the induced matrix 2-norm, that is,

$$\|\mathbf{\Pi}_A\mathbf{\Pi}_G\|_2^2 = \sigma_1^2. \tag{34}$$

In the study of angles between subspaces, there has also been some interest in a different definition of the angle between two subspaces based on exterior algebra. Specifically, this so-called higher dimensional angle θ is related to the principal angles as [38, 39]

$$\cos^p(\theta) = \prod_{k=1}^{K}\cos^p(\theta_k) = \prod_{k=1}^{K}\sigma_k^p \tag{35}$$

for $p = 1, 2, \ldots$, which for $p = 1$ can be interpreted as the volume of a certain matrix [38]. In [40], θ was shown to be an angle in the usual Euclidean sense.

Equations (33), (34), and (35) are not very convenient measures for our purpose since they cannot be calculated from the individual columns of $\mathbf{A}$ but rather depend on all of them. This means that optimization of any of these measures would require multidimensional nonlinear optimization over the frequencies $\{\omega_l\}$.

We will now investigate how the various measures relate to each other, and in doing so, we will arrive at some interesting bounds. First, we note that the arithmetic mean of the singular values can be related to the geometric mean and (35) as

$$\frac{1}{K}\sum_{k=1}^{K}\sigma_k^2 \geq \left(\prod_{k=1}^{K}\sigma_k^2\right)^{1/K} \geq \prod_{k=1}^{K}\sigma_k^2, \tag{36}$$

where the right-most expression follows from $\sigma_k \leq 1$. We can now establish the following set of inequalities that relate

the various measures based on angles between subspaces to the Frobenius norm:

$$\prod_{k=1}^{K}\sigma_k^2 \leq \sigma_K^2 \leq \sigma_1^2 \leq \sum_{k=1}^{K}\sigma_k^2. \tag{37}$$

It follows that the Frobenius norm can be seen as a majorizing function for the other measures. Therefore, finding the frequencies using (12) can be seen to minimize the upper bound of the other measures. Similarly, we obtain the following set of inequalities for the normalized measure involving the average over the squared cosine terms in (19), that is,

$$\prod_{k=1}^{K}\sigma_k^2 \leq \sigma_K^2 \leq \frac{1}{K}\sum_{k=1}^{K}\sigma_k^2 \leq \sigma_1^2. \tag{38}$$

In this case, the normalized Frobenius norm is still an upper bound for two of the measures, but it is lower than or equal to the measure in (19). In this sense, the measure in (19) can be seen as a majorizing function for the measures in (33) and (35). It can be seen from (38) that the measures are identical when all singular values $\{\sigma_k\}$ are either one or zero, that is, when the subspaces have a K dimensional intersection or are orthogonal in all directions. The only measure, however, that ensures orthogonality in all directions for a value of zero, is the proposed measure in (25). Clearly, this is a desirable property for our application.

3.4. Application to Sinusoidal Order Estimation. As can be seen, (10) can only be expected to hold when the eigenvectors of $\mathbf{R}$ are partitioned into a signal and a noise subspace such that the rank of the signal subspace is equal to the true number of sinusoids. Based on the proposed orthogonality measure, the order is found by evaluating the measure for various candidate orders $1 \leq L \leq M-1$ and then picking the order for which the measure is minimized, that is,

$$\hat{L} = \arg\min_{L}\min_{\{\omega_l\}}\frac{||\mathbf{A}^H\mathbf{G}||_F^2}{MK} \tag{39}$$

$$= \arg\min_{L}\sum_{l=1}^{L}\min_{\omega_l}\frac{||\mathbf{a}^H(\omega_l)\mathbf{G}||_F^2}{MK} \tag{40}$$

with $K = \min\{L, M-L\}$. As before, the frequencies should be distinct and satisfy (14). The set of candidate orders does not include zero (as no angles can be measured then), meaning that the measure cannot be used for determining whether only noise is present. This is also the case for the related ESTER and SAMOS methods.

3.4.1. Consistency. Regarding the consistency of the proposed method, it can easily be verified that the covariance matrix model and the orthogonality property hold for the noise-free case. We will here make the following simple argument for the consistency of the method for noisy signals based on [31]: since a consistent estimate of the covariance matrix is used, the eigenvector estimates are consistent too and the covariance matrix model in (3) holds asymptotically in N and M (which is here assumed to be chosen proportional to N) [31, 32]. Therefore, the orthogonality criterion in (10) holds as N tends to infinity. Provided that the sinusoids are linearly independent but not all orthogonal, (10) holds only for the combination of the true set of frequencies $\{\omega_l\}$ and order L. Regarding the finite length performance of MUSIC, it is well known to perform well for high SNR and N being consistent but suboptimal [31, 32] while exhibiting thresholding behavior below certain SNR or number of samples N. This thresholding behavior can largely be attributed to the occurrence of "subspace swapping" [41, 42].

3.4.2. Computational Complexity. The major contributor to the computational complexity of a direct implementation of (40) is the EVD of the covariance matrix, and this is also the case for the ESTER and SAMOS methods and [18, 19]. This can be lessened by the use of recursive computation of the covariance matrix eigenvectors over time, also known as subspace tracking. However, for our method and the ESTER and SAMOS methods, it is critical that a subspace tracker is chosen that tracks the eigenvectors and not just an arbitrary basis of the subspace. The reasons is that a subpartitioning of an arbitrary basis is not necessarily the same as a subpartitioning of the eigenvectors and the methods may therefore fail to provide accurate order estimates. Examples of subspace trackers that are suited for this purpose are, for example, [43–45] (see [46] for more on this). Aside from the EVD, our method also requires nonlinear optimization for finding the frequencies. This is by no means a particular property of our methodl; indeed most other methods for finding the order of the model in (1), including [4, 10–13], require this as well, with the methods of [19, 20] being notable exceptions. For (40), this can be done either by FFTs (see [22, 46]) or by polynomial rooting methods [28]. In the FFT-based implementation of [22], the Fourier transform of the eigenvectors is calculated once per segment and this information is simply reused in the subsequent optimization. The complexity is therefore similar to that of spectral [14] or root MUSIC [28], two methods that have a rich history in spectral estimation and array processing. In practice, the complexity can be reduced considerably by applying certain approximations, that is, by either (1) using the min-norm solution, which can be calculated recursively over the orders, instead of the full noise subspace [5, 47], or by (2) finding approximate solutions using a number of the least significant eigenvectors that are known with certainty to belong to the noise subspace (usually an upper bound on number of possible sinusoids can be identified from the application).

3.4.3. Comparison to ESTER and SAMOS. There appears to be a number of advantages to our method compared to the related methods ESTER and SAMOS that are also based on the eigenvectors. It can be seen from (40) that the method can find orders in a wider range than both the ESTER and SAMOS methods, with those methods being able to find orders in the intervals $1 \leq L \leq M-2$ and

$1 \leq L \leq (M-1)/2$, respectively. The class of shift-invariant signal models also includes damped sinusoids and the ESTER and SAMOS methods hold also for this model and so does the orthogonality property of MUSIC. At first sight it may appear that an efficient implementation of the nonlinear optimization in (40) does not exist. However, either the rooting approach of [28] may be used or the principle of unitary ESPRIT can be applied by using a forward-backward estimate of the covariance matrix whereby the FFT-based implementation is applicable (see [24]). We here stress that an additional advantage of the MUSIC-based method presented here is that it is more general than those based on the shift-invariance property [20, 21]; that is, the relation (10) can be used for a more general class of signal models. It is, however, not certain that there exits an efficient implementation of the nonlinear optimization required by this approach.

4. Experimental Results

4.1. Details and Reference Methods. We now proceed to evaluate the performance of the proposed estimator (denoted MUSIC (new) in the figures) under various conditions using Monte Carlo simulations comparing to a number of other methods that have appeared in literature. The reference methods are listed in Table 1. It should be noted that the model selection criteria of the MDL [13] and the MAP [4] methods are in fact identical for this problem, although derived from different perspectives. The difference between these two methods is then, essentially, that one uses high-resolution estimates of the frequencies while the other uses the computationally simple periodogram. Note that it is possible to refine the initial frequency estimates obtained from the periodogram in several ways, for example, [48, 49], but to retain the computational simplicity, we refrain from doing this here.

In the experiments, signals are generated according to the model in (1) with Gaussian noise. Furthermore, all amplitudes are set to unity, that is, $A_l = 1$ for all l and the signal-to-noise ratio (SNR) is defined as $\text{SNR} = 10\log_{10}(\sum_{l=1}^{L} A_l^2/\sigma^2)$ [dB]. Note that similar results have been obtained for other amplitude distributions. For example, the general conclusions are the same for a Rayleigh pdf, but in the interest of brevity we will focus on the simple case of unit amplitudes. The sinusoidal phases and frequencies are generated according to a uniform pdf in the interval $(-\pi, \pi]$ which will result in spectrally overlapping sinusoids sometimes. For each combination of the parameters, 500 Monte Carlo simulations were run. Unless otherwise stated, we will use $L = 5$ and $M = N/2$.

4.2. Statistical Evaluation. First, we will evaluate the performance in terms of the percentage of correctly estimated orders under various conditions. We start out by varying the number of observations N while keeping the SNR fixed at 20 dB and then we will keep N fixed at 200 while varying the SNR. The partitioning of the EVD into signal and noise subspaces in (7) and (8) depends on the sorting of the eigenvalues resulting in the right ordering of the eigenvectors. As a result, the performance of the methods is expected to depend on the SNR. The results are shown in Figures 2 and 3. Next, we evaluate the performance as a function of the true model order for $N = 100$ and SNR = 20 dB. Note that the choice of M also limits the number of possible sinusoids that can be found using MUSIC since $M > L$. The results are depicted in Figure 4. An experiment to investigate the dependency of the performance on the choice of M while keeping $N = 100$ constant has also been

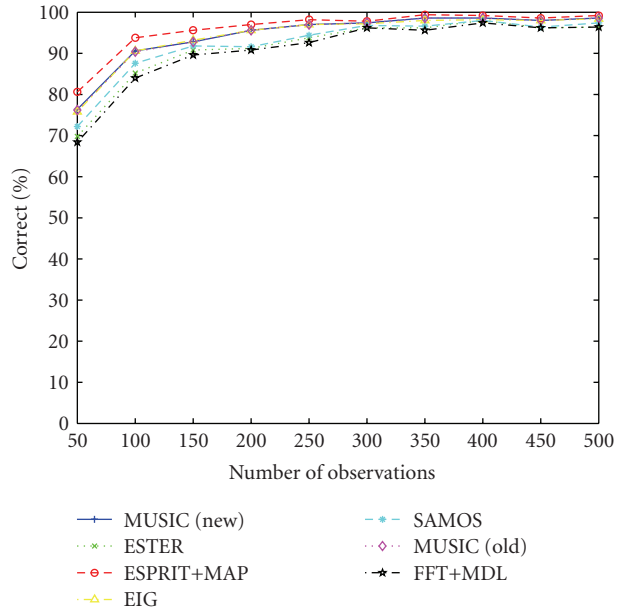


Figure 2: Percentage of correctly estimated model orders as a function of the number of observations for an SNR of 20 dB.

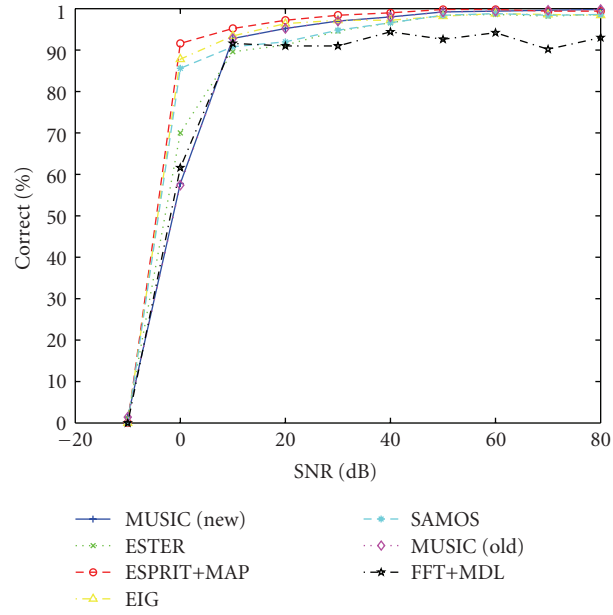


Figure 3: Percentage of correctly estimated model orders versus the SNR for $N = 200$.

Table 1: List of reference methods used in the experiments with short descriptions and references to literature.

Name	Reference	Description
ESTER	[20]	Subspace-based method based on the shift-invariance property of the signal model
ESPRIT+MAP	[4, 16]	Frequencies estimated using ESPRIT, amplitudes using least-squares, model selection using the MAP criterion
EIG	[19]	Method based on the ratio between the arithmetic and geometric means of the eigenvalues
SAMOS	[21]	Same as ESTER except for measure
MUSIC (old)	[22, 23]	Same as the proposed method except for the normalization
FFT+MDL	[1, 12, 13]	Statistical method based on MDL, with parameters estimated using the periodogram

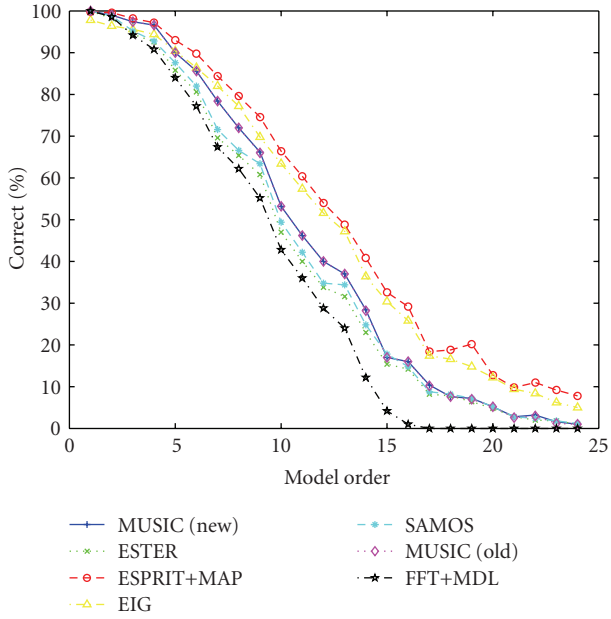


Figure 4: Percentage of correctly estimated model orders as a function of the true order with SNR = 20 dB and $N = 100$.

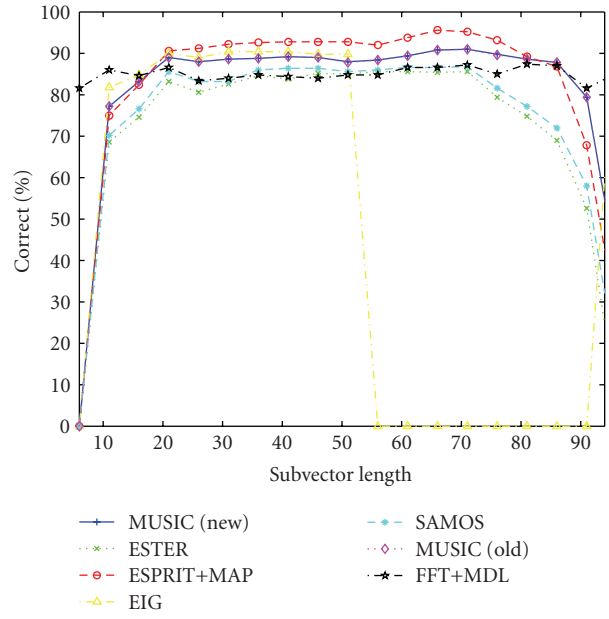


Figure 5: Percentage of correctly estimated model orders as a function of subvector length with SNR = 20 dB and $N = 100$.

conducted with an SNR of 20 dB. The results are shown in Figure 5. The reason that the method of [19] fails here is that the covariance matrix is rank deficient for $M > N/2$. This can of course easily be fixed by modifying the range over which the geometric and arithmetic means of the eigenvalues are calculated. Since the gap between the signal and noise subspace eigenvalues depends not only on the SNR but also on how closely spaced the sinusoids are in frequency, the importance of the difference in frequency between the sinusoids will now be investigated. We do this by distributing the frequencies evenly as $2\pi\Delta l$ and then vary Δ for $L = 5$ sinusoids, $N = 100$, $M = 25$, and an SNR of 20 dB. All other experimental conditions are as described earlier. The results are shown in Figure 6. In a final experiment, we illustrate the applicability of the estimators in the presence of colored Gaussian noise. The percentages of correctly estimated orders are shown in Figure 7 as a function of the SNR. To generate the colored noise, a second-order autoregressive process was used having the transfer function $H(z) = 1/(1 - 0.25z^{-1} + 0.5z^{-2})$. Other than the noise color, the experimental conditions are the same as for Figure 3, that is, with $N = 200$. Note that for a fair comparison, the white noise model selection criterion has been used for all the

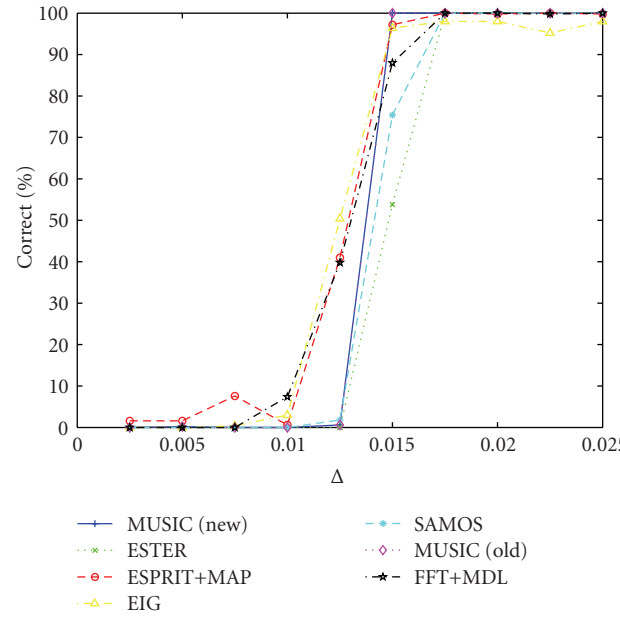


Figure 6: Percentage of correctly estimated model orders as a function of the difference between frequencies distributed uniformly as $2\pi\Delta l$ with SNR = 20 dB and $N = 100$.

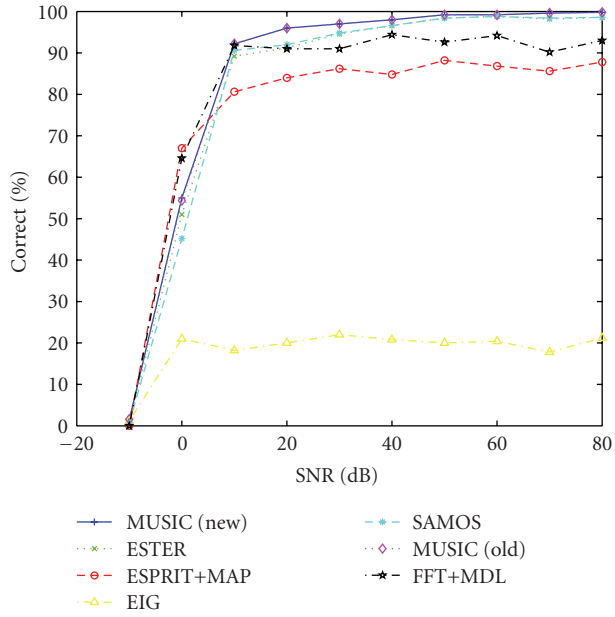


Figure 7: Percentage of correctly estimated model orders as a function of the SNR for colored Gaussian noise for $N = 200$.

methods. In other words, this experiment can be seen as an evaluation of the sensitivity to the white noise assumption. It is of course possible to modify the methods to take the colored noise into account in various ways, one way that can be applied to all the methods being prewhitening [18], but all such ways require that the statistics of the noise be known.

5. Discussion

From the experiments the following general observations can be made. First of all, it can be observed that, with one exception, all the methods exhibit the same dependencies on the tested variables, although they sometimes exhibit quite different thresholding behavior. The one exception is for colored Gaussian noise. It can be seen from these figures that the proposed estimator has the desirable properties that the performance improves as the SNR and/or the number of observations increases and that the model order can be determined with high probability for a high SNR and/or a high number of observations, and this is generally the case of all the tested methods. MUSIC can also be observed to consistently outperform the other subspace methods based on the eigenvectors, namely, ESTER and SAMOS. Curiously, the new MUSIC criterion performs similarly to the old one in all the simulations, which indicates that the orthogonality criterion does not depend strongly on the normalization. The MAP criterion of [4] combined with ESPRIT and the method based on the eigenvalues [19] can be seen to generally perform the best, outperforming the measure based on angles between subspaces when the noise is white Gaussian. This is, most likely, due to these methods making use of the assumption that the noise is not only white but also Gaussian; this assumption is not used in the proposed method. Despite their good performance for white Gaussian noise, both aforementioned methods appear to be rather sensitive to the white noise assumption and their performance is rather poor for colored noise. The poor performance of the eigenvalue-based method of [19] for colored noise is no surprise. In fact, for colored noise, the method of [19] can be shown to overestimate the model order with probability 1 [50, 51]. That the MAP criterion in combination with ESPRIT outperforms the method of [13] can only be attributed to the former method resulting in superior parameter estimates to the periodogram, which will fail to resolve adjacent sinusoids for a low number of samples. We observe from Figure 4 that the performance of all the methods deteriorates as the number of parameters approaches M. That the MAP-based method fails in this case cannot be solely attributed to the MAP rule since it relies on sinusoidal parameter estimates being accurate. However, the MAP rule was derived in [4] based on the assumption that the likelihood function is highly peaked around the parameters estimates, which is usually the case when N is high relative to the number of parameters. We have observed from order estimation error histograms that while the orders are not estimated correctly for high orders, the estimated order is still generally close to the true one and may thus still be useful. From Figure 5, it appears that the methods are not very sensitive to the choice of M as long as it is not chosen too low or too high, that is, not too close to either L or N.

6. Conclusion and Future Work

In this paper, we have considered the problem of finding the number of complex sinusoids in white noise, and a new measure for solving this problem has been derived based on angles between the noise subspace and the candidate model. The measure is essentially the mean of the cosine to all nontrivial angle squared, which is asymptotically closely related to the original MUSIC cost function as defined for direction-of-arrival and frequency estimation. The derivations in this paper put order estimation using the orthogonality property of MUSIC on a firm mathematical ground. Numerical simulations show that the correct order can be determined for a high number of observations and/or a high signal-to-noise ratio (SNR) with a high probability. Additionally, experiments show that the performance of the proposed method exhibits the same functional dependencies on the SNR, the number of observations, and the model order as statistical methods. The experiments showed that the proposed method outperforms other previously published subspace methods and that the method is more robust to the noise being colored than all the other methods. Future work includes a rigorous statistical analysis of the proposed method along the lines of [33].

Appendix

Alternative Derivation of the Old Measure

We will now derive the normalized MUSIC cost function first proposed in [22] for finding the number of sinusoids.

Note that this derivation differs from the one in [22]. The following can be established for the acute angle $0 \leq \theta_{l,m} \leq \pi/2$ between two vectors $\mathbf{a}(\omega_l)$ and $\mathbf{q}_m$:

$$\cos^2\theta_{l,m} = \frac{\left|\mathbf{a}^H(\omega_l)\mathbf{q}_m\right|^2}{\|\mathbf{a}(\omega_l)\|_2^2\|\mathbf{q}_m\|_2^2}. \tag{A.1}$$

Averaging over $\cos^2\theta_{l,m}$ for all vector pairs, we get

$$\begin{aligned} J &= \frac{1}{L(M-L)}\sum_{l=1}^{L}\sum_{m=L+1}^{M}\cos^2\theta_{l,m} \\ &= \frac{1}{L(M-L)}\sum_{l=1}^{L}\sum_{m=L+1}^{M}\frac{\left|\mathbf{a}^H(\omega_l)\mathbf{q}_m\right|^2}{\|\mathbf{a}(\omega_l)\|_2^2\|\mathbf{q}_m\|_2^2}. \end{aligned} \tag{A.2}$$

Noting that all the columns of $\mathbf{A}$ and $\mathbf{G}$ have the same norms, this can be written as

$$\begin{aligned} J &= \sum_{l=1}^{L}\sum_{m=L+1}^{M}\frac{\left|\mathbf{a}^H(\omega_l)\mathbf{q}_m\right|^2}{L\|\mathbf{a}(\omega_l)\|_2^2(M-L)\|\mathbf{q}_m\|_2^2} \\ &= \frac{\|\mathbf{A}^H\mathbf{G}\|_F^2}{\|\mathbf{A}\|_F^2\|\mathbf{G}\|_F^2} = \frac{\|\mathbf{A}^H\mathbf{G}\|_F^2}{LM(M-L)}, \end{aligned} \tag{A.3}$$

and, clearly, we have the following inequalities:

$$0 \leq \frac{\|\mathbf{A}^H\mathbf{G}\|_F^2}{LM(M-L)} \leq 1, \tag{A.4}$$

which also follow from the Cauchy-Schwartz inequality. The orthogonality measure in (A.3) has the desirable properties that it facilitates optimization over the individual columns of $\mathbf{A}$ and is invariant to the dimensions of the matrices. This measure is different than the original measure proposed in [14] due to the scaling of the cost function. Note that the MUSIC cost function originally was introduced as the reciprocal of the Euclidean distance between the signal model vectors and the signal subspace.

Acknowledgments

This research was supported in part by the Parametric Audio Processing project, Danish Research Council for Technology and Production Sciences Grant no. 274-06-0521.

References

[1] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[2] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[3] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[4] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Transactions on Signal Processing*, vol. 46, no. 10, pp. 2726–2735, 1998.

[5] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Pearson Prentice Hall, Englewood Cliffs, NJ, USA, 2005.

[6] E. G. Larsson and Y. Selen, "Linear regression with a sparse parameter vector," *IEEE Transactions on Signal Processing*, vol. 55, no. 2, pp. 451–460, 2007.

[7] J.-J. Fuchs, "Estimating the number of sinusoids in additive white noise," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 12, pp. 1846–1853, 1988.

[8] J.-J. Fuchs, "Estimation of the number of signals in the presence of unknown correlated sensor noise," *IEEE Transactions on Signal Processing*, vol. 40, no. 5, pp. 1053–1061, 1992.

[9] A. T. James, "Test of equality of latent roots of a covariance matrix," in *Multivariate Analysis*, pp. 205–218, 1969.

[10] B. G. Quinn, "Estimating the number of terms in a sinusoidal regression," *Journal of Time Series Analysis*, vol. 10, no. 1, pp. 71–75, 1989.

[11] X. Wang, "An AIC type estimator for the number of cosinusoids," *Journal of Time Series Analysis*, vol. 14, no. 4, pp. 434–440, 1993.

[12] L. Kavalieris and E. J. Hannan, "Determining the number of terms in a trigonometric regression," *Journal of Time Series Analysis*, vol. 15, no. 6, pp. 613–625, 1994.

[13] E. J. Hannan, "Determining the number of jumps in a spectrum," in *Developments in Time Series Analysis*, pp. 127–138, Chapman and Hall, London, UK, 1993.

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

[15] G. Bienvenu, "Influence of the spatial coherence of the background noise on high resolution passive methods," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '79)*, pp. 306–309, April 1979.

[16] R. Roy and T. Kailath, "ESPRIT—estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[17] V. F. Pisarenko, "The retrieval of harmonics from a covariance function," *Geophysical Journal of the Royal Astronomical Society*, vol. 33, pp. 347–366, 1973.

[18] G. Bienvenu and L. Kopp, "Optimality of high resolution array processing using the eigensystem approach," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 31, no. 5, pp. 1235–1248, 1983.

[19] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985.

[20] R. Badeau, B. David, and G. Richard, "A new perturbation analysis for signal enumeration in rotational invariance techniques," *IEEE Transactions on Signal Processing*, vol. 54, no. 2, pp. 450–458, 2006.

[21] J.-M. Papy, L. De Lathauwer, and S. van Huffel, "A shift invariance-based order-selection technique for exponential data modelling," *IEEE Signal Processing Letters*, vol. 14, no. 7, pp. 473–476, 2007.

[22] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1635–1644, 2007.

[23] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Sinusoidal order estimation using the subspace orthogonality and shift-invariance properties," in *Proceedings of the 41st Asilomar Conference on Signals, Systems and Computers (ACSSC '07)*, pp. 651–655, Pacific Grove, Calif, USA, November 2007.

[24] M. Haardt and J. A. Nossek, "Unitary ESPRIT: how to obtain increased estimation accuracy with a reduced computational

burden," *IEEE Transactions on Signal Processing*, vol. 43, no. 5, pp. 1232–1242, 1995.

[25] P. C. Hansen and S. H. Jensen, "Prewhitening for rank-deficient noise in subspace methods for noise reduction," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3718–3726, 2005.

[26] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: survey and analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 92953, 24 pages, 2007.

[27] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, vol. 5 of *Synthesis Lectures on Speech & Audio Processing*, Morgan & Claypool, San Rafael, Calif, USA, 2009.

[28] A. J. Barabell, "Improving the resolution performance of eigenstructure-based direction-finding algorithms," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '83)*, vol. 1, pp. 336–339, Boston, Mass, USA, April 1983.

[29] H. Saarnisaari, "Robustness of the MUSIC algorithm to errors in estimation the dimensions of the subspaces: delay estimation in DS/SS in the presence of interference," in *Proceedings of IEEE Military Communications Conference (MILCOM '99)*, vol. 2, pp. 851–854, 1999.

[30] A. Eriksson, P. Stoica, and T. Töderström, "Asymptotical analysis of MUSIC and ESPRIT frequency estimates," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '93)*, vol. 4, pp. 556–559, Minneapolis, Minn, USA, April 1993.

[31] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramer-Rao bound," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 5, pp. 720–741, 1989.

[32] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramer-Rao bound: further results and comparisons," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, no. 12, pp. 2140–2150, 1990.

[33] B. G. Quinn and E. J. Hannan, *The Estimation and Tracking of Frequency*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, UK, 2001.

[34] D. R. Farrier, D. J. Jeffries, and R. Mardani, "Theoretical performance prediction of the music algorithm," *IEE Proceedings F: Communications, Radar and Signal Processing*, vol. 135, no. 3, pp. 216–224, 1988.

[35] G. H. Golub and C. F. V. Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, Md, USA, 3rd edition, 1996.

[36] A. V. Knyazev and M. E. Argentati, "Principal angles bewteen subspaces in an a-based sclar product: algorithms and pertubation estimates," *SIAM Journal on Scientific Computing*, vol. 23, no. 6, pp. 2009–2041, 2002.

[37] P.-A. Absil, A. Edelman, and P. Koev, "On the largest principal angle between random subspaces," *Linear Algebra and Its Applications*, vol. 414, no. 1, pp. 288–294, 2006.

[38] J. Miao and A. Ben-Israel, "On principal angles between subspaces in $\mathbb{R}^n$," *Linear Algebra and Its Applications*, vol. 171, pp. 81–98, 1992.

[39] J. Miao and A. Ben-Israel, "Product cosines of angles between subspaces," *Linear Algebra and Its Applications*, vol. 237-238, pp. 71–81, 1996.

[40] S. Jiang, "Angles between euclidean subspaces," *Geometriae Dedicata*, vol. 63, no. 2, pp. 113–121, 1996.

[41] M. Hawkes, A. Nehorai, and P. Stoica, "Performance breakdown of subspace-based methods: prediction and cure," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 6, pp. 4005–4008, Salt Lake, Utah, USA, May 2001.

[42] J. K. Thomas, L. L. Scharf, and D. W. Tufts, "The probability of a subspace swap in the SVD," *IEEE Transactions on Signal Processing*, vol. 43, no. 3, pp. 730–736, 1995.

[43] B. Yang, "Projection approximation subspace tracking," *IEEE Transactions on Signal Processing*, vol. 43, no. 1, pp. 95–107, 1995.

[44] D. J. Rabideau, "Fast, rank adaptive subspace tracking and applications," *IEEE Transactions on Signal Processing*, vol. 44, no. 9, pp. 2229–2244, 1996.

[45] P. Strobach, "Low-rank adaptive filters," *IEEE Transactions on Signal Processing*, vol. 44, no. 12, pp. 2932–2947, 1996.

[46] J. R. Jensen, J. K. Nielsen, M. G. Christensen, S. H. Jensen, and T. Larsen, "On fast implementation of harmonic MUSIC for known and unknown model orders," in *Proceedings of the 16th European Signal Processing Conferenc (EURASIP '08)*, Lausanne, Switzerland, August 2008.

[47] R. Kumaresan and D. W. Tufts, "Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 30, no. 6, pp. 833–840, 1982.

[48] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 4, pp. 477–489, 1988.

[49] B. G. Quinn and J. M. Fernandes, "A fast efficient technique for the estimation of frequency," *Biometrika*, vol. 78, no. 3, pp. 489–497, 1991.

[50] W. Xu and M. Kaveh, "Analysis of the performance and sensitivity of eigendecomposition-based detectors," *IEEE Transactions on Signal Processing*, vol. 43, no. 6, pp. 1413–1426, 1995.

[51] A. P. Liavas and P. A. Regalia, "On the behavior of information theoretic criteria for model order selection," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1689–1695, 2001.

Paper B

Optimal Filter Designs for Separating and Enhancing Periodic Signals

M. G. Christensen and A. Jakobsson,

The paper has been published in the
IEEE Transactions on Signal Processing, vol. 58, no. 12, pp. 5969–5983, 2010.

Optimal Filter Designs for Separating and Enhancing Periodic Signals

Mads Græsbøll Christensen* and Andreas Jakobsson

***Abstract*— In this paper, we consider the problem of separating and enhancing periodic signals from single-channel noisy mixtures. More specifically, the problem of designing filters for such tasks is treated. We propose a number of novel filter designs that 1) are specifically aimed at periodic signals, 2) are optimal given the observed signal and thus signal-adaptive, 3) offer full parametrizations of periodic signals, and 4) reduce to well-known designs in special cases. The found filters can be used for a multitude of applications including processing of speech and audio signals. Some illustrative signal examples demonstrating its superior properties as compared to other related filters are given and the properties of the various designs are analyzed using synthetic signals in Monte Carlo simulations.**

I. Introduction

Many natural signals that are of interest to mankind are periodic by nature or approximately so. In mathematics and engineering sciences, such periodic signals are often described by Fourier series, i.e., a sum of sinusoids, each described by an amplitude and a phase, having frequencies that are integer multiples of a fundamental frequency. In mathematical descriptions of periodic functions, the period which is inversely proportional to the fundamental frequency is assumed to be known and the function is observed over a single period over which the sinusoids form an orthogonal basis. When periodic signals are observed over arbitrary intervals, generally have unknown fundamental frequencies, and are corrupted by some form of observation noise, the problem of parametrizing the signals is a different and much more difficult one. The problem of estimating the fundamental frequency from such an observed signal is referred to as fundamental frequency or pitch estimation. Additionally, some signals contain many such periodic signals, in which case the problem is referred to as multi-pitch estimation. Strictly speaking, the word pitch originates in the perception of acoustical signals and is defined as "that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale" [1], but since this attribute in most cases is the same as the fundamental frequency of a Fourier series, these terms are often used synonymously. Some pathological examples do exist, however, where it is not quite that simple. The pitch estimation problem has received much attention in the fields of speech and audio processing, not just because it is an interesting and challenging problem, but also because it is the key, or, perhaps more correctly, a key to many fundamental problems such as separation of periodic sources [2], enhancement, and compression of periodic sources [3] as Fourier series constitute naturally compact descriptions of such signals. A fundamental problem in signal processing is the source separation problem, as many other problems are trivially, or at least more easily, solved once a complicated mixture has been broken into its basic parts (for examples of this, see [4], [5]). We remark that for periodic signals, this problem is different from that of blind source separation, as assumptions have been made as to the nature of the sources (for an overview of classical methods for blind source separation, see, e.g., [6], [7]). For periodic signals, once the fundamental frequencies of the periodic sources have been found, it is comparably easy to estimate either the individual periodic signals directly [8]–[11] or their remaining unknown parameters, i.e., the amplitudes, using methods like those in [12]. With amplitudes and the fundamental frequency found, the signal parametrization is complete. Some representative methodologies that have been employed in fundamental frequency estimators are: linear prediction [13], correlation [14], subspace methods [15]–[17], harmonic fitting [18], maximum likelihood [19], [20], cepstral methods [21], Bayesian estimation [22]–[24], and comb filtering [8], [25], [26]. Several of these methodologies can be interpreted in several ways and one should therefore not read too much into this rather arbitrary grouping of methods. For an overview of pitch estimation methods and their relation to source separation, we refer the interested reader to [27]. It should also be noted that separation based on parametric models of the sources is closely related to source separation using sparse decompositions (for an example of such an approach, see [28]).

The scope of this paper is filtering methods with application to periodic signals in noise. We propose a number of novel filter design methods, which are aimed specifically at the processing of noisy observations of periodic signals or from single-channel mixtures of periodic signals. These filter design methods result in filters that are optimal given the observed signal, i.e., they are signal-adaptive, and contain as special cases several well-known designs. The proposed filter designs are inspired by the principle used in the Amplitude and Phase EStimation (APES) method [29], [30], a method which is well-known to have several advantages over the Capon-based estimators. The obtained filters can be used for a number of tasks involving periodic signals, including separation, enhancement, and parameter estimation. In other words, the filtering approaches proposed herein provide full

Part of this work was presented at the 43rd Annual Asilomar Conference on Signals, Systems, and Computers.

M. G. Christensen is with the Dept. of Architecture, Design and Media Technology, Aalborg University, Niels Jernes Vej 14, DK-9220 Aalborg, Denmark (phone: 99 40 97 93, email: mgc@imi.aau.)

A. Jakobsson is with the Dept. of Mathematical Statistics, Lund University, SE-221 00 Lund, Sweden (phone: +46 46 222 4520, email: andreas.jakobsson@ieee.org).

parametrizations of periodic signals through the use of filters. We will, however, focus on the application of such filters to extraction, separation, and enhancement of periodic signals. A desirable feature of the filters is that they do not require prior knowledge of the noise or interfering source but are able to automatically reject these.

The paper is organized as follows. In Section II, we introduce the fundamentals and proceed to derive the initial design methodology leading to single filter that is optimal given the observed signal in Section III. We then derive an alternative design using a filterbank in Section IV, after which, in Section V, we first illustrate the properties of the proposed design and compare the resulting filters to those obtained using previously published methods. Moreover, we demonstrate its application for the extraction of real quasi-periodic signals from mixtures of interfering periodic signals and noise, i.e., for separation and enhancement. Finally, we conclude on the work in Section VI.

II. Fundamentals

We define a model of a signal containing a single periodic component, termed a source, consisting of a weighted sum of complex sinusoids having frequencies that are integer multiples of a fundamental frequency[1] ω_k, and additive noise. Such a signal can, for $n = 0, \ldots, N-1$, be written as

$$x_k(n) = \sum_{l=1}^{L_k} a_{k,l} e^{j\omega_k l n} + e_k(n) \tag{1}$$

where $a_{k,l} = A_{k,l} e^{j\phi_{k,l}}$ is the complex amplitude of the lth harmonic of the source (indexed by k) and $e_k(n)$ is the noise which is assumed to be zero-mean and complex. The complex amplitude is composed of a real, non-zero amplitude $A_{k,l} > 0$ and a phase $\phi_{k,l}$ distributed uniformly on the interval $\{-\pi, \pi]$. The number of sinusoids, L_k, is referred to as the order of the model and is often considered known in the literature. We note that this assumption is generally not consistent with the behavior of speech and audio signals, where the number of harmonics can be observed to vary over time. In most recordings of music, the observed signal consists of many periodic signals, in which case the signal model is

$$x(n) = \sum_{k=1}^{K} x_k(n) = \sum_{k=1}^{K} \sum_{l=1}^{L_k} a_{k,l} e^{j\omega_k l n} + e(n). \tag{2}$$

Note that all noise sources $e_k(n)$ are here modeled by a single noise source $e(n)$. We refer to signals of the form (2) as multi-pitch signals and the model as the multi-pitch model. Even if a recording is only of a single instrument, the signal may be multi-pitch as only some instruments are monophonic. Even in that case, room reverberation may cause the observed signal to consist of several different tones at a particular time, i.e., the signal is effectively a multi-pitch signal.

The algorithms under consideration operate on vectors consisting of M time-reversed samples of the observed signal, defined as $\mathbf{x}(n) = [\, x(n)\ x(n-1)\ \cdots\ x(n-M+1)\,]^T$, where $M \leq N$ and $(\cdot)^T$ denotes the transpose, and similarly for the sources $x_k(n)$ and the noise $e(n)$. Defining the filter output $y_k(n)$ as

$$y_k(n) = \sum_{m=0}^{M-1} h_k(m) x(n-m), \tag{3}$$

and introducing $\mathbf{h}_k = [\, h_k(0)\ \cdots\ h_k(M-1)\,]^H$, we can express the output of the filter as $y_k(n) = \mathbf{h}_k^H \mathbf{x}(n)$, with $(\cdot)^H$ being the Hermitian transpose operator. The expected output power can thus be expressed as

$$\mathrm{E}\left\{|y_k(n)|^2\right\} = \mathrm{E}\left\{\mathbf{h}_k^H \mathbf{x}(n) \mathbf{x}^H(n) \mathbf{h}_k\right\} \tag{4}$$

$$= \mathbf{h}_k^H \mathbf{R} \mathbf{h}_k, \tag{5}$$

where $\mathrm{E}\{\cdot\}$ denotes the statistical expectation. The above expression can be seen to involve the covariance matrix defined as $\mathbf{R} = \mathrm{E}\left\{\mathbf{x}(n)\mathbf{x}^H(n)\right\}$. We will now analyze the covariance matrix a bit more in detail.

The signal model in (2) can now be written using the above definitions as

$$\mathbf{x}(n) = \sum_{k=1}^{K} \mathbf{Z}_k \begin{bmatrix} e^{-j\omega_k 1 n} & & 0 \\ & \ddots & \\ 0 & & e^{-j\omega_k L_k n} \end{bmatrix} \mathbf{a}_k^* + \mathbf{e}(n) \tag{6}$$

$$\triangleq \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{a}_k^*(n) + \mathbf{e}(n) \tag{7}$$

or, alternatively, as $\mathbf{x}(n) \triangleq \sum_{k=1}^{K} \mathbf{Z}_k(n) \mathbf{a}_k^* + \mathbf{e}(n)$. Here, $\mathbf{Z}_k \in \mathbb{C}^{M \times L_k}$ is a Vandermonde matrix, being constructed from L_k harmonically related complex sinusoidal vectors as $\mathbf{Z}_k = [\, \mathbf{z}(\omega_k)\ \cdots\ \mathbf{z}(\omega_k L_k)\,]$, with $\mathbf{z}(\omega) = [\, 1\ e^{-j\omega}\ \cdots\ e^{-j\omega(M-1)}\,]^T$, and $\mathbf{a}_k = [\, a_{k,1}\ \cdots\ a_{k,L_k}\,]^H$ is a vector containing the complex amplitudes. Introducing $z_k = e^{-j\omega_k}$, the structure of the matrix $\mathbf{Z}_k$ can be seen to be

$$\mathbf{Z}_k = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_k^1 & z_k^2 & \cdots & z_k^{L_k} \\ \vdots & \vdots & \ddots & \vdots \\ z_k^{(M-1)} & z_k^{(M-1)2} & \cdots & z_k^{(M-1)L_k} \end{bmatrix}. \tag{8}$$

From this, it can be observed that either the complex amplitude vector or the Vandermonde matrix can be thought of as time-varying quantities, i.e., $\mathbf{a}_k^*(n) = \mathbf{D}^n \mathbf{a}_k^*$ and $\mathbf{Z}_k(n) = \mathbf{Z}_k \mathbf{D}^n$ with

$$\mathbf{D}^n = \begin{bmatrix} e^{-j\omega_k 1 n} & & 0 \\ & \ddots & \\ 0 & & e^{-j\omega_k L_k n} \end{bmatrix}, \tag{9}$$

meaning that the time index n can be seen as either changing the sinusoidal basis or, equivalently, the phases of the sinusoids. Depending on the context, one perspective may be more appropriate or convenient than the other.

For statistically independent sources, the covariance matrix of the observed signal can be written as $\mathbf{R} = \sum_{k=1}^{K} \mathbf{R}_k = \sum_{k=1}^{K} \mathrm{E}\left\{\mathbf{x}_k(n)\mathbf{x}_k^H(n)\right\}$, i.e., as a summation of the covariance matrices of the individual sources. By inserting the

[1] For many signals, the frequencies of the harmonics will not be exact integer multiples of the fundamental. This can be handled in several ways by modifying the signal model (see, e.g., [27] for more on this), but this is beyond the scope of this paper and will not be discussed any further.

single-pitch signal model in this expression, we can express the covariance matrix of the multi-pitch signal $\mathbf{x}(n)$ as

$$\mathbf{R} = \sum_{k=1}^{K} \mathbf{Z}_k \mathrm{E}\left\{\mathbf{a}_k^*(n)\mathbf{a}_k^T(n)\right\}\mathbf{Z}_k^H + \mathrm{E}\left\{\mathbf{e}_k(n)\mathbf{e}_k^H(n)\right\} \quad (10)$$

$$= \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{P}_k \mathbf{Z}_k^H + \mathbf{Q}, \quad (11)$$

where the matrix $\mathbf{P}_k$ is the covariance matrix of the amplitudes, i.e., $\mathbf{P}_k = \mathrm{E}\left\{\mathbf{a}_k^*(n)\mathbf{a}_k^T(n)\right\}$. For statistically independent and uniformly distributed stochastic phases (on the interval $(-\pi, \pi]$), this matrix reduces to a diagonal matrix having the power of the sinusoidal components on the diagonal, i.e., $\mathbf{P}_k = \mathrm{diag}\left(\left[\begin{array}{ccc} A_{k,1}^2 & \cdots & A_{k,L}^2 \end{array}\right]\right)$. We note, however, that one can also arrive at the same result by considering the complex amplitudes deterministic as in (6). Moreover, the matrix $\mathbf{Q}$ is the covariance matrix of the combined noise source $e(n)$, i.e., $\mathbf{Q} = \mathrm{E}\left\{\mathbf{e}(n)\mathbf{e}^H(n)\right\} = \sum_{k=1}^{K}\mathbf{Q}_k$ also referred to as the noise covariance matrix.

In practice, the covariance matrix is unknown and is replaced by an estimate, namely the sample covariance matrix defined as $\widehat{\mathbf{R}} = \frac{1}{G}\sum_{n=M-1}^{N-1}\mathbf{x}(n)\mathbf{x}^H(n)$ where $G = N - M + 1$ is the number of samples over which we average. For the sample covariance matrix $\widehat{\mathbf{R}}$ to be invertible, we require that $M < \frac{N}{2} + 1$ so that the averaging consists of at least M rank 1 vectors (see, e.g., [31] for details). In the rest of the paper, we will assume that M is chosen proportionally to N such that when N grows, so does M. This is important for the consistency of the methods under consideration.

III. Optimal Single Filter Designs

A. Basic Principle

We will now proceed with the first design. We seek to find an optimal set of coefficients, $\{h_k(m)\}$, such that the mean square error (MSE) between the filter output, $y_k(n)$, and a desired output, a signal model if you will, $\hat{y}_k(n)$, is minimized in the following sense:

$$P = \frac{1}{G}\sum_{n=M-1}^{N-1}\left|y_k(n) - \hat{y}_k(n)\right|^2, \quad (12)$$

Since we are here concerned with periodic signals, this should be reflected in the choice of the signal model $\hat{y}_k(n)$. In fact, this should be chosen as the sum of sinusoids having frequencies that are integer multiples of a fundamental frequency ω_k weighted by their respective complex amplitudes $a_{k,l}$, i.e., $\hat{y}_k(n) = \sum_{l=1}^{L_k} a_{k,l}e^{j\omega_k l n}$. This leaves us with the following expression for the MSE:

$$P = \frac{1}{G}\sum_{n=M-1}^{N-1}\left|\sum_{m=0}^{M-1}h_k(m)x(n-m) - \sum_{l=1}^{L_k}a_{k,l}e^{j\omega_k l n}\right|^2. \quad (13)$$

In the following derivations, we assume the fundamental frequency ω_0 and the number of harmonics L_k to be known (with $L_k < M$), although the so-obtained filters can later be used for finding these quantities. Next, we proceed to find not only the filter coefficients but also the complex amplitudes $a_{k,l}$. We now introduce a vector containing the complex sinusoids at time n, i.e.,

$$\mathbf{w}_k(n) = \left[\begin{array}{ccc} e^{j\omega_k 1 n} & \cdots & e^{j\omega_k L_k n} \end{array}\right]^T. \quad (14)$$

With this, we can express (12) as

$$P = \frac{1}{G}\sum_{n=M-1}^{N-1}\left|\mathbf{h}_k^H\mathbf{x}(n) - \mathbf{a}_k^H\mathbf{w}_k(n)\right|^2, \quad (15)$$

which in turn can be expanded into

$$P = \mathbf{h}_k^H\widehat{\mathbf{R}}\mathbf{h}_k - \mathbf{a}_k^H\mathbf{G}_k\mathbf{h}_k - \mathbf{h}_k^H\mathbf{G}_k^H\mathbf{a}_k + \mathbf{a}_k^H\mathbf{W}_k\mathbf{a}_k, \quad (16)$$

where the new quantities are defined as

$$\mathbf{G}_k = \frac{1}{G}\sum_{n=M-1}^{N-1}\mathbf{w}_k(n)\mathbf{x}^H(n) \quad (17)$$

and

$$\mathbf{W}_k = \frac{1}{G}\sum_{n=M-1}^{N-1}\mathbf{w}_k(n)\mathbf{w}_k^H(n). \quad (18)$$

B. Solution

Solving for the complex amplitudes in (16) yields the following expression [31]

$$\hat{\mathbf{a}}_k = \mathbf{W}_k^{-1}\mathbf{G}_k\mathbf{h}_k, \quad (19)$$

which depends on the yet unknown filter $\mathbf{h}_k$. For $\mathbf{W}_k$ to be invertible, we require that $G \geq L_k$, but to ensure that also the covariance matrix is invertible (as already noted), we will further assume that $G \geq M$. By substituting the expression above back into (16), we get

$$P = \mathbf{h}_k^H\widehat{\mathbf{R}}_k\mathbf{h}_k - \mathbf{h}_k^H\mathbf{G}_k^H\mathbf{W}_k^{-1}\mathbf{G}_k\mathbf{h}_k. \quad (20)$$

By some simple manipulation, we see that this can be simplified somewhat as

$$P = \mathbf{h}_k^H\left(\widehat{\mathbf{R}}_k - \mathbf{G}_k^H\mathbf{W}_k^{-1}\mathbf{G}_k\right)\mathbf{h}_k \triangleq \mathbf{h}_k^H\widehat{\mathbf{Q}}_k\mathbf{h}_k \quad (21)$$

where

$$\widehat{\mathbf{Q}}_k = \widehat{\mathbf{R}}_k - \mathbf{G}_k^H\mathbf{W}_k^{-1}\mathbf{G}_k \quad (22)$$

can be thought of as a *modified* covariance matrix estimate that is formed by subtracting the contribution of the harmonics from the covariance matrix given the fundamental frequency. It must be stressed, though, that for multi-pitch signals, this estimate will differ from $\mathbf{Q}_k$ in the sense that $\widehat{\mathbf{Q}}_k$ will then also contain the contribution of the other sources. Therefore, $\widehat{\mathbf{Q}}_k$ is only truly an estimate of $\mathbf{Q}_k$ for single-pitch signals. Note also that similar observations apply to the usual use of APES [29], [30].

Solving for the unknown filter in (21) directly results in a trivial and useless result, namely the zero vector. To fix this, we will introduce some additional constraints. Not only should the output of the filter be periodic, i.e., resemble a sum of harmonically related sinusoids, the filter should also have unit gain for all the harmonic frequencies of that particular source, i.e., $\sum_{m=0}^{M-1}h_k(m)e^{-j\omega_k l m} = 1$ for $l = 1, \ldots, L_k$, or, equivalently, as $\mathbf{h}_k^H\mathbf{z}(\omega_k l) = 1$. We can now state the

filter design problem as the following constrained optimization problem:

$$\min_{\mathbf{h}_k} \mathbf{h}_k^H \widehat{\mathbf{Q}}_k \mathbf{h}_k \quad \text{s.t.} \quad \mathbf{h}_k^H \mathbf{z}(\omega_k l) = 1, \quad \text{for} \quad l = 1, \ldots, L_k. \tag{23}$$

The constraints for the L_k harmonics can also be expressed as $\mathbf{h}_k^H \mathbf{Z}_k = \mathbf{1}$, where $\mathbf{1} = [\ 1\ \cdots\ 1\ ]^T$. The problem in (23) is a quadratic optimization problem with equality constraints that can be solved using the Lagrange multiplier method. Introducing the Lagrange multiplier vector

$$\boldsymbol{\lambda} = [\ \lambda_1\ \cdots \lambda_{L_k}\ ]^T, \tag{24}$$

the Lagrangian dual function of the problem stated above can be expressed as

$$\mathcal{L}(\mathbf{h}_k, \boldsymbol{\lambda}) = \mathbf{h}_k^H \widehat{\mathbf{Q}}_k \mathbf{h}_k - \left(\mathbf{h}_k^H \mathbf{Z}_k - \mathbf{1}^T\right) \boldsymbol{\lambda}. \tag{25}$$

By taking the derivative with respect to the unknown filter vector and the Lagrange multiplier vector, we get

$$\nabla \mathcal{L}(\mathbf{h}_k, \boldsymbol{\lambda}) = \begin{bmatrix} \widehat{\mathbf{Q}}_k & -\mathbf{Z}_k \\ -\mathbf{Z}_k^H & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{h}_k \\ \boldsymbol{\lambda} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{1} \end{bmatrix}. \tag{26}$$

Equaling this to zero, i.e., $\nabla \mathcal{L}(\mathbf{h}_k, \boldsymbol{\lambda}) = \mathbf{0}$, we obtain

$$\boldsymbol{\lambda} = \left(\mathbf{Z}_k^H \widehat{\mathbf{Q}}_k^{-1} \mathbf{Z}_k\right)^{-1} \mathbf{1} \tag{27}$$

and

$$\mathbf{h}_k = \widehat{\mathbf{Q}}_k^{-1} \mathbf{Z}_k \boldsymbol{\lambda}, \tag{28}$$

which combine to yield the following optimal filters:

$$\hat{\mathbf{h}}_k = \widehat{\mathbf{Q}}_k^{-1} \mathbf{Z}_k \left(\mathbf{Z}_k^H \widehat{\mathbf{Q}}_k^{-1} \mathbf{Z}_k\right)^{-1} \mathbf{1}. \tag{29}$$

We will refer to this filter as SF-APES (single filter APES-like design). This filter is optimal in the sense that it has unit gain at the harmonic frequencies and an output that resembles a sum of harmonically related sinusoids while everything else is suppressed maximally. It can readily be used for determining the amplitudes of those sinusoids by inserting (29) into (19), which yields the following estimate:

$$\hat{\mathbf{a}}_k = \mathbf{W}_k^{-1} \mathbf{G}_k \widehat{\mathbf{Q}}_k^{-1} \mathbf{Z}_k \left(\mathbf{Z}_k^H \widehat{\mathbf{Q}}_k^{-1} \mathbf{Z}_k\right)^{-1} \mathbf{1} \tag{30}$$

$$= \mathbf{W}_k^{-1} \mathbf{G}_k \left(\widehat{\mathbf{R}} - \mathbf{G}_k^H \mathbf{W}_k^{-1} \mathbf{G}_k\right)^{-1} \mathbf{Z}_k \tag{31}$$

$$\times \left(\mathbf{Z}_k^H \left(\widehat{\mathbf{R}} - \mathbf{G}_k^H \mathbf{W}_k^{-1} \mathbf{G}_k\right)^{-1} \mathbf{Z}_k\right)^{-1} \mathbf{1}. \tag{32}$$

The output power of the filter, when this is applied to the original signal, can be expressed as $\hat{\mathbf{h}}_k^H \widehat{\mathbf{R}} \hat{\mathbf{h}}_k$, which may be used for determining the fundamental frequency by treating ω_k in $\mathbf{Z}_k$, $\mathbf{G}_k$, $\mathbf{W}_k$ as an unknown parameter and then pick as an estimate the value for which the output power is maximized, i.e.,

$$\hat{\omega}_k = \arg\max_{\omega_k} \hat{\mathbf{h}}_k^H \widehat{\mathbf{R}} \hat{\mathbf{h}}_k. \tag{33}$$

In practice, this is done in the following manner: For a segment of data, the optimal filters are found for each candidate fundamental frequency. The filters are then applied to the signal and the output power is measured. This shows how much power is passed by the filters as a function of the fundamental frequency, and the fundamental frequency estimate is then picked as the fundamental frequency for which the most power is passed. One can also obtain an estimate of the number of harmonics L by estimating the noise variance by filtering out the harmonics and applying one of the many statistical model order estimation tools, like, e.g., the MAP-rule of [32], as shown in [33]. From the optimal filter, it is thus possible to obtain a full parametrization of periodic signals as was claimed in the introduction.

The proposed filter design leads to filters that are generally also much well-behaved for high SNRs, where Capon-like filters are well-known to perform poorly and require that diagonal loading or similar techniques be applied [31]. The proposed filter also holds several advantages over traditional methods, like the comb filtering approach or sinusoidal filters (also known as FFT filters), namely that it is 1) optimal given the observed signal, and 2) optimized for periodic filter output. To quantify further what exactly is meant by the filter being optimal, one has to take a look back at (12). The found filter is optimal in the sense that it minimizes the difference in (12), the exact time interval being determined by the summation limits, under the constraint that it should pass the content at specific frequencies undistorted and the output should to the extent possible resemble a periodic signal.

We will now discuss some simplified designs that are all special cases of the optimal single filter design.

1) Simplification No. 1: We remark that it can be shown that $\mathbf{W}_k$ is asymptotically identical to the identity matrix. By replacing $\mathbf{W}_k$ by $\mathbf{I}$ in (21), one obtains the usual noise covariance matrix estimate, used, for example, in [12]. As before, the optimal filters are

$$\hat{\mathbf{h}}_k = \widehat{\mathbf{Q}}_k^{-1} \mathbf{Z}_k \left(\mathbf{Z}_k^H \widehat{\mathbf{Q}}_k^{-1} \mathbf{Z}_k\right)^{-1} \mathbf{1}, \tag{34}$$

but the modified covariance matrix estimate is now determined as

$$\widehat{\mathbf{Q}}_k = \widehat{\mathbf{R}} - \mathbf{G}_k^H \mathbf{G}_k, \tag{35}$$

which is computationally simpler as it does not require the inversion of the matrix $\mathbf{W}_k$ for each candidate frequency. We refer to this design as SF-APES (appx). It must be stressed that for finite N, this is only an approximation that, nonetheless, may still be useful for practical reasons as it is much simpler. This approximation is actually equivalent to estimating the noise covariance matrix by subtracting from $\widehat{\mathbf{R}}_k$ an estimate of the covariance matrix model (for a single source) in (11) based on periodogram-like amplitude estimates.

2) Simplification No. 2: Interestingly, the Capon-like filters of [19], [34] can be obtained as a special case of the solution presented here by setting the modified covariance matrix equal to the sample covariance matrix of the observed signal, i.e., $\widehat{\mathbf{Q}}_k = \widehat{\mathbf{R}}$. More specifically, the optimal filter is then

$$\hat{\mathbf{h}}_k = \widehat{\mathbf{R}}^{-1} \mathbf{Z}_k \left(\mathbf{Z}_k^H \widehat{\mathbf{R}}^{-1} \mathbf{Z}_k\right)^{-1} \mathbf{1}, \tag{36}$$

which is the design that we will refer to as Capon in the experiments. The main difference between the design proposed here and the Capon-like designs previously proposed is that the modified covariance matrix $\widehat{\mathbf{Q}}_k$ is used in (23) in place of

$\widehat{\mathbf{R}}$, i.e., the difference is essentially in terms of the output of the filter being periodic.

3) Simplification No. 3: A simpler set of filters yet are obtained from (36) by assuming that the input signal is white, i.e., $\widehat{\mathbf{R}} = \sigma^2\mathbf{I}$. These filters are then no longer signal adaptive, but they also only have to be calculated once. The optimal filters are then given by

$$\hat{\mathbf{h}}_k = \mathbf{Z}_k \left(\mathbf{Z}_k^H \mathbf{Z}_k\right)^{-1} \mathbf{1}, \tag{37}$$

which is thus fully specified by the pseudo-inverse of $\mathbf{Z}_k$.

4) Simplification No. 4: Curiously, the filters defined in (37) can be further simplified as follows: complex sinusoids are asymptotically orthogonal for any set of distinct frequencies, which means that the pseudo-inverse of $\mathbf{Z}_k$ can be approximated as

$$\lim_{M\to\infty} M\mathbf{Z}_k \left(\mathbf{Z}_k^H \mathbf{Z}_k\right)^{-1} = \mathbf{Z}_k \lim_{M\to\infty} \left(\frac{1}{M}\mathbf{Z}_k^H \mathbf{Z}_k\right)^{-1} \tag{38}$$

$$= \mathbf{Z}_k. \tag{39}$$

This means that the filter becomes particularly simple. In fact, it is just

$$\hat{\mathbf{h}}_k = \frac{1}{M}\mathbf{Z}_k \mathbf{1}, \tag{40}$$

i.e., the normalized sum over a set of filters defined by Fourier vectors.

IV. Optimal Filterbank Designs

A. Basic Principle

We will now consider a different approach to designing optimal filters for periodic signals. Suppose that we design a filter not for the entire periodic signal, but one for each of the harmonics of the signal. In that case, we seek to find a set of filter coefficients that depend on the harmonic number l, i.e., $\{h_{k,l}(m)\}$. The corresponding output of such a filter, we denote $y_{k,l}(n)$. The output of each filter should resemble a signal model $\hat{y}_{k,l}(n)$ exhibiting certain characteristics. As was the case with the single filter, we propose a cost function defined as

$$P_l = \frac{1}{G}\sum_{n=M-1}^{N-1} \left|y_{k,l}(n) - \hat{y}_{k,l}(n)\right|^2, \tag{41}$$

which measures the extent to which the filter output $y_{k,l}(n)$ resembles $\hat{y}_{k,l}(n)$. Adding this cost up across all harmonics of the kth source, we obtain an estimate of the discrepancy as

$$P = \sum_{l=1}^{L_k} P_l = \frac{1}{G}\sum_{l=1}^{L_k}\sum_{n=M-1}^{N-1} \left|y_{k,l}(n) - \hat{y}_{k,l}(n)\right|^2. \tag{42}$$

For the single filter design, the output of each filter should resemble a periodic function having possibly a number of harmonics. In the present case, however, the output of the filter should be just a single sinusoid, i.e., $\hat{y}_{k,l}(n) = a_{k,l}e^{j\omega_k l n}$. Defining

$$y_{k,l}(n) = \sum_{m=0}^{M-1} h_{k,l}(m)x(n-m) \triangleq \mathbf{h}_{k,l}^H \mathbf{x}(n), \tag{43}$$

we can express (42) as

$$P = \frac{1}{G}\sum_{l=1}^{L_k}\sum_{n=M-1}^{N-1} \left|\mathbf{h}_{k,l}^H \mathbf{x}(n) - a_{k,l}e^{j\omega_k l n}\right|^2. \tag{44}$$

To form an estimate of the kth source from the output of the filterbank, we simply sum over all the outputs of the individual filters, as each output is an estimate of the lth harmonic, i.e.,

$$y_k(n) = \sum_{l=1}^{L_k} y_{k,l}(n) = \sum_{m=0}^{M-1}\sum_{l=1}^{L_k} \mathbf{h}_{k,l}^H \mathbf{x}(n), \tag{45}$$

which shows that the filters of the filterbank can be combined to yield the single filter needed to extract the source. As before, we proceed in our derivation of the optimal filters by expanding this expression

$$P = \sum_{l=1}^{L_k} \mathbf{h}_{k,l}^H \widehat{\mathbf{R}} \mathbf{h}_{k,l}^H + \sum_{l=1}^{L_k} |a_{k,l}|^2 \tag{46}$$

$$- \sum_{l=1}^{L_k} \mathbf{h}_{k,l}^H \mathbf{g}(\omega_k l) a_{k,l}^* - \sum_{l=1}^{L_k} a_{k,l}\mathbf{g}^H(\omega_k l)\mathbf{h}_{k,l}, \tag{47}$$

where the $\widehat{\mathbf{R}}$ is defined as before and the only new quantity is

$$\mathbf{g}(\omega) = \frac{1}{G}\sum_{n=M-1}^{N-1} \mathbf{x}(n)e^{-j\omega n}. \tag{48}$$

B. Solution

With all the basic definitions in place, we can now derive the optimal filterbank. First, however, we must solve for the amplitudes. Differentiating (47) by $\hat{a}_{k,l}$ and setting the result equal to zero, we obtain

$$\hat{a}_{k,l} = \mathbf{h}_{k,l}^H \mathbf{g}(\omega_k l) \quad \text{for} \quad l = 1, \ldots, L_k. \tag{49}$$

Inserting this back into (47), we are left with an expression that depends only on the filters $\{\mathbf{h}_{k,l}\}$:

$$P = \sum_{l=1}^{L_k} \mathbf{h}_{k,l}^H \widehat{\mathbf{R}} \mathbf{h}_{k,l} - \sum_{l=1}^{L_k} \mathbf{h}_{k,l}^H \mathbf{g}(\omega_k l)\mathbf{g}^H(\omega_k l)\mathbf{h}_{k,l} \tag{50}$$

$$= \sum_{l=1}^{L_k} \mathbf{h}_{k,l}^H \left(\widehat{\mathbf{R}} - \mathbf{g}(\omega_k l)\mathbf{g}^H(\omega_k l)\right)\mathbf{h}_{k,l} \tag{51}$$

$$\triangleq \sum_{l=1}^{L_k} \mathbf{h}_{k,l}^H \widehat{\mathbf{Q}}_{k,l} \mathbf{h}_{k,l}, \tag{52}$$

where $\widehat{\mathbf{Q}}_{k,l}$ is a modified covariance matrix estimate as before, only it now depends on the individual harmonics. We can now move on to the problem of solving for the filters. As before, we must introduce some constraints to solve this problem. It is natural to impose that each filter $\mathbf{h}_{k,l}$ should have unit gains for the lth harmonic. However, one can take additional knowledge into account in the design by also requiring that the other harmonics are canceled by the filter. Mathematically, we can state this as

$$\mathbf{h}_{k,l}^H \mathbf{Z}_k = \mathbf{b}_l, \tag{53}$$

where

$$\mathbf{b}_l = [\ \underbrace{0\ \cdots\ 0}_{l-1}\ 1\ \underbrace{0\ \cdots\ 0}_{L_k-l}\ ]. \tag{54}$$

We can now state the design problem for the lth filter of the filterbank as

$$\min_{\mathbf{h}_{k,l}} \mathbf{h}_{k,l}^H \widehat{\mathbf{Q}}_{k,l} \mathbf{h}_{k,l} \quad \text{s.t.} \quad \mathbf{h}_{k,l}^H \mathbf{Z}_k = \mathbf{b}_l. \tag{55}$$

For this problem, the Lagrangian dual function is

$$\mathcal{L}(\mathbf{h}_{k,l}, \boldsymbol{\lambda}) = \mathbf{h}_{k,l}^H \widehat{\mathbf{Q}}_{k,l} \mathbf{h}_{k,l} - \left(\mathbf{h}_{k,l}^H \mathbf{Z}_k - \mathbf{b}_l^T\right) \boldsymbol{\lambda}. \tag{56}$$

By taking the derivative with respect to the unknown filter vector and the Lagrange multiplier vector, we get

$$\nabla \mathcal{L}(\mathbf{h}_{k,l}, \boldsymbol{\lambda}) = \begin{bmatrix} \widehat{\mathbf{Q}}_{k,l} & -\mathbf{Z}_k \\ -\mathbf{Z}_k^H & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{h}_{k,l} \\ \boldsymbol{\lambda} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{b}_l \end{bmatrix}. \tag{57}$$

By the usual method, we obtain

$$\boldsymbol{\lambda} = \left(\mathbf{Z}_k^H \widehat{\mathbf{Q}}_{k,l}^{-1} \mathbf{Z}_k\right)^{-1} \mathbf{b}_l \tag{58}$$

and

$$\mathbf{h}_{k,l} = \widehat{\mathbf{Q}}_{k,l}^{-1} \mathbf{Z}_k \boldsymbol{\lambda}. \tag{59}$$

This, finally, results in the following optimal filters for $l = 1, \ldots, L_k$

$$\hat{\mathbf{h}}_{k,l} = \widehat{\mathbf{Q}}_{k,l}^{-1} \mathbf{Z}_k \left(\mathbf{Z}_k^H \widehat{\mathbf{Q}}_{k,l}^{-1} \mathbf{Z}_k\right)^{-1} \mathbf{b}_l. \tag{60}$$

We will refer to this design as FB-APES (filterbank APES-like design). The individual filters can now be applied to obtain amplitude estimates as

$$\hat{a}_{k,l} = \hat{\mathbf{h}}_{k,l}^H \mathbf{g}(\omega_k l) \tag{61}$$

$$= \mathbf{b}_l^H \left(\mathbf{Z}_k^H \widehat{\mathbf{Q}}_{k,l}^{-1} \mathbf{Z}_k\right)^{-1} \mathbf{Z}_k^H \widehat{\mathbf{Q}}_{k,l}^{-1} \mathbf{g}(\omega_k l). \tag{62}$$

Organizing all the filters for the kth source in a matrix, we get

$$\widehat{\mathbf{H}}_k = [\ \mathbf{h}_{k,1}\ \cdots\ \mathbf{h}_{k,L_k}\ ]. \tag{63}$$

The optimal filters in (60) can also be rewritten using the matrix inversion lemma to obtain an expression that does not require direct inversion of $\widehat{\mathbf{Q}}_{k,l}$ of each l:

$$\widehat{\mathbf{Q}}_{k,l}^{-1} = \left(\widehat{\mathbf{R}} - \mathbf{g}(\omega_k l)\mathbf{g}^H(\omega_k l)\right)^{-1} \tag{64}$$

$$= \widehat{\mathbf{R}}^{-1} + \frac{\widehat{\mathbf{R}}^{-1}\mathbf{g}(\omega_k l)\mathbf{g}^H(\omega_k l)\widehat{\mathbf{R}}^{-1}}{1 - \mathbf{g}^H(\omega_k l)\widehat{\mathbf{R}}^{-1}\mathbf{g}(\omega_k l)}, \tag{65}$$

which can then be inserted into (60). As with the single filter approach, this design can also be used for estimating the fundamental frequency by summing over the output powers of all the filters, i.e.,

$$\hat{\omega}_k = \arg\max_{\omega_k} \sum_{l=1}^{L_k} \hat{\mathbf{h}}_{k,l}^H \widehat{\mathbf{R}} \hat{\mathbf{h}}_{k,l} \tag{66}$$

$$= \arg\max_{\omega_k} \sum_{l=1}^{L_k} \mathrm{Tr}\left\{\widehat{\mathbf{H}}_k^H \widehat{\mathbf{R}} \widehat{\mathbf{H}}_k\right\}. \tag{67}$$

Note that the filters can also be applied in a different way, or, rather, the output power can be measured differently. In (66), the output power is determined as the sum of output power of the individual filters. If, instead, the output power is measured on the estimated source obtained as in (45), one obtains

$$\mathrm{E}\left\{|y_k(n)|^2\right\} = \left(\sum_{l=1}^{L_k} \hat{\mathbf{h}}_{k,l}^H\right) \widehat{\mathbf{R}} \left(\sum_{l=1}^{L_k} \hat{\mathbf{h}}_{k,l}\right). \tag{68}$$

However, assuming that the output of the individual filters is uncorrelated, the two estimates will be identical (see [34] for more details about this).

At this point some remarks are in order. For the Capon-like filters of [19], [34], the single filter and the filterbank approaches are closely related. This is, however, not the case for the designs considered here in that they operate on different covariance matrix estimates, $\widehat{\mathbf{Q}}_k$ and $\widehat{\mathbf{Q}}_{k,l}$, respectively. While it is more complicated to compute the former than the latter, the latter must be computed a number of times, once for each harmonic l. This suggests that, in fact, the single filter should be preferable from a complexity point of view if the number of harmonics is high.

As with the single filter design, it is possible to obtain some simplified versions of the optimal design. Next, we will look more into some of these.

1) Simplification No. 1: By posing the optimization problem in (55) in a slightly different way, we obtain an important special case. More specifically, by changing the constraints of (55) such that each filter only has to have unit gain for the corresponding harmonic, we obtain the following problem:

$$\min_{\mathbf{h}_{k,l}} \mathbf{h}_{k,l}^H \widehat{\mathbf{Q}}_{k,l} \mathbf{h}_{k,l} \quad \text{s.t.} \quad \mathbf{h}_{k,l}^H \mathbf{z}(\omega_k l) = 1, \tag{69}$$

where, as before, $\widehat{\mathbf{Q}}_{k,l} = \widehat{\mathbf{R}} - \mathbf{g}(\omega_k l)\mathbf{g}(\omega_k l)^H$. The solution to this problem is, in fact, the usual single sinusoid APES filter [29], [30], which is

$$\hat{\mathbf{h}}_{k,l} = \frac{\widehat{\mathbf{Q}}_{k,l}^{-1}\mathbf{z}(\omega_k l)}{\mathbf{z}^H(\omega_k l)\widehat{\mathbf{Q}}_{k,l}^{-1}\mathbf{z}(\omega_k l)}. \tag{70}$$

This design takes only the individual harmonics into account in the design of the individual filters. Essentially, the filter that is obtained from (55) takes the presence of all the harmonics of the kth source into account, while the present one does not.

2) Simplification No. 2: Taking this one step further and replacing $\widehat{\mathbf{Q}}_{k,l}$ by $\widehat{\mathbf{R}}$, one obtains the well-known single sinusoid Capon filter [35]

$$\hat{\mathbf{h}}_{k,l} = \frac{\widehat{\mathbf{R}}^{-1}\mathbf{z}(\omega_k l)}{\mathbf{z}^H(\omega_k l)\widehat{\mathbf{R}}^{-1}\mathbf{z}(\omega_k l)}. \tag{71}$$

As with the prior simplification, this design leaves it for the algorithm to automatically cancel out the contribution of the other harmonics.

3) Simplification No. 3: Similarly, replacing $\widehat{\mathbf{Q}}_{k,l}$ by $\widehat{\mathbf{R}}$ in (60) results in the filters

$$\hat{\mathbf{h}}_{k,l} = \widehat{\mathbf{R}}^{-1}\mathbf{Z}_k \left(\mathbf{Z}_k^H \widehat{\mathbf{R}}^{-1} \mathbf{Z}_k\right)^{-1} \mathbf{b}_l, \tag{72}$$

which are identical to the filters of the optimal Capon-like filterbank of [19]. Interestingly, when summed, it result in the

optimal single Capon-like filter as

$$\sum_{l=1}^{L_k} \hat{\mathbf{h}}_{k,l} = \hat{\mathbf{h}}_k. \tag{73}$$

4) Simplification No. 4: The previous design can, of course, be simplified further by assuming that the covariance matrix is white, i.e., $\widehat{\mathbf{R}} = \sigma^2\mathbf{I}$, which results in static filters that have to be calculated only once. The filters are then given by

$$\hat{\mathbf{h}}_{k,l} = \mathbf{Z}_k \left(\mathbf{Z}_k^H \mathbf{Z}_k\right)^{-1} \mathbf{b}_l, \tag{74}$$

which when organized in a filterbank matrix can be written as

$$\widehat{\mathbf{H}}_k = \mathbf{Z}_k \left(\mathbf{Z}_k^H \mathbf{Z}_k\right)^{-1}. \tag{75}$$

Source estimates obtained using this filterbank, as described in (45), will be exactly the same estimates as one would get using (37)–this can easily be verified by inserting the right hand side of (74) in (45). The resulting fundamental frequency estimators are, however, generally different, but are equivalent under certain conditions. In the experimental part of this paper, we will refer to this method as the FB-WNC design (filterbank white noise Capon-like design).

5) Simplification No. 5: Applying the asymptotic approximation in (39) to the filters in (74), we obtain even simpler filters. More specifically, (74) reduces to

$$\hat{\mathbf{h}}_{k,l} = \frac{1}{M}\mathbf{Z}_k\mathbf{b}_l, \tag{76}$$

and the filterbank matrix is then simply given by

$$\widehat{\mathbf{H}}_k = \frac{1}{M}\mathbf{Z}_k. \tag{77}$$

When applied to the problem of fundamental frequency estimation, as in (66), this leads to the familiar approximate non-linear least squares (NLS) method–it is nonlinear in the fundamental frequency, hence the name; it is also sometimes referred to as the harmonic summation method [27]. Note that when source estimates are obtain using this filterbank as described in (45), one will obtain exactly the same estimate as with (40). We will refer to this method as FB-WNC (appx) in the experiments, where it will serve as a method representative of the usual way filters are designed. A large class of methods exist for enhancement and separation of signals that operate on the coefficients of the short-time Fourier Transform (STFT) (see, e.g., [36], [37]). The individual bases of the STFT are the same as the individual filters of the filterbank (76), in fact, this will be the case for all methods that operate directly on the coefficients of the STFT, including mask-based methods like [38] and non-negative matrix factorization-based methods like [39].

6) Simplification No. 6: We will close this section by introducing one final simplification. If in lieu of $\widehat{\mathbf{Q}}_{k,l}$ we use $\widehat{\mathbf{Q}}_k$ as obtained for the single filter approach in (22) in (60), the optimal filters of the filterbank are then given by

$$\hat{\mathbf{h}}_{k,l} = \widehat{\mathbf{Q}}_k^{-1}\mathbf{Z}_k \left(\mathbf{Z}_k^H \widehat{\mathbf{Q}}_k^{-1}\mathbf{Z}_k\right)^{-1} \mathbf{b}_l. \tag{78}$$

It can be seen that the only difference between the different filters of the filterbank is then the vector $\mathbf{b}_l$, which serves to extract the filter for the individual harmonics. The filterbank matrix containing these filters can then be expressed as

$$\widehat{\mathbf{H}}_k = \widehat{\mathbf{Q}}_k^{-1}\mathbf{Z}_k \left(\mathbf{Z}_k^H \widehat{\mathbf{Q}}_k^{-1}\mathbf{Z}_k\right)^{-1}. \tag{79}$$

It is then also easy to see that these filters are related to the optimal single filter in (29) in a trivial way as

$$\hat{\mathbf{h}}_k = \widehat{\mathbf{H}}_k\mathbf{1}. \tag{80}$$

A similar relationship exists for the corresponding Capon-like filters [34]. Curiously, one would also obtain these filters by modifying (42) by moving the summation over the harmonics inside the absolute value, which would also be consistent with the formation of the source estimates according to (45).

V. Results

A. Practical Considerations

Before moving on to the experimental parts of the present paper, we will now go a bit more into details of how to apply the proposed filters and what issues one has to consider in doing so.

Given a segment of new data $\{x(n)\}$, the procedure is as follows:

1) Estimate the fundamental frequencies $\{\omega_k\}$ of all sources of interest for the data $\{x(n)\}$.
2) Determine or update recursively the sample covariance matrix $\widehat{\mathbf{R}}$.
3) Compute a noise covariance matrix estimate $\mathbf{Q}_k$ for each source (or for its harmonics $\mathbf{Q}_{k,l}$) and the inverse.
4) Compute the optimal single filter $\mathbf{h}_k$ or filterbank $\mathbf{H}_k$ for each source of interest k using one of the proposed designs.
5) Perform block filtering on the data $\{x(n)\}$ to obtain source estimates $y_k(n)$ for each source of interest k (using the observed signal from the previous segment as filter states as appropriate).

In performing the above, there are a number of user parameters that must be chosen. The following may serve as a basis for choosing these. Generally speaking, the higher the filter length M, the better the filter will be in attenuating noise and canceling interference from other sources as the filter has more degrees of freedom. This also means that the higher the model order, the more interfering sources the filter can deal with. However, there are several concerns that limit the filter length. First of all, the validity of the signal model. If the signal is not approximately stationary over the duration of the segment, the filters cannot possibly capture the signal of interest, neither can it deal with noise and other sources. On a related issue, the filter length M must be chosen, as mentioned, with $M < N/2 + 1$ to yield a well-conditioned problem. This means that the signal should be stationary over N and not just M. It should of course also be taken into account that the higher the filter order, the more computationally complex the design will also be. Regarding how often one should compute the optimal filters, i.e., how high the update-rate should be relative to M and N, it should be noted that for the filter outputs to be well-behaved, the filters must not

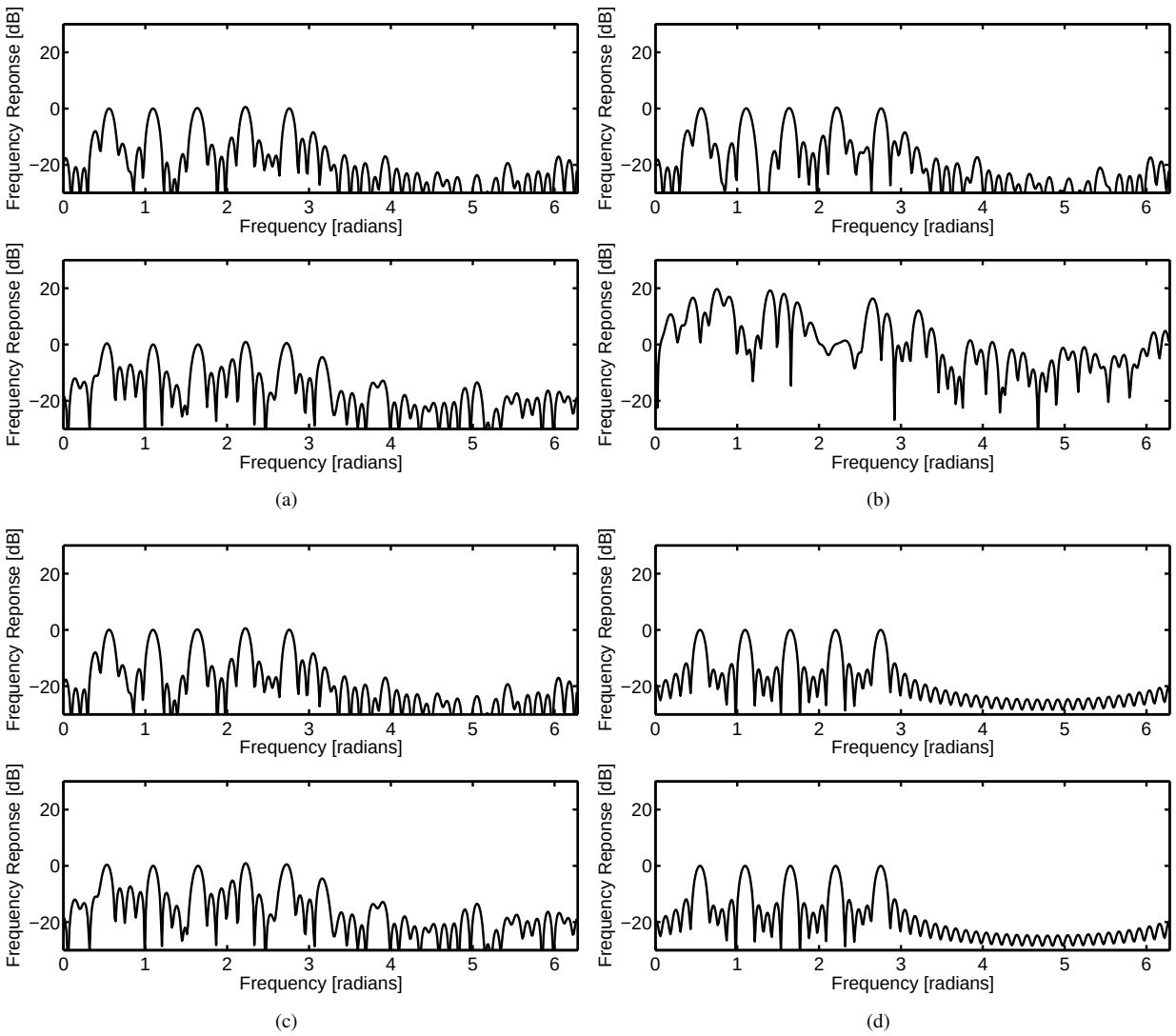


Fig. 1. Frequency responses of the various filters for a set of harmonically related sinusoids in white Gaussian noise at an SNR of -20 dB (top panels) and 20 dB (bottom panels). The designs shown here are (a) SF-APES, (b) SF-Capon, (c) FB-APES, (d) FB-WNC.

change abruptly. Consequently, it is advantageous to update the filters as often as possible by computing a new covariance matrix and subsequently new filters at the cost of increased computational complexity. In this process, one may also just as well update the fundamental frequency. In fact, it may also be advantageous to estimate a new fundamental frequency frequently relative to M and N to track changes in the signal of interest. This all suggests that it should be preferable in most situations to update the fundamental frequency, the covariance matrix and filters frequently.

Regarding numerical issues, as we have seen, the Capon-design suffers from bad conditioning of the covariance matrix for high SNRs, and it may thus be reasonable to use a regularized estimate of the covariance matrix, like $\widetilde{\mathbf{R}} = \widehat{\mathbf{R}} + \delta \mathbf{I}$, where δ is a small positive constant, before computing inverses. It is also possible that the APES-like designs may benefit from such modified estimates under extreme conditions.

B. Tested Designs

In the tests to follow, we will compare the proposed design methods to a number of existing FIR design methods. More specifically, we will compare the following:

- SF-APES which is the optimal single filter design given by (29).
- SF-Capon, i.e., the single filter design proposed in [19], [34], which is based on a generalization of the Capon principle. The optimal filter is given by (36).
- SF-APES (appx) is an approximation of SF-APES based on the simpler modified covariance matrix estimate in (35). It is thus a computationally simpler approximation to SF-APES.
- FB-APES is the optimal filterbank design given by (60).
- FB-WNC is a static single filter design based on Fourier vectors. The filter is given by (74). It serves as reference method as such filters are often used for processing of periodic signals.
- FB-WNC (appx) is an approximation of the FB-WNC

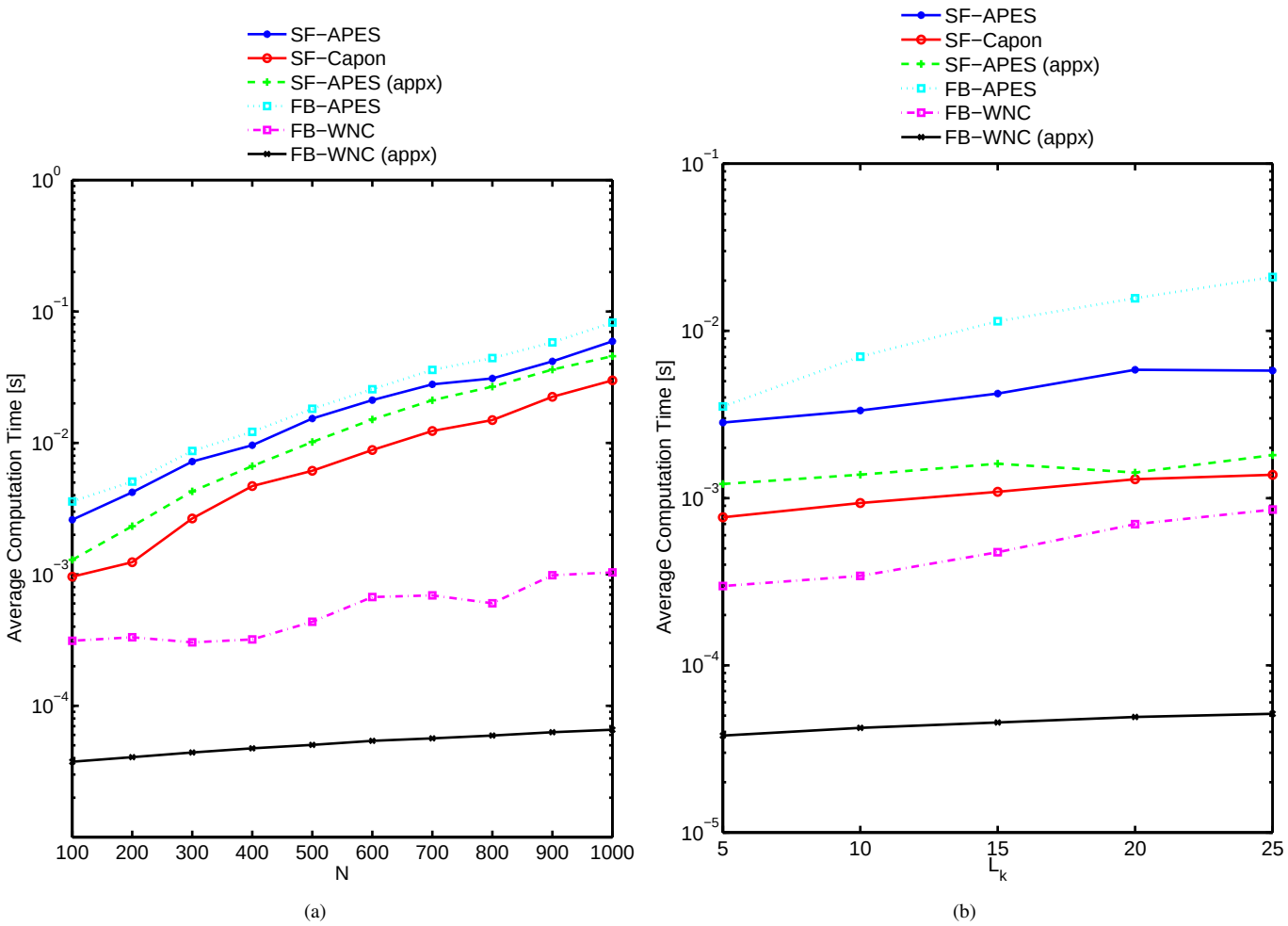


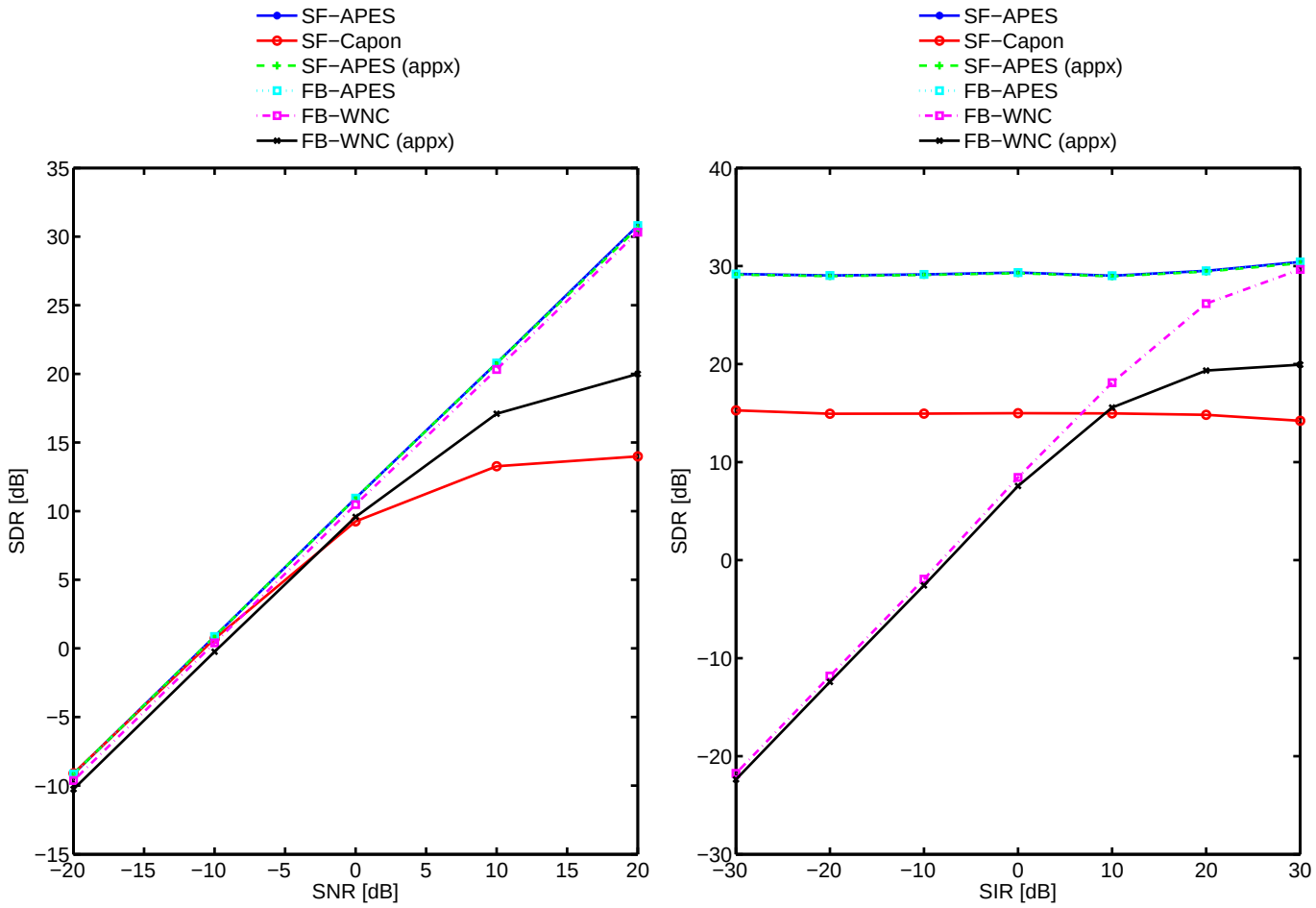
Fig. 2. Estimated computation times for the various filters designs (a) as a function of the number of observations N with $M = N/4$ and $L_k = 5$, and (b) as a function of the number of harmonics L_k with $N = 100$ and $M = 25$. For each data point, each filter was computed 1000 times and the average was computed.

filters with the filters being defined in (76). It is based on the asymptotic orthogonality of complex sinusoids. It is perhaps the most commonly used filter design method for processing periodic signals and is sometimes also referred to as the frequency sampling design method or the resulting filters as FFT filters.

Note that we do not include all the simplifications of Sections III and IV as some of them are trivially related.

C. Frequency Response

We will start out the experimental part of this paper by showing an example of the optimal filters obtained using some of the proposed methods and their various simplifications and the Capon-like filters of [19], [34]. More specifically, we will show the frequency response of the filters obtained using some of the various designs for a synthetic signal. In Figure 1, these are shown for a synthetic signal having $\omega_0 = 0.6283$, $L = 5$, Rayleigh distributed amplitudes and uniformly distributed phases with white Gaussian noise added at a -20 dB SNR (top panels) and 20 dB (bottom panels). The filters all have length 50 in these examples and were estimated from 200 samples. All the filters can be seen to exhibit the expected response for -20 dB SNR following the harmonic structure of the signal having 0 dB gain for the harmonic frequencies, and several of them are also quite similar. For an SNR of 20 dB, however, it can clearly be seen that the proposed filters still exhibit the desired response emphasizing the harmonics of the signal. The Capon-like design, SF-Capon, however, behaves erratically for 20 dB SNR, and this is typical of the Capon-like filters. Comparing the response of this method to the proposed ones, namely SF-APES, and FB-APES, it can be seen that this problem is overcome by the new design methodology. The erratic behavior of the Capon-like filter can be understood by noting that for high SNR, the Capon method will generally suffer from poor conditioning of the sample covariance matrix (as the eigenvalues only due to the noise tending toward zero), explaining the low accuracy of the resulting filter, and as the SNR increases, the filters obtained using the SF-Capon design will get progressively worse. We also remark that for the example considered here, SF-APES (appx) will be quite similar to SF-APES and FB-WNC (appx) to FB-WNC, for which reason these designs are not shown. This is because the asymptotic approximations that these derivative methods are based on are quite accurate in this case. This is also the likely



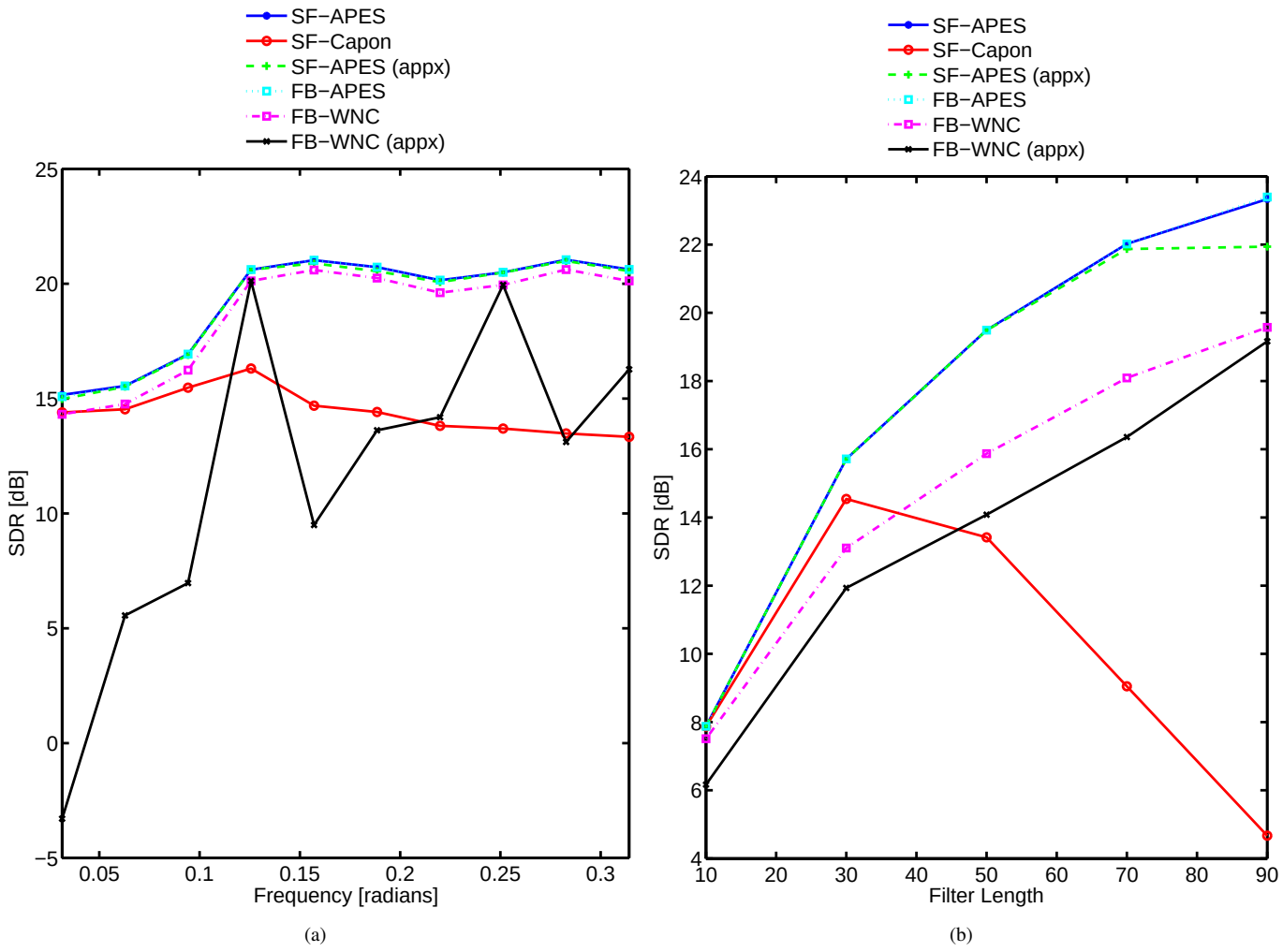
(a) (b)

Fig. 3. Performance of the various filters in SDR (a) as a function of the SNR and (b) the SIR with an interfering source present (with noise added at a fixed SNR of 20 dB).

explanation for the frequency responses of SF-APES and FB-APES looking extremely similar for both SNRs. We remark that while the adaptive designs will change with the observed signal, FB-WNC and its simplification will remain the same.

D. Computational Complexity and Computation Times

In comparing the performance of the various methods, it is of course also important to keep the computational complexity of the various methods in mind. All the tested methods, except the FB-WNC (appx) design, have cubic complexities involving operations of complexity $\mathcal{O}(M^3)$, $\mathcal{O}(L_k^3)$, $\mathcal{O}(M^2 L_k)$, and $\mathcal{O}(M L_k^2)$, as they involve matrix inversions and matrix-matrix multiplications. Some of the designs avoid some matrix inversions, like the SF-APES (appx) design, but such details cannot be differentiated with these asymptotic complexities. We therefore have measured average computation times of the various designs in MATLAB. More specifically, we have computed the average computation times over 1000 trials as a function of L_k and N as M is assumed to be chosen proportionally to N. The measurements were obtained on an Intel(R) Core(TM)2 CPU 6300 @ 1.86GHz with 2GB of RAM running MATLAB 7.6.0 (R2008a) and Linux 2.6.31-17 (Ubuntu). Note that the current implementations do not take into account the structure of the various matrices like, e.g., Toeplitz structure of the covariance matrix. The obtained results are shown in Figure 2(a) as a function of N with $M = N/4$ and $L_k = 5$ and as a function of L_k with $N = 100$ and $M = 25$ in Figure 2(b) for typical ranges of these quantities. From Figure 2(a) it can be observed that the computational complexity of the the designs SF-APES, SF-APES (appx), FB-APES, and SF-Capon indeed are cubic in M (and thus N), the difference essentially being a scaling. It can be observed that the FB-APES design is the most complex, owing to the different noisy covariance matrix estimates that must be determined for each harmonic. Note that for a very low number of harmonics, this design is less complex than SF-APES and SF-APES (appx). It can also be seen that, as expected, the SF-Capon design is the least complex of the adaptive designs, as it does not require the computation of a noise covariance matrix estimate. The general picture is the same in Figure 2(b), although it can be observed that the difference in computation time between the FB-APES method and the others appear to increase on the logarithmic scale as the number of harmonics is increased, the reason again being that the higher the number of harmonics, the more noise covariance matrices (and their inverses) must be determined.



(a) (b)

Fig. 4. Performance of the various filters in SDR (a) as function of the fundamental frequency, (b) and the filter length with an interfering source present.

E. Enhancement and Separation

Next, we will consider the application of the various filter designs to extracting periodic signals from noisy mixtures containing other periodic signals and noise or just noise. We will test the performance under various conditions by generating synthetic signals and then use the filters for extracting the desired signal. More specifically, the signals are generated in the following manner: A desired signal $\mathbf{s}_1(n)$ that we seek to extract from an observed signal $\mathbf{x}(n)$ is buried in a stochastic signal, i.e., noise $\mathbf{e}(n)$; additionally, an interfering source $\mathbf{s}_2(n)$ is also present, here in the form of a single sinusoid. The observed signal is thus constructed as

$$\mathbf{x}(n) = \mathbf{s}_1(n) + \mathbf{s}_2(n) + \mathbf{e}(n). \tag{81}$$

We will measure the extent to which the various filter designs are able to extract $\mathbf{s}_1(n)$ from $\mathbf{x}(n)$ using the signal-to-distortion ratio (SDR) defined as:

$$SDR = 20\log_{10}\frac{\|\mathbf{s}_1(n)\|_2}{\|\mathbf{s}_1(n) - \mathbf{y}_1(n)\|_2} \quad [\text{dB}]\,, \tag{82}$$

where $\mathbf{y}_1(n)$ is the signal extracted by applying the obtained filters to $\mathbf{x}(n)$. The ultimate goal is of course to reconstruct $\mathbf{s}_1(n)$ as closely as possible and, therefore, to maximize the SDR.

As a measure of the power of the interfering signal $\mathbf{s}_2(n)$ relative to the desired signal $\mathbf{s}_1(n)$, we use the following measure:

$$SIR = 20\log_{10}\frac{\|\mathbf{s}_1(n)\|_2}{\|\mathbf{s}_2(n)\|_2} \quad [\text{dB}]\,, \tag{83}$$

which we refer to as the signal-to-interference ratio (SIR) (for a discussion of performance measures for assessment of separation algorithms see, e.g., [38], [40]). It is expected that the higher the SIR, the worse the SDR will be. Finally, we measure how noisy the signal is using the signal-to-noise ratio (SNR) defined as

$$SNR = 20\log_{10}\frac{\|\mathbf{s}_1(n)\|_2}{\|\mathbf{e}(n)\|_2} \quad [\text{dB}]\,. \tag{84}$$

The reader should be aware that our definitions of SDR and SIR are consistent with those of [40], but also that our definition of SNR differs but is consistent with its use in estimation theory. In the experiments reported next, unless otherwise stated, the conditions were as follows; the above quantities were calculated by applying the found filters to the observed

signal and the SDR was then measured. This was then repeated 100 times for each test condition, i.e., the quantities are determined using Monte Carlo simulations. In doing this, the zero-state responses of the filters were ignored. Segments of length $N = 200$ were used with filter lengths of $M = N/4$ (for all designs) and an SNR of 20 dB was used. The desired signal was generated with a fundamental frequency of 0.5498 and five harmonics. The real and imaginary values of the complex amplitudes were generated as realizations of i.i.d. Gaussian random variables, leading to Rayleigh distributed amplitudes and uniformly distributed phases. The interfering source was a periodic signal having a fundamental frequency of 0.5890, five harmonics and with Rayleigh distributed amplitudes and uniformly distributed phases. Its amplitudes were then scaled to match the desired SIR in each realization. In these experiments, we will assume that the fundamental frequency of the desired signal is known while the fundamental frequency of the interference is unknown. As has already been mentioned, it is possible to estimate the fundamental frequency using the proposed filters, but this is beyond the scope of this paper, and we will just assume that the fundamental frequency has been estimated a priori using one of the methods of [27].

In the first experiment, only the desired signal and the noise are present, i.e., no interfering source was added, and the performance of the filters is observed as a function of the SNR. The resulting measurements are plotted in Figure 3(a). It can be seen that the Capon-like filter design, SF-Capon, that was the starting point of this work, performs poorly in this task. In fact, it is worse than the static designs FB-WNC and FB-WNC (appx). It can also be observed that the APES-like filters, SF-APES, SF-APES (appx) and FS-APES, all perform well, achieving the highest SDR. In [19], it was shown that the Capon-like filters perform well in terms of multi-pitch estimation under adverse conditions compared to the alternatives. This was especially true when multiple periodic sources were present at the same time as the signal-adaptive optimal designs were able to cancel out the interference without prior knowledge of it. It appears that with this particular setup, there is a 10 dB reduction in the noise regardless of the SNR for the proposed filters, and, interestingly, all the filter designs seem to tend perform similarly for low SNRs. This means that there appears to be no reason to prefer one method over the others for low SNRs, in which case the simplest design then should be chosen.

The next experiment is, therefore, concerned with the performance of the filters when interference is present. Here, the noise level, i.e., the SNR, is kept constant at 20 dB while the SIR is varied. The results are depicted in Figure 3(b). This figure clearly shows the advantage that the adaptive designs, SF-APES, SF-APES (appx), FB-APES, and SF-Capon, hold over the static ones, FB-WNC and FB-WNC (appx) in that the former perform well even when the interference is very strong, while the latter does not. The advantages of the designs proposed herein are also evident as the APES-like filters, SF-APES, SF-APES (appx), and FS-APES, outperform all others for the entire tested range of SIR values. We remark that in several of these figures, it may be hard to distinguish the performance of SF-APES, SF-APES (appx), and FB-APES as the curves are very close; indeed they appear to have similar performance in terms of SDR.

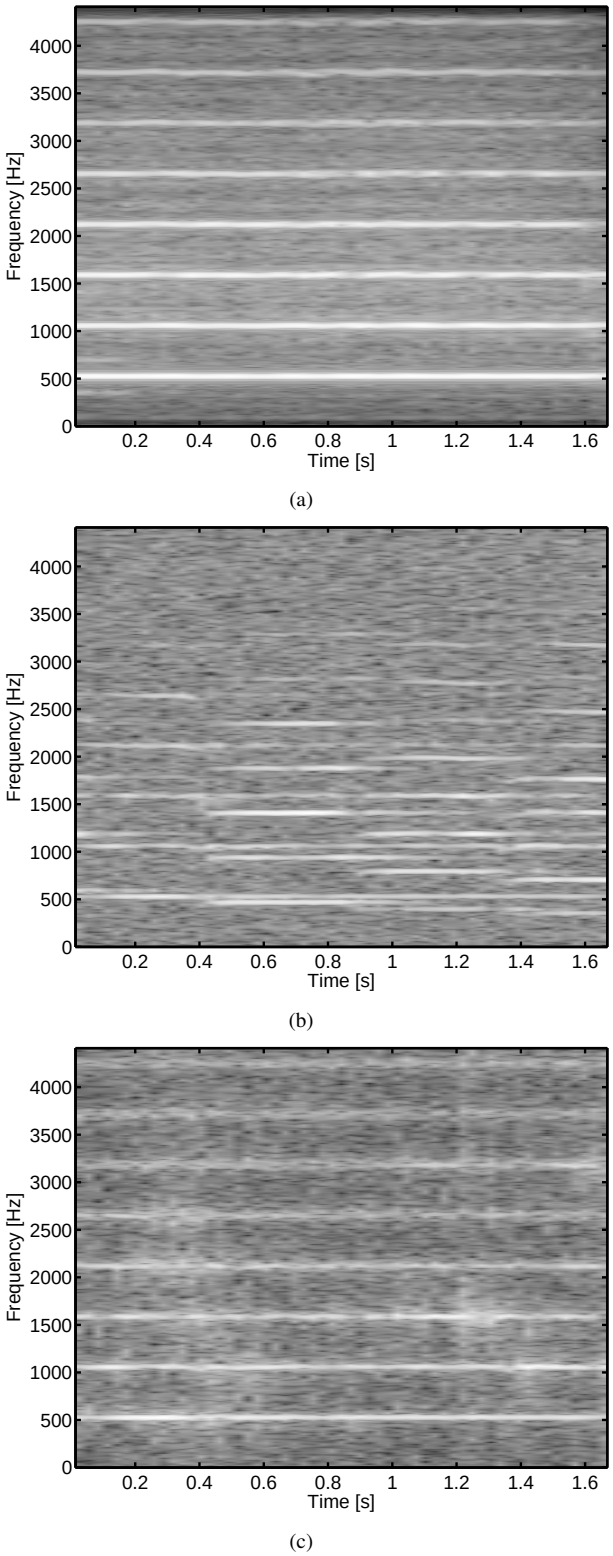


Fig. 5. Shown are: (a) the spectrogram of the original signal, (b) with noise and interference added with $SNR = 0$ dB and $SIR = -10$ dB (c) and the signal extracted using the FS-APES design.

As some of the simpler designs are based on sinusoids being asymptotically orthogonal, namely SF-APES (appx) and FB-

WNC (appx), it is interesting to see how the various filters perform when this is not the case. We do this by lowering the fundamental frequency for a given N, as for a given N, the fundamental frequency has to be high, relatively speaking, for the asymptotic approximation to hold. In this case, only noise is added to the desired signal at an SNR of 10 dB. The results are shown in Figure 4(a). As could be expected, the aforementioned approximate designs perform poorly (as does the Capon-like filters SF-Capon), but, generally, the performance of all the methods degrades as the fundamental frequency is lowered. This is, however, to be expected. Note that the reason FB-WNC (appx) performs well for certain fundamental frequencies is that the harmonics may be close to (or exactly) orthogonal, but this would merely be a coincidence in all practical situations.

Now we will investigate the influence of the filter length by varying M while keeping N fixed at 200, here in the presence of an interfering source. In this case, noise is added at an SNR of 10 dB while the SIR was 10 dB. In Figure 4(b), the results are shown. The conclusions are essentially the same as for the other experiments; the proposed filter designs perform the best, the SF-Capon filters behave erratically, and the static designs FB-WNC and FB-WNC (appx) perform poorly when interference is present. We note that for the respective matrices to be invertible, the filter lengths cannot be too long. On the other hand, one would expect that the longer the filters, the better the performance as the filters have more degrees of freedom to capture the desired signal while canceling noise and interference, and this indeed seems to be the case.

These experiments generally show that the proposed filter designs have a number of advantages over previous designs and static designs alike when applied to the problem of separating periodic signals. Among the proposed designs, SF-APES and FB-APES appear to perform the best and equally well while SF-APES (appx) is sometimes slightly worse.

F. Some Speech and Audio Examples

We will now demonstrate the applicability of the proposed methods to real signals. In the experiments to follow, we will use the SF-APES design. In the first such experiment, we will use the filters obtained using the said method to extract a real trumpet signal, a single tone sampled at ~ 8 kHz using 50 ms segments and a filter length of 100 and the filter is updated every 5 ms. Note that both the signal and the filters are complex by mapping the input signal to its analytic counterpart using the Hilbert transform. For each segment the fundamental frequency and the model order was found using the approximate non-linear least squares method of [27] and the optimal filter was updated every 1 ms. The single tone has been buried in noise at an SNR of 0 dB and interfering tones, which were also trumpet tones (both signals are from the SQAM database [41]), have been added with an SIR of -10 dB. The spectrogram of the original signal is shown in Figure 5(a) and the same signal with noise and interference added is depicted in Figure 5(b). The spectrogram of the extracted signal is shown in Figure 5(c). These figures clearly demonstrate the ability of the APES-like designs to extract the signal while rejecting not only noise, but also strong periodic interference even when these are fairly close to the harmonics of the desired signal. Note that for this particular example, because the SIR and SNR are quite low, the FS-Capon method would also perform quite well.

Regarding the application of the proposed filters to speech signals, an interesting question is whether the filters are suitable for such signals, as they exhibit non-stationarity. To address this question, we apply the SF-APES method to a voiced speech signal, this particular signal being from the SQAM database [41] and sampled at 11025 Hz. As with the prior example, we estimate the pitch for each segment, which are here of size 30 ms (corresponding to 165 complex samples), a size commonly used in speech processing and coding. From these segments, the optimal filterbank is then also determined using the estimated pitch. In this example, the complex filters of length 40 are updated every 2.5 ms. The signal is depicted in Figure 6(a) and the extracted signal is shown in Figure 6(b). The difference between the original signal and the extracted one is shown in Figure 6(c) and the estimated pitch is shown in Figure 6(d). A number of observations can be made regarding the original signal. Firstly, it is non-stationary at the beginning and the end with a time-varying envelope, and the pitch can be observed to vary too. It can, however, be observed from the extracted signal and the corresponding error signal that the filters are indeed able to track this signal, resulting in an SDR of 20 dB. This demonstrates that the filters may be useful even if the signal is not completely stationary.

Our final example involves the separation of two speech signals, more specifically two quasi-stationary segments of voiced speech mixed at an SIR of 0 dB. These signals are sampled at 8 kHz and are from the EUROM.1 corpus [42]. As before 30 ms segments are used for determining the pitch and the optimal filters resulting in segments consisting of 120 complex samples along with filters of length 30. We here update the filters every 2.5 ms. In Figures 7(a) and 7(b), the two signals are shown along with their mixture in Figure 7(c). As before, the fundamental frequencies of the two sources are estimated with the approximate non-linear least squares method [27], and the resulting estimates are shown in Figure 7(d). It can be seen that one source has an average pitch of approximately 162 Hz while that of the other is about 200 Hz. The two extracted signals are shown in Figures 7(e) and 7(f), respectively. As can be seen, the filters are able to separate the signals achieving SDRs of 14 and 12 dB, respectively. Of course, some errors occur, as can also clearly be seen, as parts of the other interfering source will be passed by the filters.

VI. Conclusion

In this paper, new filter designs for extracting and separating periodic signals have been proposed, a problem occurring frequently in, for example, speech and audio processing. The proposed filters are designed such that they have unit gain at the frequencies of the harmonics of the desired signal and suppress everything else. The novel part of the present designs is that they are optimized for having an output that is

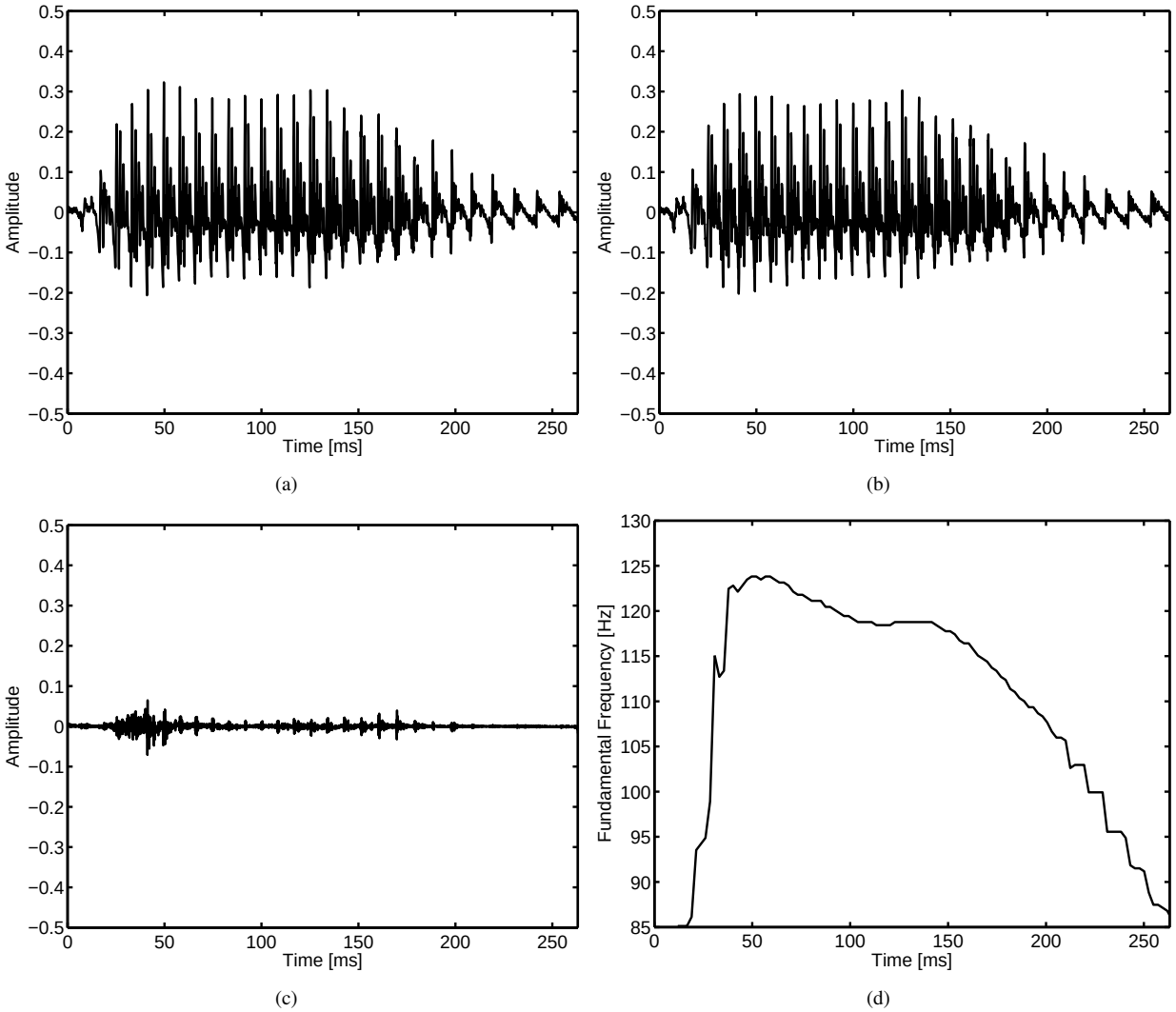


Fig. 6. Shown are: (a) the original voiced speech signal, (b) the extracted signal, (c) the difference between the two signals, i.e., the part of the signal that was not extracted, and (d) the estimated pitch used in the filters.

approximately periodic too. Additionally, the obtained filters are optimal for a segment of the observed signal and are thus signal-adaptive. The filter designs can be used not only for the aforementioned applications but also for estimating the parameters of periodic signals. The designs have been demonstrated to overcome the shortcomings of previous designs while retaining their desirable properties, like the ability to cancel out interfering signals. We have shown how the new designs reduce to a number of well-known designs under certain conditions and they can thus be seen as generalizations of previous methods. In simulations, we have demonstrated the superior performance of the obtained filters in enhancement and separation applications.

REFERENCES

[1] American Standards Association (ASA), "Acoustical Terminology, SI, 1-1960," 1960.

[2] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 27–30 April 1993, vol. 2, pp. 728–731.

[3] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., chapter 4, pp. 121–174. Elsevier Science B.V., 1995.

[4] B. Kostek, "Musical instrument classification and duet analysis employing music information retrieval techniques," *Proc. IEEE*, vol. 92, no. 4, pp. 712–729, Apr 2004.

[5] J. H. Jensen, M. G. Christensen, D. P. W. Ellis, and S. H. Jensen, "Quantitative analysis of a common audio similarity measure," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17(4), pp. 693–703, May 2009.

[6] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 9(10), pp. 2009–2025, Oct. 1998.

[7] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley, 2001.

[8] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34(5), pp. 1124–1138, Oct. 1986.

[9] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis," *IEEE Trans. Signal Process.*, vol. 41(10), pp. 3024–3051, Oct. 1993.

[10] M.-Y. Zou, C. Zhenming, and R. Unbehauen, "Separation of periodic signals by using an algebraic method," in *Proc. IEEE Int. Symp. Circuits and Systems*, 1991, vol. 5, pp. 2427–2430.

[11] B. Santhanam and P. Maragos, "Demodulation of Discrete Multicomponent AM-FM Signals using Periodic Algebraic Separation and Energy Demodulation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1997.

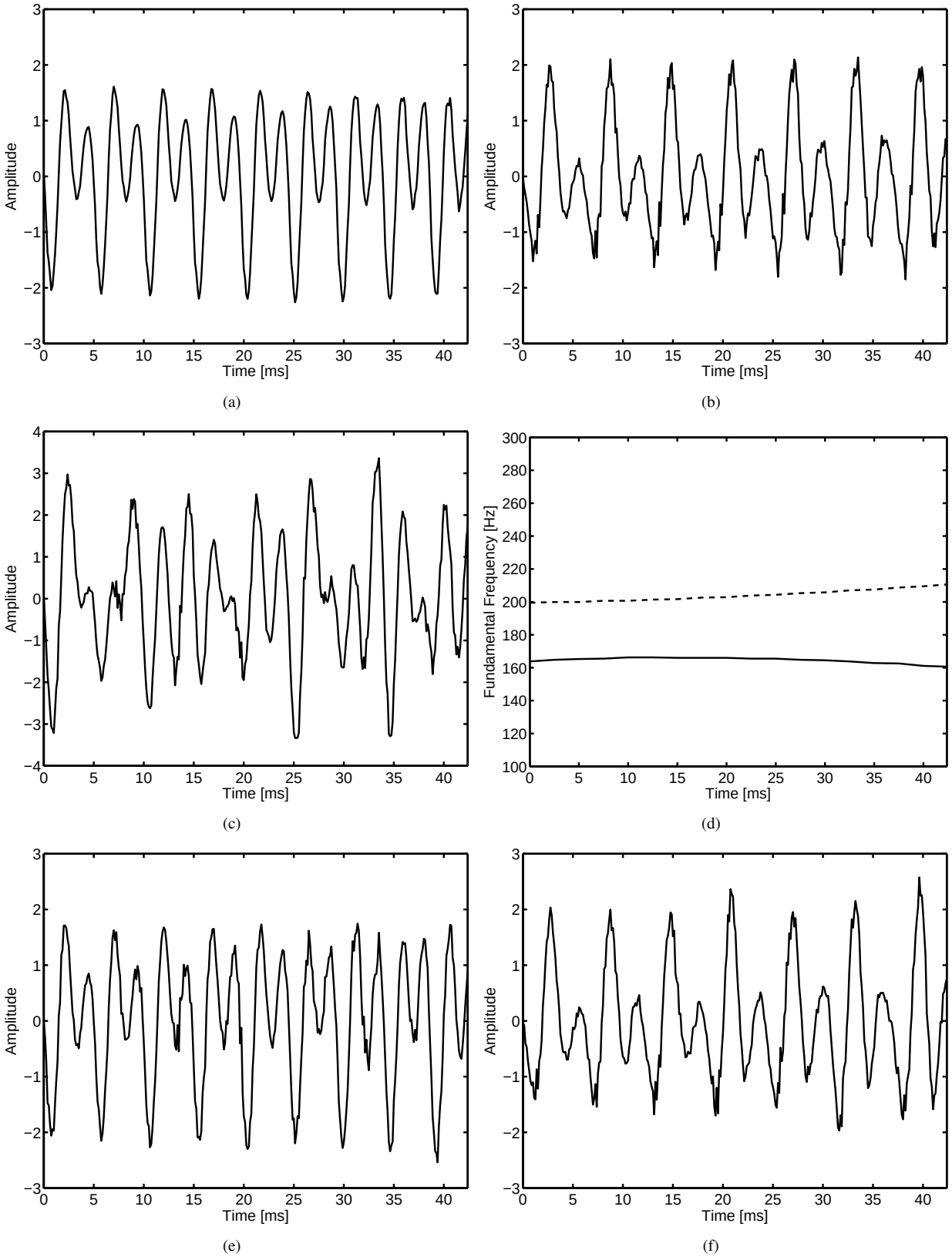


Fig. 7. Shown are the following signals: (a) voiced speech signal of source 1, (b) voiced speech signal of source 2, (c) the mixture of the two signals, (d) the estimated pitch tracks for source 1 (dashed) and 2 (solid), (e) the estimate of source 1 obtained from the mixture, and (f) the estimate of source 2 extracted from the mixture.

[12] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: Survey, new results and an application," *IEEE Trans. Signal Process.*, vol. 48(2), pp. 338–352, Feb. 2000.

[13] K. W. Chan and H. C. So, "Accurate frequency estimation for real

harmonic sinusoids," *IEEE Signal Process. Lett.*, vol. 11(7), pp. 609–612, July 2004.
[14] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111(4), pp. 1917–1930, Apr. 2002.
[15] M. G. Christensen, S. H. Jensen, S. V. Andersen, and A. Jakobsson, "Subspace-based fundamental frequency estimation," in *Proc. European Signal Processing Conf.*, 2004, pp. 637–640.
[16] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15(5), pp. 1635–1644, July 2007.
[17] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Fundamental frequency estimation using the shift-invariance property," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2007, pp. 631–635.
[18] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Signal Processing*, vol. 80, pp. 1937–1944, 2000.
[19] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Elsevier Signal Processing*, vol. 88(4), pp. 972–983, Apr. 2008.
[20] M. Noll, "Pitch determination of human speech by harmonic product spectrum, the harmonic sum, and a maximum likelihood estimate," in *Proc. Symposium on Computer Processing Communications*, 1969, pp. 779–797.
[21] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, vol. 41(2), pp. 293–309, 1967.
[22] A. T. Cemgil, H. J. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 679–694, March 2006.
[23] A. T. Cemgil, *Bayesian Music Transcription*, Ph.D. thesis, Nijmegen University, 2004.
[24] S. Godsill and M. Davy, "Bayesian harmonic models for musical pitch estimation and analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, vol. 2, pp. 1769–1772.
[25] J. Moorer, "The optimum comb method of pitch period analysis of continuous digitized speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 22, no. 5, pp. 330–338, Oct 1974.
[26] J. Lim, A. Oppenheim, and L. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 4, pp. 354–358, Aug 1978.
[27] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, vol. 5 of *Synthesis Lectures on Speech & Audio Processing*, Morgan & Claypool Publishers, 2009.
[28] C. Fevotte and S. J. Godsill, "A Bayesian approach for blind separation of sparse sources," *IEEE Trans. Signal Process.*, vol. 14(6), pp. 2174–2188, Nov. 2006.
[29] J. Li and P. Stoica, "An adaptive filtering approach to spectral estimation and SAR imaging," *IEEE Trans. Signal Process.*, vol. 44(6), pp. 1469–1484, June 1996.
[30] P. Stoica, H. Li, and J. Li, "A new derivation of the APES filter," *IEEE Signal Process. Lett.*, vol. 6(8), pp. 205–206, Aug. 1999.
[31] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Pearson Prentice Hall, 2005.
[32] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, pp. 2726–2735, Oct. 1998.
[33] M. G. Christensen, J. H. Jensen, A. Jakobsson, and S. H. Jensen, "Joint fundamental frequency and order estimation using optimal filtering," in *Proc. European Signal Processing Conf.*, 2009, pp. 1334–1338.
[34] M. G. Christensen, J. H. Jensen, A. Jakobsson, and S. H. Jensen, "On optimal filter designs for fundamental frequency estimation," *IEEE Signal Process. Lett.*, vol. 15, pp. 745–748, 2008.
[35] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57(8), pp. 1408–1418, 1969.
[36] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.
[37] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 9(5), pp. 504–512, June 2001.
[38] O. Yilmaz and S. Reckard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Signal Process. Lett.*, vol. 52(7), July 2004.
[39] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18(3), pp. 538–549, Mar. 2010.
[40] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Signal Process.*, vol. 14(4), pp. 1462–1469, July 2006.
[41] European Broadcasting Union, *Sound Quality Assessment Material Recordings for Subjective Tests*, EBU, Apr. 1988, Tech. 3253.
[42] Center for PersonKommunikation, *Documentation of the Danish EU-ROM.1 Database*, Institute of Electronic Systems, Aalborg University, 1995.

Mads Græsbøll Christensen Mads Græsbøll Christensen was born in Copenhagen, Denmark in March 1977. He received the M.Sc. and Ph.D. degrees in 2002 and 2005, respectively, from Aalborg University in Denmark, where he is also currently employed at the Department of Architecture, Design and Media Technology as Associate Professor. He was previously with the Department of Electronic Systems, Aalborg University and has been a visiting researcher at Philips Research Labs, Ecole Nationale Supérieure des Télécommunications (ENST), and Columbia University. Dr. Christensen has received several awards, namely an IEEE Int. Conf. Acoust. Speech, and Signal Proc. Student Paper Contest Award, the Spar Nord Foundation's Research Prize awarded annually for an excellent Ph.D. thesis, and a Danish Independent Research Council's Young Researcher's Award. He has published more than 75 papers in peer-reviewed conference proceedings and journals is author (with A. Jakobsson) of the book "Multi-Pitch Estimation", Morgan & Claypool Publishers, 2009. He is a Member of the IEEE and an Associate Editor for the IEEE Signal Processing Letters. His research interests include digital signal processing theory and methods with application to speech and audio, in particular parametric analysis, modeling, and coding.

Andreas Jakobsson received his M.Sc. from Lund Institute of Technology and his Ph.D. in Signal Processing from Uppsala University in 1993 and 2000, respectively. Since, he has held positions with Global IP Sound AB, the Swedish Royal Institute of Technology, King's College London, and Karlstad University, as well as an Honorary Research Fellowship at Cardiff University. He is currently Professor of Mathematical Statistics at Lund University, Sweden. He has published his research findings in over 100 refereed journal and conference papers, and has filed four patents. He has also co-authored (together with M. G. Christensen) a recent book on multi-pitch estimation (Morgan & Claypool, 2009). He is a Senior Member of IEEE, a member of the IEEE Sensor Array and Multichannel (SAM) Signal Processing Technical Committee, and an Associate Editor for the IEEE Transactions on Signal Processing, the IEEE Signal Processing Letters and the Research Letters in Signal Processing. His research interests include statistical and array signal processing, detection and estimation theory, and related application in remote sensing, telecommunication and biomedicine.

Paper C

Accurate Estimation of Low Fundamental Frequencies

M. G. Christensen

The paper has been published in the
IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 10, pp. 2042–2056, October 2013.

Accurate Estimation of Low Fundamental Frequencies from Real-Valued Measurements

Mads Græsbøll Christensen, *Senior Member, IEEE*

***Abstract*— In this paper, the difficult problem of estimating low fundamental frequencies from real-valued measurements is addressed. The methods commonly employed do not take the phenomena encountered in this scenario into account and thus fail to deliver accurate estimates. The reason for this is that they employ asymptotic approximations that are violated when the harmonics are not well-separated in frequency, something that happens when the observed signal is real-valued and the fundamental frequency is low. To mitigate this, we analyze the problem and present some exact fundamental frequency estimators that are aimed at solving this problem. These estimators are based on the principles of nonlinear least-squares, harmonic fitting, optimal filtering, subspace orthogonality, and shift-invariance, and they all reduce to already published methods for a high number of observations. In experiments, the methods are compared and the increased accuracy obtained by avoiding asymptotic approximations is demonstrated.**

I. Introduction

Signals that are periodic can be decomposed into a sum of sinusoids having frequencies that are integer multiples of a fundamental frequency, much like the well-known Fourier series, except that real-life signals are noisy and are not observed over an integer number of periods. The problem of finding this fundamental frequency is referred to as fundamental frequency estimation or sometimes as pitch estimation, with the latter term referring to the perceptual attribute that is associated with sound waves exhibiting periodicity. Many signals that can be encountered by the signal processing practitioner are periodic or approximately so. This is, for example, the case in speech processing, where voiced speech exhibits such characteristics, and in music processing for tones produced by musical instruments. Also in the analysis of some bird calls and various other biological signals, like vital signs [1], such signals can be encountered. Moreover, they occur in radar applications for rotating targets [2] and in passive detection, localization, and identification of boats and helicopters [3]. It is then also not surprising that a host of methods have been proposed over the years including methods based on the principles of maximum likelihood, least-squares (LS), and weighted least-squares (WLS) [4]–[8], auto-/cross-correlation and related methods [9]–[13], linear prediction [14], filtering [2], [15]–[17], and subspace methods [18], [19]. We note in passing that several of the cited methods can be interpreted in more than one way and may therefore be considered as belonging to several of the categories above. For an introduction to the fundamental frequency estimation problem and an overview of fundamental frequency estimators, we refer the interested reader to [20].

We are here concerned with a specific problem in determining the fundamental frequency under certain circumstances. When the fundamental frequency of a periodic signal is low as compared to the number of samples, the harmonics of the signal are closely spaced in its spectrum, as the distance between harmonics is given by the fundamental frequency. A similar effect comes into play when the observed signal is real (when we say that some quantity is real, we mean that it is real-valued, i.e., its imaginary part is zero). In this case, harmonics occur in the spectrum not only at positive integer multiples of a fundamental, but also for negative, as complex conjugate pairs of complex sinusoids combine to yield real signals. Again, the distance between the individual complex sinusoids is given by the fundamental frequency. The problem here is that when harmonics are close in frequency, and they are far from being orthogonal, they will interact. As such, this is not really a problem, but most of the parametric methods in the literature ignore this. The reason for this is simple: by ignoring the interaction, one obtains simpler estimators that can be implemented efficiently using, for example, the fast Fourier transform (FFT) or polynomial rooting methods. An example of this is the so-called harmonic summation method [4], in which an approximate maximum likelihood estimate of the fundamental frequency is obtained by summing the power spectral density sampled at candidate fundamental frequencies and picking the one that yields the highest power. This method is accurate when the number of samples approaches infinity, but it fails to take the interaction into account for finite length signals. From the above discussion, it should also be clear that when the fundamental frequency is high relative to the number of available samples, there is essentially no error in using a complex model for a real-valued signal.

Interestingly, the problem of taking the nature of real signals into account has been addressed in the frequency estimation literature, i.e., for the case where sinusoids are not constrained to being integer multiples of a fundamental frequency. Some examples of adaptations of well-known estimators to this problem are for maximum likelihood methods [21], [22], subspace methods [23], [24], Capon's method [25], and the linear prediction [26] method.

It is possible to bound the performance of estimators by computing the Cramér-Rao lower bound (CRLB), which is a lower bound on the variance of an unbiased estimator. This has also been done for the problem of estimating the fundamental frequency [2], [18]. These show that the expected

Part of this work has been presented at the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics 2011 and at the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2013.

M. G. Christensen is with Audio Analysis Lab,Dept. of Architecture, Design and Media Technology, Aalborg University, Denmark, (phone: +45 99 40 97 93, email: mgc@create.aau.dk).

performance (of an optimal estimator) does not depend on the fundamental frequency. At first glance, this seems to contradict the premise of this paper. However, upon closer inspection, it turns out that these bounds were derived based on asymptotic approximations relying on the number of samples approaching infinity or being *sufficiently* large. In former case, the support of spectrum of the sinusoids reduces to a single point, and, hence, the interaction between sinusoids will be zero as long as the fundamental frequency is different from zero, a trivial case that is of no interest anyway.

In this paper, we aim to analyze and solve this problem in a systematic manner. We define the problem of interest with complex and real signal models and analyze it using using what we refer to as the exact CRLB. Then, a number of solutions to the problem are presented, some of which are new, some of which are known, namely a nonlinear least-squares method, an optimal filtering method, a subspace method based on angles between subspaces, and, finally, a method based on a WLS fitting of unconstrained frequencies (called harmonic fitting). The presented methods have in common that they avoid the use of asymptotic approximations, whenever possible, and they take the real-valued nature of the observed signal into account. The nonlinear least-squares method is well-documented in the literature having been applied to many problems, including also frequency estimation and fundamental frequency estimation [5], [6], [8]. The optimal filtering method, which is based on constrained optimization, was originally proposed in [8], but only for complex signals. Here, the underlying constraints are modified to fit real signals. The method based on angles between subspaces is an exact version of the MUSIC-based methods of [8], [18] both of which employ an approximate measure of subspace orthogonality as introduced in [27]. The connection between the exact and approximate measures of the angles between subspaces was first analyzed in [28], but was only used for deriving an approximate, normalized measure for order estimation and, hence, not for fundamental frequency estimation. The harmonic fitting method was originally proposed in [6], but employed a weighting of the individual harmonics derived based on asymptotic properties, while we here avoid using these. In simulations, the effectiveness of these methods is then investigated and their performance compared to the exact CRLB, and the problem is analyzed via comparisons of the asymptotic and exact CRLBs.

The remainder of the present paper is organized as follows: In Section II, we introduce the problem and the signal models and proceed to derive the corresponding CRLB. In the section that follows, namely Section III, we present some methods for solving the problem. We then present the experimental results in Section IV, after which we conclude on our work in Section V.

II. Preliminaries

A. Model and Problem Definition

We will now proceed to define the problem of interest and the associated signal model. The observed real signal $x(n)$ is composed of a set of L sinusoids having frequencies that are integer multiples of a fundamental frequency ω_0, real amplitude $A_l > 0$, and phases $\phi_l \in [0, 2\pi)$. Aside from the sinusoids, we assume that an additive noise source $e(n)$ is present. This noise source represents all stochastic signal components, even those that are integral parts of natural signals that may be of interest to us in other cases. It is here assumed to be white Gaussian distributed having variance σ^2 and zero mean, although this is strictly speaking not necessary for all the presented methods. Mathematically, the observed signal can be expressed for $n = 0, \ldots, N-1$ as

$$x(n) = \sum_{l=1}^{L} A_l \cos\left(\omega_0 l n + \phi_l\right) + e(n). \tag{1}$$

The problem is then to estimate ω_0 from $x(n)$. For a given L, the fundamental frequency can be in the range $\omega_0 \in (0, \frac{\pi}{L})$. Regarding the remaining unknown parameters, some comments are in order. The model order, L, (also referred to as the number of harmonics) can be found a variety of ways and it is possible to solve jointly for the fundamental frequency and the model order, something that has been done for all the methodologies employed here (see [20]), and the extension of these principles to the estimator presented herein is fairly straightforward for which reason we defer from any further discussion of this problem. Once the fundamental frequency and the model order L has been found, the corresponding phases and amplitudes can be found using one of the many existing amplitude estimators [20], [29]. Compared to the problem of estimating the fundamental frequency, this is fairly easy, as these parameters are linear. We note that for $L = 1$, the model above reduces to a single real sinusoid and the associated estimation problem to the usual frequency estimation problem.

Regarding the realism of the model (1), there are several issues that may be a concern. First, the amplitudes, phases and frequencies are assumed to be constant for the duration of the N samples. Since natural sources most often are time-varying, N should be chosen sufficiently low so that the model is a good approximation of the observed signal. Second, the frequencies of the harmonics are assumed to be integer multiples of the fundamental frequency. This should be considered an approximation too, as natural signals may exhibit deviations from this for variety of reasons. We note in passing that a number of modified signal models that take this into account exist [20], [30]. Since these are widely application and signal specific and we wish to retain the generality of the presented material, we will not go further into details on this matter. Third, the noise was assumed to be Gaussian and white. Regarding the Gaussian assumption, this appears to be the norm in the literature, and, in our experience, it does not appear to be a major shortcoming of existing methods used in speech and audio processing. It should also be noted that even though several of the estimators herein are derived based on this assumption, the estimators may still be accurate, at least asymptotically so, even if the assumption does not hold [31]. Moreover, the white Gaussian distribution can be shown to be the one that maximizes the entropy of the noise [32], i.e., it is a worst case scenario. For colored noise, one can apply pre-

whitening [5], [33], i.e., a filtering, to render the noise white, or, at least, more close to being white than it was prior to the pre-whitening. Fourth, the noise was assumed to have zero mean and no DC offset (0 frequency component) is included in the deterministic part of (1). This is mostly done for simplicity. The presence of such a component can, though, be addressed in several ways: a) the presented estimators can be extended by including the zero frequency component having an unknown amplitude [31]; b) the mean can be estimated (and removed) a priori as it is typically caused by calibration errors in microphones and constant outside $n = 0, \ldots, N-1$; c) the signal of interest can be preprocessed using a simple DC blocking filter.

This signal model in (1) can also be expressed using complex sinusoids as

$$x(n) = \sum_{l=-L}^{L} a_l e^{j\omega_0 ln} + e(n), \tag{2}$$

with $a_l = a_{-l}^*$ and $a_0 = 0$. In this notation, the phase and amplitude have been combined into a complex amplitude as $a_l = \frac{A_l}{2} e^{j\phi_l}$ and $(\cdot)^*$ denotes the complex-conjugate. It should be stressed that no additional assumptions have been used in going from (1) to (2), which means that (2) is exact. The error in applying a complex model arises when modifying (2) into $x(n) \approx \sum_{l=1}^{L} a_l e^{j\omega_0 ln} + e(n)$, i.e., when assuming that only half the complex sinusoids are there. This essentially ignores the interaction between the complex sinusoids having frequencies $\{\omega_0 l\}_{l=1}^{L}$ and $\{-\omega_0 l\}_{l=1}^{L}$. Another frequently used approach is to convert (1) into a complex model via the Hilbert transform, which can be used to compute the so-called discrete-time analytic signal. However, the error committed in this process is essentially the same (aside from the suboptimality of the finite-length Hilbert transform), and they are both accurate under the same conditions, namely that ω_0 is not close to 0 relative to N.

B. Cramér-Rao Lower Bound and Further Definitions

An estimator is said to be unbiased if the expected value of its estimate $\hat{\theta}_i$ of the ith parameter θ_i of the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^P$ is identical to the true parameter for all possible values of the true parameter, i.e., $\mathrm{E}\left\{\hat{\theta}_i\right\} = \theta_i \ \forall \theta_i$. The difference, i.e., $\theta_i - \mathrm{E}\left\{\hat{\theta}_i\right\}$, is referred to as the bias. The CRLB is a lower bound on the variance of an unbiased estimate of a parameter, say θ_i, and it is given by $\mathrm{var}(\hat{\theta}_i) \geq \left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{ii}$. Here, the notation $[\mathbf{I}(\boldsymbol{\theta})]_{il}$ means the ilth entry of the matrix $\mathbf{I}(\boldsymbol{\theta})$ and $\mathrm{var}(\cdot)$ denotes the variance. Furthermore, $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information matrix defined as

$$\mathbf{I}(\boldsymbol{\theta}) = -\mathrm{E}\left\{\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right\}, \tag{3}$$

where $p(\mathbf{x};\boldsymbol{\theta})$ is the likelihood function of the observed signal parametrized by the parameters $\boldsymbol{\theta}$. For the case of Gaussian signals with $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}), \mathbf{Q})$ where $\mathbf{Q}$ is the noise covariance matrix (which is not parametrized by any of the parameters in $\boldsymbol{\theta}$) and $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the mean, the likelihood function is given by

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{\det\left(2\pi\mathbf{Q}\right)^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))^T \mathbf{Q}^{-1}(\mathbf{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))}. \tag{4}$$

For this case, Slepian-Bang's formula [34] can be used for determining a more specific expression for the Fisher information matrix. More specifically, it is given by

$$[\mathbf{I}(\boldsymbol{\theta})]_{nm} = \frac{\partial \boldsymbol{\mu}^T(\boldsymbol{\theta})}{\partial \theta_n} \mathbf{Q}^{-1} \frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_m}. \tag{5}$$

For the problem and signal model considered here, the involved quantities are given by:

$$\begin{aligned}
\mathbf{x} &\triangleq \left[\ x(0)\ \cdots\ x(N-1)\ \right]^T \\
\mathbf{Q} &\triangleq \sigma^2 \mathbf{I} \\
\boldsymbol{\theta} &\triangleq \left[\ \omega_0\ A_1\ \phi_1\ \cdots\ A_L\ \phi_L\ \right]^T \\
\boldsymbol{\mu}(\boldsymbol{\theta}) &\triangleq \mathbf{Z}\mathbf{a} \\
\mathbf{Z} &\triangleq \left[\ \mathbf{z}(\omega_0)\ \mathbf{z}^*(\omega_0)\ \cdots\ \mathbf{z}(\omega_0 L)\ \mathbf{z}^*(\omega_0 L)\ \right], \\
\mathbf{a} &\triangleq \frac{1}{2}\left[A_1 e^{j\phi_1}\ A_1 e^{-j\phi_1}\ \cdots\ A_L e^{j\phi_L}\ A_L e^{-j\phi_L}\ \right]^T \\
\mathbf{z}(\omega_0 l) &\triangleq \left[\ 1\ e^{j\omega_0 l 1}\ \cdots\ e^{j\omega_0 l(N-1)}\ \right]^T.
\end{aligned}$$

Note that we will make extensive use of these definitions later. In relation to the problem at hand, some observations about the nature of the matrix $\mathbf{Z}$ can be made: Firstly, for $\omega_0 \neq 0$ and $\omega_0 \in (0, \frac{\pi}{L})$, $\mathbf{Z}$ has full rank. However, for $\omega_0 = 0$, it will be rank deficient and as $\omega_0 \to 0$, the condition number of $\mathbf{Z}$ will tend to infinity and the involved estimation problem is basically ill-posed.

With the above in place, we now have to determine the following derivatives:

$$\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \omega_0} = \frac{\partial \mathbf{Z}}{\partial \omega_0}\mathbf{a}, \frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial A_l} = \mathbf{Z}\frac{\partial \mathbf{a}}{\partial A_l}, \frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \phi_l} = \mathbf{Z}\frac{\partial \mathbf{a}}{\partial \phi_l}, \tag{6}$$

which, in turn, require that the following be computed:

$$\begin{aligned}
\frac{\partial \mathbf{Z}}{\partial \omega_0} &= \left[\ \frac{\partial \mathbf{z}(\omega_0)}{\partial \omega_0}\ \frac{\partial \mathbf{z}^*(\omega_0)}{\partial \omega_0}\ \cdots\ \frac{\partial \mathbf{z}(\omega_0 L)}{\partial \omega_0}\ \frac{\partial \mathbf{z}^*(\omega_0 L)}{\partial \omega_0}\ \right] \\
\frac{\partial \mathbf{z}(\omega_0 l)}{\partial \omega_0} &= \left[\ 0\ jle^{j\omega_0 l}\ \cdots\ j(N-1)le^{j\omega_0 l(N-1)}\ \right]^T \\
\frac{\partial \mathbf{a}}{\partial A_l} &= \frac{1}{2}\left[\ 0\ \cdots\ 0\ e^{j\phi_l}\ e^{-j\phi_l}\ 0\ \cdots\ 0\ \right]^T \\
\frac{\partial \mathbf{a}}{\partial \phi_l} &= \frac{1}{2}\left[\ 0\ \cdots\ 0\ jA_l e^{j\phi_l}\ -jA_l e^{-j\phi_l}\ 0\ \cdots\ 0\ \right]^T.
\end{aligned} \tag{7}$$

For simplicity, we introduce the following definitions:

$$\begin{aligned}
\frac{\partial \mathbf{Z}}{\partial \omega_0}\mathbf{a} &\triangleq \boldsymbol{\alpha}_0 \\
\mathbf{Z}\frac{\partial \mathbf{a}}{\partial A_l} &= \mathrm{Re}\left\{e^{j\phi_l}\mathbf{z}(\omega_0 l)\right\} \triangleq \boldsymbol{\beta}_l \\
\mathbf{Z}\frac{\partial \mathbf{a}}{\partial \phi_l} &= -A_l\,\mathrm{Im}\left\{e^{j\phi_l}\mathbf{z}(\omega_0 l)\right\} \triangleq \boldsymbol{\gamma}_l.
\end{aligned} \tag{8}$$

Here, $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ denote the real and imaginary values, respectively. Note that all the quantities above are real. The

entries in the Fisher information matrix can now be expressed in terms of inner products between these quantities as:

$$\mathbf{I}(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\begin{bmatrix} \boldsymbol{\alpha}_0^T\boldsymbol{\alpha}_0 & \boldsymbol{\alpha}_0^T\boldsymbol{\beta}_1 & \boldsymbol{\alpha}_0^T\boldsymbol{\gamma}_1 & \cdots & \boldsymbol{\alpha}_0^T\boldsymbol{\beta}_L & \boldsymbol{\alpha}_0^T\boldsymbol{\gamma}_L \\ \boldsymbol{\beta}_1^T\boldsymbol{\alpha}_0 & \boldsymbol{\beta}_1^T\boldsymbol{\beta}_1 & \boldsymbol{\beta}_1^T\boldsymbol{\gamma}_1 & \cdots & \boldsymbol{\beta}_1^T\boldsymbol{\beta}_L & \boldsymbol{\beta}_1^T\boldsymbol{\gamma}_L \\ \boldsymbol{\gamma}_1^T\boldsymbol{\alpha}_0 & \boldsymbol{\gamma}_1^T\boldsymbol{\beta}_1 & \boldsymbol{\gamma}_1^T\boldsymbol{\gamma}_1 & \cdots & \boldsymbol{\gamma}_1^T\boldsymbol{\beta}_L & \boldsymbol{\gamma}_1^T\boldsymbol{\gamma}_L \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{\beta}_L^T\boldsymbol{\alpha}_0 & \boldsymbol{\beta}_L^T\boldsymbol{\beta}_1 & \boldsymbol{\beta}_L^T\boldsymbol{\gamma}_1 & \cdots & \boldsymbol{\beta}_L^T\boldsymbol{\beta}_L & \boldsymbol{\beta}_L^T\boldsymbol{\gamma}_L \\ \boldsymbol{\gamma}_L^T\boldsymbol{\alpha}_0 & \boldsymbol{\gamma}_L^T\boldsymbol{\beta}_1 & \boldsymbol{\gamma}_L^T\boldsymbol{\gamma}_1 & \cdots & \boldsymbol{\gamma}_L^T\boldsymbol{\beta}_L & \boldsymbol{\gamma}_L^T\boldsymbol{\gamma}_L \end{bmatrix} \tag{9}$$

The CRLB can now be determined from this by computing the inverse of this matrix and inspecting its diagonal elements. The simple closed form expressions for CRLBs obtained in [2], [18] can be found using the asymptotic orthogonality of complex sinusoids in computing the inner products above. However, we here do not employ this technique as we wish to take into account that the sinusoids are not orthogonal for low fundamental frequencies, and we therefore refer to this CRLB as the exact CRLB. For reference, the asymptotic CRLB for the problem at hand is given by

$$\mathrm{var}(\hat{\omega}_0) \geq \frac{24\sigma^2}{N^3\sum_{l=1}^{L} A_l^2 l^2}. \tag{10}$$

The lower bound can be seen to be determined by the signal-to-noise ratio (SNR) defined (in dB) as

$$SNR = 20\log_{10}\frac{\sum_{l=1}^{L} A_l^2 l^2}{\sigma^2}\,[\mathrm{dB}]\,. \tag{11}$$

An interesting observation can be made from (9): it can be seen that the noise variance is simply a constant factor, and the effect of noise is, hence, unrelated to the problem of low fundamental frequencies. In this connection, it should be noted that this is also the case when the noise variance is uknown [35].

III. Methods

A. Nonlinear Least-Squares

We will now present a number of estimators for solving the problem of interest. The first such method is the nonlinear least-squares (NLS) method, which is based on the principle of maximum likelihood estimation. It is an adaptation of a type of estimator that has appeared in many forms and contexts throughout the years to the problem at hand [4], [5], [8]. The maximum likelihood estimator for the parameters $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x};\boldsymbol{\theta}). \tag{12}$$

Under the assumption that $\mathbf{x}$ is Gaussian distributed and the noise is white, i.e., $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}), \sigma^2\mathbf{I})$, the likelihood function is given by (4). By inserting (4) into (12), taking the logarithm and dropping all constant terms, we obtain:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\|\mathbf{x} - \boldsymbol{\mu}(\boldsymbol{\theta})\right\|^2, \tag{13}$$

where $\|\cdot\|^2$ denotes the vector 2-norm. This shows the well-known result that when the noise is white and Gaussian distributed, the LS method is the maximum likelihood estimator. As before, the mean is determined by the harmonic signal model, i.e., $\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbf{Z}\mathbf{a}$ and the unknown parameters are in this case the fundamental frequency ω_0 that completely characterizes $\mathbf{Z}$ and the vector $\mathbf{a}$ containing the complex amplitudes. This results in the following problem:

$$(\hat{\omega}_0, \hat{\mathbf{a}}) = \arg\min_{\omega_0,\mathbf{a}} \left\|\mathbf{x} - \mathbf{Z}\mathbf{a}\right\|^2. \tag{14}$$

Since we are not really interested in the complex amplitudes, we will substitute these by their maximum likelihood estimate (for a given ω_0), which is $\hat{\mathbf{a}} = \left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{x}$, with $(\cdot)^H$ denoting the Hermitian-transposition. The resulting estimator depends only on ω_0:

$$\hat{\omega}_0 = \arg\max_{\omega_0} \mathbf{x}^T\boldsymbol{\Pi}_Z\mathbf{x}. \tag{15}$$

with $\boldsymbol{\Pi}_Z$ being the orthogonal projection matrix for the space spanned by the columns of $\mathbf{Z}$, i.e., $\boldsymbol{\Pi}_Z = \mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H$. This is the estimator that we will here refer to as the NLS estimator. For each fundamental frequency candidate it involves operations of complexity $\mathcal{O}(L^2N) + \mathcal{O}(L^3) + \mathcal{O}(LN^2) + \mathcal{O}(N^2)$. The estimator does not, however, require any initialization[1], unlike the methods to follow. It should be noted that in assessing the complexity of the various methods, we treat the involved variables, here N and L, as independent variables, although they may not be. The matrix $\mathbf{Z}$ has full rank as long as $\omega_0 \neq 0$ and that $N \geq L$ for the inverse $\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}$ to exist. However, for very small ω_0, numerical effects may render the estimates useless.

The harmonic summation method [4] follows from this by using *the fact* that the columns of $\mathbf{Z}$ are orthogonal asymptotically in N [20]. Although this leads to a computationally efficient implementation based on the fast Fourier transform, this ultimately also leads to the failure of this method for low ω_0 and N.

B. Harmonic Fitting

The idea behind the following method is quite intuitive and appealing due to its simplicity. It is based on the principle of [36] used in [6]. Many different and good methods exist for finding frequencies of sinusoids in an unconstrained manner, meaning that they find frequencies that are not constrained to being integer multiples of a fundamental frequency. The question is then how to find an estimate of the fundamental frequency from these frequencies.

Suppose we find a set of parameter estimates $\hat{\boldsymbol{\eta}}$ from $\mathbf{x}$, and assuming that a maximum likelihood estimator with sufficiently large N is used (and that some regularity conditions are satisfied), the estimates $\hat{\boldsymbol{\eta}}$ are distributed as (see, e.g., [34])

$$\hat{\boldsymbol{\eta}} \sim \mathcal{N}(\boldsymbol{\eta}, \mathbf{I}^{-1}(\boldsymbol{\eta})) \tag{16}$$

where $\mathbf{I}(\boldsymbol{\eta})$ is the Fisher information matrix for the likelihood function for $\boldsymbol{\eta}$ (here, $\boldsymbol{\eta}$ are the true values). Now, suppose that we are not interested in these parameters, but rather in a

[1]In the context of complexity analysis, by initialization we mean that quantities that have to be computed before numerical optimization can be performed to obtain the parameters of interest, i.e. the computation of quantities other than the signal of interest.

different set $\boldsymbol{\theta}$ and that we can find a linear transformation $\mathbf{S}$ that relates these two. Mathematically, this can be stated as

$$\boldsymbol{\eta} = \mathbf{S}\boldsymbol{\theta}. \tag{17}$$

In the following, we assume that $\mathbf{S}$ is real, has full rank and is tall and that both parameter sets are real. Since $\hat{\boldsymbol{\eta}}$ are estimates of $\boldsymbol{\eta}$ and are distributed according to (16), the difference $\hat{\boldsymbol{\eta}} - \mathbf{S}\boldsymbol{\theta}$ is distributed as $\hat{\boldsymbol{\eta}} - \mathbf{S}\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\eta}))$. We can now use this to pose a probability density function of $\hat{\boldsymbol{\eta}}$ as

$$p(\hat{\boldsymbol{\eta}}; \boldsymbol{\theta}) = \frac{1}{\det\left(2\pi \mathbf{I}^{-1}(\boldsymbol{\eta})\right)^{\frac{1}{2}}} e^{-\frac{1}{2}(\hat{\boldsymbol{\eta}} - \mathbf{S}\boldsymbol{\theta})^T \mathbf{I}(\boldsymbol{\theta})(\hat{\boldsymbol{\eta}} - \mathbf{S}\boldsymbol{\theta})}, \tag{18}$$

which can be seen to be parametrized by the unknown parameters $\boldsymbol{\theta}$ and is, hence, a likelihood function. Proceeding now as in Subsection III-A, we can state the maximum likelihood estimate of $\boldsymbol{\theta}$ as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ln p(\hat{\boldsymbol{\eta}}; \boldsymbol{\theta}) \tag{19}$$

$$= \arg\min_{\boldsymbol{\theta}} \left(\hat{\boldsymbol{\eta}} - \mathbf{S}\boldsymbol{\theta}\right)^T \mathbf{I}(\boldsymbol{\eta}) \left(\hat{\boldsymbol{\eta}} - \mathbf{S}\boldsymbol{\theta}\right), \tag{20}$$

which can be seen to be a WLS estimator. Since the signal model is linear, the problem has a closed form solution, which is given by $\hat{\boldsymbol{\theta}} = \left(\mathbf{S}^T \mathbf{I}(\boldsymbol{\eta}) \mathbf{S}\right)^{-1} \mathbf{S}^T \mathbf{I}(\boldsymbol{\eta}) \hat{\boldsymbol{\eta}}$. At this point, some remarks are in order. Firstly, the estimator takes on the form of the solution to a linear LS problem regardless of the original distribution of $\mathbf{x}$. Secondly, the estimates $\hat{\boldsymbol{\eta}}$ need not follow the exact distribution in (16) for (20) to hold; the estimate covariance can be off by a multiplicative factor without affecting the form of the estimator. The principle used in arriving at (20) is known as the extended invariance principle (EXIP), or just the invariance principle, depending on the exact problem [36] (see also [31], [34], [37]). The principle has been applied to fundamental frequency estimation for a complex model and using asymptotic approximations of $\mathbf{I}(\boldsymbol{\eta})$ in [6]. Here, we will use it for a real model and without making use of the aforementioned asymptotic approximation.

The problem now remains to cast the problem of interest in this framework and determine $\boldsymbol{\eta}$, $\mathbf{S}$, and $\mathbf{I}(\boldsymbol{\eta})$. Firstly, for the case of a sinusoidal model with no harmonic constraint, we obtain a set of frequencies $\{\Omega_l \in (0, \pi)\}_{l=1}^{L}$. Moreover, we assume that the frequencies are organized as $\Omega_1 < \ldots < \Omega_L$. Next, we define a parameter set containing the corresponding parameters as $\boldsymbol{\eta} \triangleq \left[\, \Omega_1 \; C_1 \; \Phi_1 \; \cdots \; \Omega_L \; C_L \; \Phi_L \,\right]^T$, where $\{C_l\}$ are the corresponding amplitudes and $\{\Phi_l\}$ the phases. It should be noted that there is no reason to include both positive and negative frequencies as these will be identical (as will the corresponding amplitudes and phases) for estimators tailored to real measurements. The transformation $\mathbf{S} \in \mathbb{R}^{3L \times 2L+1}$ relating these to $\boldsymbol{\theta}$ can easily be confirmed to be given by

$$\mathbf{S} = \begin{bmatrix} 1 & & & & & & \\ 0 & 1 & & & & & \\ 0 & & 1 & & & & \\ 2 & & & & & & \\ 0 & & & 1 & & & \mathbf{0} \\ 0 & & & & 1 & & \\ \vdots & & & & & \vdots & \\ L & & & \mathbf{0} & & & \\ 0 & & & & & 1 & \\ 0 & & & & & & 1 \end{bmatrix}. \tag{21}$$

We can now express $\boldsymbol{\eta}$ in terms of $\boldsymbol{\theta}$ as in (17). The estimator in (20) requires that the true parameters are known to find $\mathbf{I}(\boldsymbol{\eta})$. Instead, we can use an approximation based on the parameter estimates $\hat{\boldsymbol{\eta}}$ (see [37], [38]), i.e., $\mathbf{I}(\boldsymbol{\eta}) \approx \mathbf{I}(\hat{\boldsymbol{\eta}})$, which, for Gaussian signals, is given by

$$\left[\mathbf{I}(\hat{\boldsymbol{\eta}})\right]_{nm} = \left. \frac{\partial \boldsymbol{\mu}^T(\boldsymbol{\eta})}{\partial \eta_n} \mathbf{Q}^{-1} \frac{\partial \boldsymbol{\mu}(\boldsymbol{\eta})}{\partial \eta_m} \right|_{\boldsymbol{\eta} = \tilde{\boldsymbol{\eta}}}, \tag{22}$$

with $\mathbf{Q}$ being the covariance matrix of the observation noise and $\boldsymbol{\mu}(\boldsymbol{\eta})$ being the mean of the same signal parametrized in terms of $\boldsymbol{\eta}$. The approximation above is essentially valid due to the maximum likelihood estimates $\hat{\boldsymbol{\eta}}$ being consistent estimates of $\boldsymbol{\eta}$ and $\mathbf{I}(\boldsymbol{\eta})$ being a continuous function. For the particular parametrization used here, i.e., the unconstrained model, $\mathbf{I}(\hat{\boldsymbol{\eta}})$ can be shown to be the following (see, e.g., [34]):

$$\mathbf{I}(\hat{\boldsymbol{\eta}}) = \frac{1}{\sigma^2} \begin{bmatrix} \boldsymbol{\Xi}_{11} & \cdots & \boldsymbol{\Xi}_{1L} \\ \vdots & & \vdots \\ \boldsymbol{\Xi}_{L1} & \cdots & \boldsymbol{\Xi}_{LL} \end{bmatrix}, \tag{23}$$

where the individual blocks are given by

$$\boldsymbol{\Xi}_{kl} = \begin{bmatrix} \boldsymbol{\delta}_k^T \boldsymbol{\delta}_l & \boldsymbol{\delta}_k^T \boldsymbol{\epsilon}_l & \boldsymbol{\delta}_k^T \boldsymbol{\zeta}_l \\ \boldsymbol{\epsilon}_k^T \boldsymbol{\delta}_l & \boldsymbol{\epsilon}_k^T \boldsymbol{\epsilon}_l & \boldsymbol{\epsilon}_k^T \boldsymbol{\zeta}_l \\ \boldsymbol{\zeta}_k^T \boldsymbol{\delta}_l & \boldsymbol{\zeta}_k^T \boldsymbol{\epsilon}_l & \boldsymbol{\zeta}_k^T \boldsymbol{\zeta}_l \end{bmatrix}. \tag{24}$$

The entries in these blocks involve a number of quantities defined as

$$\begin{aligned} \boldsymbol{\delta}_l &\triangleq \hat{C}_l \operatorname{Re}\left\{ e^{j\hat{\Phi}_l} \left[\, 0 \;\; j e^{j\hat{\Omega}_l} \;\; \cdots \;\; j(N-1) e^{j\hat{\Omega}_l (N-1)} \,\right]^T \right\} \\ \boldsymbol{\epsilon}_l &\triangleq \operatorname{Re}\left\{ e^{j\hat{\Phi}_l} \mathbf{z}(\hat{\Omega}_l) \right\} \\ \boldsymbol{\zeta}_l &\triangleq -\hat{C}_l \operatorname{Im}\left\{ e^{j\hat{\Phi}_l} \mathbf{z}(\hat{\Omega}_l) \right\}. \end{aligned} \tag{25}$$

We note that the usually used expression for the CRLB for the unconstrained model is obtained, as before, by applying asymptotic approximations. More specifically, this leads to a block-diagonal structure in (23) as the off-diagonal blocks are approximately equal to zero, i.e., $\boldsymbol{\Xi}_{kl} = \mathbf{0}$,for $l \neq k$. Moreover, the individual blocks on the diagonal exhibit a block-diagonal structure themselves, hence leading to simple closed-form expressions.

Returning to the task at hand, we, finally, arrive at the estimator:

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{S}^T \mathbf{I}(\hat{\boldsymbol{\eta}}) \mathbf{S}\right)^{-1} \mathbf{S}^T \mathbf{I}(\hat{\boldsymbol{\eta}}) \hat{\boldsymbol{\eta}}. \tag{26}$$

The processing steps of the estimator can be summarized as follows: First, estimate the parameters in $\hat{\boldsymbol{\eta}}$ and, second,

compute $\mathbf{I}(\hat{\boldsymbol{\eta}})$ from these parameters. Third, compute the parameter of $\hat{\boldsymbol{\theta}}$ from the aforementioned quantities along with $\mathbf{S}$, which is not signal-dependent. The fundamental frequency can now simply be extracted from the first element of $\hat{\boldsymbol{\theta}}$. Obviously, this process can be simplified somewhat if only the fundamental frequency is desired by determining only the first row of the matrix $\left(\mathbf{S}^T\mathbf{I}(\hat{\boldsymbol{\eta}})\mathbf{S}\right)^{-1}\mathbf{S}^T\mathbf{I}(\hat{\boldsymbol{\eta}})$. As was demonstrated in [6], this methodology proved quite successful even with a number of asymptotic approximation, and we thus also expect it to perform well for our problem. Given the initial estimates $\hat{\boldsymbol{\eta}}$, the estimator has complexity $\mathcal{O}(L^3)$, but unlike the NLS method, it is in closed-form.

C. Optimal Filtering

The next solution to the problem under consideration is based on optimal filtering, which was first used for fundamental frequency estimation in [8] (see also [16]). Before providing more details on this, we introduce some notation and definitions. First, we define the output signal $\hat{x}(n)$ of the length M filter having real coefficients $h(n)$ as

$$\hat{x}(n) = \sum_{m=0}^{M-1} h(m)x(n-m) \triangleq \mathbf{h}^T\mathbf{x}(n), \tag{27}$$

with $\mathbf{h}$ being a vector containing the filter coefficients of the filter defined as $\mathbf{h} = [\ h(0)\ \cdots\ h(M-1)\ ]^T$ and $\mathbf{x}(n) = [\ x(n)\ x(n-1)\ \cdots\ x(n-M+1)\ ]^T$. For our signal model, the output signal $\hat{x}(n)$ can be thought of as an estimate of the periodic parts of the signal. The output power of the filter can be expressed in terms of the covariance matrix $\mathbf{R}$ as $\mathrm{E}\left\{|\hat{x}(n)|^2\right\} = \mathbf{h}^T\mathbf{R}\mathbf{h}$. The question is now how to design the filter such that $\hat{x}(n)$ actually resembles a periodic signal. Such a filter should have a frequency response that allows the periodic components to pass undistorted while suppressing everything else. This means that the frequency response should be one for all the harmonic frequencies, and, since we are here concerned with real signals, this should be the case also for the negative frequencies. One can think of filters having these properties as a kind of comb filter. Mathematically, we can state this as the following optimization problem:

$$\min_{\mathbf{h}} \mathbf{h}^T\mathbf{R}\mathbf{h} \quad \text{s.t.} \quad \mathbf{Z}^H\mathbf{h} = \mathbf{1} \tag{28}$$

with $\mathbf{1} = [\ 1\ \cdots\ 1\ ]^T \in \mathbb{R}^{2L}$. We here remind the reader that $\mathbf{Z} \in \mathbb{C}^{M\times 2L}$ contains all the sinusoids of the real signal model, so the constraints state that the frequency response of the filter must be one for both positive and negative frequencies.

To solve the optimization problem, we introduce the Lagrange multipliers $\boldsymbol{\lambda} = [\ \lambda_1\ \cdots\lambda_{2L}\ ]^T$, and the Lagrangian dual function associated with the problem, which can be written as $\mathcal{L}(\mathbf{h},\boldsymbol{\lambda}) = \mathbf{h}^T\mathbf{R}\mathbf{h} - \boldsymbol{\lambda}^T\left(\mathbf{Z}^H\mathbf{h}-\mathbf{1}\right)$. Taking the derivative with respect to the filter coefficients and the Lagrange multipliers and setting the result equal to zero and solving for the unknowns, leads to the optimal filter $\mathbf{h} = \mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{R}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1}$. The output power of this filter can then be expressed compactly as $\mathbf{h}^T\mathbf{R}\mathbf{h} = \mathbf{1}^H\left(\mathbf{Z}^H\mathbf{R}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1}$. Since the optimal filter depends on the observed signal via $\mathbf{R}$, the resulting filter can be thought of as an adaptive comb filter.

The filter can be used for determining the fundamental frequency in the following way: for a candidate fundamental frequency, the filter passes the candidate harmonics while it suppresses everything else. Therefore, the fundamental frequency can be identified as the value for which the output power of the filter is the highest. In math, this can be stated as

$$\hat{\omega}_0 = \arg\max_{\omega_0} \mathbf{1}^H\left(\mathbf{Z}^H\mathbf{R}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1}. \tag{29}$$

For complex signals, this type of solution was demonstrated to have excellent performance under very adverse conditions in [8], effectively decoupling the multi-pitch estimation problem into a set of single-pitch problems. The estimator in (29) requires initialization of complexity $\mathcal{O}(M^3)$ for computing $\mathbf{R}^{-1}$ while for each fundamental frequency candidate, it requires computations of complexity $\mathcal{O}(L^3)+\mathcal{O}(ML^2)+\mathcal{O}(M^2L)$.

The method requires that the covariance matrix is replaced by an estimate. We use here the usual estimator, the sample covariance matrix, i.e.,

$$\mathbf{R} \approx \frac{1}{N-M+1}\sum_{n=M-1}^{N-1}\mathbf{x}(n)\mathbf{x}^T(n). \tag{30}$$

Since the method also requires that this matrix is invertible, it follows that the filter length must be chosen such that $M < \frac{N}{2}+1$, although it is well-documented in the literature that M in practice should not be chosen too close to this bound. Moreover, we also require that $M \geq L$ for the matrix inverse in (29) to exist. Combined, this allows us to bound M as $2L \leq M \leq \frac{N}{2}+1$. It should also be noted that M should be chosen proportionally to N for the estimator to be consistent. This is also the case for the other methods presented later.

D. Angles between Subspace

The next method is a subspace method reminiscent of MUSIC [27], a method that has previously been applied to the fundamental frequency estimation problem in [8], [18]. It builds on more recent ideas presented in [20], [28]. In MUSIC, an estimate of a basis for the noise subspace is obtained via the eigenvalue decomposition of the sample covariance matrix. This is then used for estimation purposes by choosing the candidate model that is the closest to being orthogonal to that space. This is also the idea we here pursue, although the present methods differs in a fundamental way, namely in terms of how the angles between the subspaces are measured. Let $\mathbf{x}(n) = [\ x(n)\ x(n+1)\ \cdots\ x(n+M-1)\ ]^T$. We can then express this vector as

$$\mathbf{x}(n) = \mathbf{Z}\mathbf{a} + \mathbf{e}(n), \tag{31}$$

with $\mathbf{Z} \in \mathbb{C}^{M\times 2L}$ being defined as in (6) except that the columns have length M and $\mathbf{e}(n) = [\ e(n)\ e(n+1)\ \cdots\ e(n+M-1)\ ]^T$. The covariance matrix[2] of this vector is given by

$$\mathbf{R} = \mathrm{E}\left\{\mathbf{x}(n)\mathbf{x}^H(n)\right\} = \mathbf{Z}\mathbf{P}\mathbf{Z}^H + \sigma^2\mathbf{I} \tag{32}$$

[2]The reader should be aware that our definitions of $\mathbf{x}(n)$ and $\mathbf{R}$ here differ from those in Section III-C.

where $\mathrm{E}\left\{\mathbf{a}\mathbf{a}^H\right\} = \mathbf{P}$, which is given by

$$\mathbf{P} = \mathrm{E}\left\{\begin{bmatrix} a_1a_1^* & a_1^*a_1^* & \dots & a_1a_L^* & a_1^*a_L^* \\ a_1a_1 & a_1^*a_1 & \dots & a_1a_L & a_1^*a_L \\ \vdots & \vdots & & \vdots & \vdots \\ a_La_1^* & a_L^*a_1^* & \dots & a_La_L^* & a_L^*a_L^* \\ a_La_1 & a_L^*a_1 & \dots & a_La_L & a_L^*a_L \end{bmatrix}\right\}. \quad (33)$$

This matrix can be seen to involve block-matrices of the following form:

$$\mathbf{P}_{kl} = \mathrm{E}\left\{\begin{bmatrix} a_ka_l^* & a_k^*a_l^* \\ a_ka_l & a_k^*a_l \end{bmatrix}\right\}. \quad (34)$$

Next, we will analyze the behavior of this matrix assuming that the phases ϕ_l are uniformly distributed and independent over l. This means that $\mathrm{E}\left\{\frac{A_k}{2}e^{j\phi_k}\right\} = 0$ and that $\mathrm{E}\left\{\frac{A_k}{2}e^{j\phi_k}\frac{A_l}{2}e^{-j\phi_l}\right\} = \frac{A_k}{2}\mathrm{E}\left\{e^{j\phi_k}\right\}\frac{A_l}{2}\mathrm{E}\left\{e^{-j\phi_l}\right\} = 0$ for $k \neq l$. Hence, we obtain that for $k \neq l$, the matrix $\mathbf{P}_{kl}$ is simply $\mathbf{P}_{kl} = \mathbf{0}$. For $k = l$, we obtain

$$\mathbf{P}_{ll} = \begin{bmatrix} \frac{A_l^2}{4} & 0 \\ 0 & \frac{A_l^2}{4} \end{bmatrix}, \quad (35)$$

as $\mathrm{E}\left\{\frac{A_l}{2}e^{j\phi_l}\frac{A_l}{2}e^{-j\phi_l}\right\} = \frac{A_l^2}{4}$ and $\mathrm{E}\left\{\frac{A_l}{2}e^{j\phi_l}\frac{A_l}{2}e^{j\phi_l}\right\} = \frac{A_l^2}{4}\mathrm{E}\left\{e^{2j\phi_l}\right\} = 0$. Therefore, the amplitude covariance matrix $\mathbf{P}$ takes on the form $\mathbf{P} = \frac{1}{4}\mathrm{diag}\left(\begin{bmatrix} A_1^2 & A_1^2 & \cdots & A_L^2 & A_L^2 \end{bmatrix}\right)$, which means that the diagonal structure obtained for complex signals is retained for real signals, and the so-called covariance matrix model, therefore, still holds. We note that the assumptions that lead to this model are sufficient but not necessary conditions.

The eigenvalue decomposition (EVD) of the covariance matrix is $\mathbf{R} = \mathbf{U}\mathbf{\Gamma}\mathbf{U}^H$, where $\mathbf{\Gamma}$ is a diagonal matrix containing the positive eigenvalues, γ_k, ordered as $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_M$. Moreover, it can easily be seen that $\gamma_{2L+1} = \ldots = \gamma_M = \sigma^2$. The covariance matrix is positive definite and symmetric by construction. Therefore, $\mathbf{U}$ contains the M orthonormal vectors, which are eigenvectors of $\mathbf{R}$. We will denote these as $\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_M \end{bmatrix}$. Let $\mathbf{S}$ be formed from a subset of the columns of this matrix as

$$\mathbf{S} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_{2L} \end{bmatrix}. \quad (36)$$

We denote the subspace spanned by the columns of $\mathbf{S}$ as $\mathcal{S} = \mathcal{R}(\mathbf{S})$ and refer to it as the signal subspace. Similarly, let $\mathbf{G}$ be formed from the remaining eigenvectors as

$$\mathbf{G} = \begin{bmatrix} \mathbf{u}_{2L+1} & \cdots & \mathbf{u}_M \end{bmatrix}. \quad (37)$$

We refer to the space $\mathcal{G} = \mathcal{R}(\mathbf{G})$ as the noise subspace. Using these definitions, we now obtain $\mathbf{U}\left(\mathbf{\Gamma} - \sigma^2\mathbf{I}\right)\mathbf{U}^H = \mathbf{Z}\mathbf{P}\mathbf{Z}^H$ as the identity matrix is diagonalized by an arbitrary orthonormal basis. Introducing $\mathbf{\Psi} = \mathrm{diag}(\begin{bmatrix} \gamma_1 - \sigma^2 & \cdots & \gamma_{2L} - \sigma^2 \end{bmatrix})$, this leads to the following partitioning of the EVD:

$$\mathbf{R} = \begin{bmatrix} \mathbf{S} & \mathbf{G} \end{bmatrix}\left(\begin{bmatrix} \mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \sigma^2\mathbf{I}\right)\begin{bmatrix} \mathbf{S}^H \\ \mathbf{G}^H \end{bmatrix}, \quad (38)$$

which shows that we may write $\mathbf{S}\mathbf{\Psi}\mathbf{S}^H = \mathbf{Z}\mathbf{P}\mathbf{Z}^H$. As the columns of $\mathbf{S}$ and $\mathbf{G}$ are orthogonal and $\mathcal{R}(\mathbf{Z}) = \mathcal{R}(\mathbf{S})$, it follows that $\mathbf{Z}^H\mathbf{G} = \mathbf{0}$, which is the subspace orthogonality principle used in the MUSIC algorithm [27], [39].

In practice, the estimated noise subspace eigenvectors will not be perfect due to the observation noise and finite observation lengths. The above relation is, therefore, only approximate and a measure must be introduced to determine how close a candidate model $\mathbf{Z}$ is to being orthogonal to $\mathbf{G}$. Traditionally, this has been done using the Frobenius norm. However, this only measures the sum of cosine to the non-trivial angles squared between the two spaces for orthogonal vectors in both $\mathbf{Z}$ and $\mathbf{G}$, and, since we are here concerned with low frequencies, the asymptotic orthogonality of the column of $\mathbf{Z}$ is not accurate. We therefore measure the orthogonality as follows. The principal angles $\{\xi_k\}$ between the two subspaces $\mathcal{Z}$ and $\mathcal{G}$ are defined recursively for $k = 1, \ldots, K$ as [40]

$$\cos(\xi_k) = \max_{\mathbf{u}\in\mathcal{Z}}\max_{\mathbf{v}\in\mathcal{G}}\frac{\mathbf{u}^H\mathbf{v}}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2} \triangleq \mathbf{u}_k^H\mathbf{v}_k, \quad (39)$$

where K is the minimal dimension of the two subspaces, i.e., $K = \min\{2L, M - 2L\}$ and $\mathbf{u}^H\mathbf{u}_i = 0$ and $\mathbf{v}^H\mathbf{v}_i = 0$ for $i = 1, \ldots, k-1$. The angles are bounded and ordered as $0 \leq \xi_1 \leq \ldots \leq \xi_K \leq \frac{\pi}{2}$. Given the orthogonal projection matrices for $\mathcal{Z}$ and $\mathcal{G}$, denoted $\mathbf{\Pi}_Z$ and $\mathbf{\Pi}_G$, respectively, the expression in (39) can be written as

$$\cos(\xi_k) = \max_{\mathbf{y}}\max_{\mathbf{z}}\frac{\mathbf{y}^H\mathbf{\Pi}_Z\mathbf{\Pi}_G\mathbf{z}}{\|\mathbf{y}\|_2\|\mathbf{z}\|_2} \quad (40)$$

$$= \mathbf{y}_k^H\mathbf{\Pi}_Z\mathbf{\Pi}_G\mathbf{z}_k = \kappa_k. \quad (41)$$

As can be seen, $\{\kappa_k\}$ are the ordered singular values of the matrix product $\mathbf{\Pi}_Z\mathbf{\Pi}_G$, and the two sets of vectors $\{\mathbf{y}\}$ and $\{\mathbf{z}\}$ are the left and right singular vectors of the matrix product, respectively. The singular values are related to the Frobenius norm of $\mathbf{\Pi}_Z\mathbf{\Pi}_G$ and hence its trace, denoted with $\mathrm{Tr}\{\cdot\}$, as $\|\mathbf{\Pi}_Z\mathbf{\Pi}_G\|_F^2 = \sum_{k=1}^K \kappa_k^2$ which shows that if the Frobenius norm of the product is zero, then all the non-trivial angles are $\frac{\pi}{2}$, i.e., the two subspaces are orthogonal. This expression can be used to find the fundamental frequency as

$$\hat{\omega}_0 = \arg\min_{\omega_0}\|\mathbf{\Pi}_Z\mathbf{\Pi}_G\|_F^2, \quad (42)$$

and the estimate can be seen to be the value for which the sum of cosine to the angles squared is the least. Finally, (42) can be expressed as

$$\hat{\omega}_0 = \arg\min_{\omega_0}\mathrm{Tr}\left\{\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{G}\mathbf{G}^H\right\}, \quad (43)$$

which is asymptotically equivalent to the fundamental frequency estimator in [18] but different for finite M and N in that it takes the non-orthogonality of the sinusoids for low M and ω_0 into account. Hence, it can be expected to yield superior estimates for low fundamental frequencies. This estimator requires that a number of quantities are computed in the initialization, i.e., only once, namely the EVD of $\mathbf{R}$ and the projection matrix for the noise subspace, which results in a complexity of $\mathcal{O}((M-L)M^2) + \mathcal{O}(M^3)$ (which is obviously only valid for $L < M$). For each candidate fundamental frequency, operations having complexity $\mathcal{O}(L^2M) + \mathcal{O}(M^2L) + \mathcal{O}(L^3)$ are computed.

As for the covariance matrix, it has to be estimated and its dimensions chosen. For this method, this is done as described in (30), only with a different definition of $\mathbf{x}(n)$ as described earlier in this section. Unlike the optimal filtering method, it is not required for this method that the estimated matrix has full rank. It must, however, allow for the estimation of a basis for the signal subspace, which requires that $M \leq N - 2L + 1$. Additionally, for the orthogonal complement to the signal subspace to be non-empty, $M \geq 2L + 1$, which means that we obtain the following inequality for M:

$$2L + 1 \leq M \leq N - 2L + 1. \tag{44}$$

E. Shift-Invariance

The final estimator is also a subspace method and thus builds on the same matrix covariance model as in Section III-D. The last method was based on the noise subspace eigenvectors, while the present one is based on the signal subspace eigenvectors. More specifically, it is based on the principle used in [19]. The signal subspace is given by $\mathcal{S} = \mathcal{R}(\mathbf{S})$ with the matrix $\mathbf{S}$ being defined as in (36). As established earlier, the columns of $\mathbf{S}$ span the same space as the columns of $\mathbf{Z}$, i.e., $\mathcal{R}(\mathbf{S}) = \mathcal{R}(\mathbf{Z})$. Therefore, we may express the relation between these matrices as $\mathbf{S} = \mathbf{Z}\mathbf{B}$ where

$$\mathbf{B} = \mathbf{P}\mathbf{Z}^H\mathbf{S}\mathbf{\Psi}^{-1}, \tag{45}$$

with $\mathbf{B}$ being a square and full rank matrix as both $\mathbf{S}$ and $\mathbf{Z}$ do, and it is hence invertible, something that we will make use of later. The matrix $\mathbf{Z}$ exhibits a particular structure, known as shift-invariance. This property can be expressed in the following way. Define the matrices $\underline{\mathbf{Z}}$ and $\overline{\mathbf{Z}}$ by removing the last and first rows of $\mathbf{Z}$, i.e., $\underline{\mathbf{Z}} = [\ \mathbf{I}\ \mathbf{0}\ ]\,\mathbf{Z}$ and $\overline{\mathbf{Z}} = [\ \mathbf{0}\ \mathbf{I}\ ]\,\mathbf{Z}$ where it follows that $\mathbf{I}$ is $(M-1)\times(M-1)$. Now, doing the same for $\mathbf{S}$ we obtain $\underline{\mathbf{S}} = [\ \mathbf{I}\ \mathbf{0}\ ]\,\mathbf{S}$ and $\overline{\mathbf{S}} = [\ \mathbf{0}\ \mathbf{I}\ ]\,\mathbf{S}$. From these definitions and (45), it can easily be seen that $\underline{\mathbf{S}}$ and $\underline{\mathbf{Z}}$ are related as $\underline{\mathbf{S}} = \underline{\mathbf{Z}}\mathbf{B}$. More importantly, however, due to the particular structure of the model, the matrices $\overline{\mathbf{Z}}$ and $\underline{\mathbf{Z}}$ can be related as $\overline{\mathbf{Z}} = \underline{\mathbf{Z}}\mathbf{D}$ where

$$\mathbf{D} = \operatorname{diag}\left(\left[\ e^{j\omega_0}\ e^{-j\omega_0}\ \cdots\ e^{j\omega_0 L}\ e^{-j\omega_0 L}\ \right]\right). \tag{46}$$

This property is known as shift-invariance. However, since we are interested in finding the parameters that characterize $\mathbf{Z}$, this is of little use by itself. From the above it also follows that $\overline{\mathbf{S}} = \underline{\mathbf{S}}\mathbf{\Sigma}$ and the matrix relating $\underline{\mathbf{S}}$ to $\overline{\mathbf{S}}$ can be shown to be (see, e.g., [41])

$$\mathbf{\Sigma} = \mathbf{B}^{-1}\mathbf{D}\mathbf{B}, \tag{47}$$

i.e., the matrix $\mathbf{\Sigma}$ has the frequencies of the harmonics as the arguments of its eigenvalues. Since $\mathbf{S}$ and hence $\overline{\mathbf{S}}$ and $\underline{\mathbf{S}}$ are known from the EVD of the sample covariance matrix, this is useful in the following way: Given $\underline{\mathbf{S}}$ to $\overline{\mathbf{S}}$, we can solve for $\mathbf{\Sigma}$, from which we can find the frequencies via the EVD. Since the sample covariance will be corrupted by noise in practice, so will $\underline{\mathbf{S}}$ and $\overline{\mathbf{S}}$, and, consequently the above relations will only hold approximately, i.e., $\overline{\mathbf{S}} \approx \underline{\mathbf{S}}\mathbf{\Sigma}$, which means we have to introduce some way of finding $\mathbf{\Sigma}$. Here, we proceed by estimating $\mathbf{\Sigma}$ using total least-squares (TLS) as follows. Define $\underline{\mathbf{\Delta}}$ and $\overline{\mathbf{\Delta}}$ as the minimal perturbations of $\underline{\mathbf{S}}$ to $\overline{\mathbf{S}}$, respectively:

$$\min \left\| \left[\overline{\mathbf{\Delta}}\ \underline{\mathbf{\Delta}}\right] \right\|_F \quad \text{s.t.} \quad \overline{\mathbf{S}} + \overline{\mathbf{\Delta}} = (\underline{\mathbf{S}} + \underline{\mathbf{\Delta}})\,\mathbf{\Sigma}. \tag{48}$$

An estimate $\widehat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$ is then obtained as the solution to $\overline{\mathbf{S}} + \overline{\mathbf{\Delta}} = (\underline{\mathbf{S}} + \underline{\mathbf{\Delta}})\,\mathbf{\Sigma}$ for the perturbations solving (48) (see [41] for further details).

The frequencies obtained from the eigenvalues of $\widehat{\mathbf{\Sigma}}$ are not constrained to being integer multiples of a fundamental frequency, i.e., they are unconstrained frequencies, and, hence, cannot be used directly for estimating the fundamental frequency. Much like for the WLS method in Section III-B, we must fit a fundamental frequency to these frequencies. We now proceed to express $\widehat{\mathbf{\Sigma}}$ in terms of the empirical EVD as

$$\widehat{\mathbf{\Sigma}} = \mathbf{C}\widehat{\mathbf{D}}\mathbf{C}^{-1} \tag{49}$$

with $\mathbf{C}$ containing the empirical eigenvectors of $\widehat{\mathbf{\Sigma}}$ and

$$\widehat{\mathbf{D}} = \operatorname{diag}\left(\left[\ e^{j\hat{\Omega}_1^+}\ e^{j\hat{\Omega}_1^-}\ \cdots\ e^{j\hat{\Omega}_L^+}\ e^{j\hat{\Omega}_{2L}^-}\ \right]\right). \tag{50}$$

We here denote the estimated frequencies as $\{\hat{\Omega}_l^+ \in (0,\pi)\}_{l=1}^L$ and $\{\hat{\Omega}_l^- \in (-\pi,0)\}_{l=1}^L$. Moreover, we assume that they are ordered $\hat{\Omega}_1^+ < \ldots < \hat{\Omega}_L^+$ and $\hat{\Omega}_1^- > \ldots > \hat{\Omega}_L^-$ and that the corresponding eigenvectors in $\mathbf{C}$ are ordered accordingly.

Recall that $\overline{\mathbf{S}} = \underline{\mathbf{S}}\mathbf{B}^{-1}\mathbf{D}\mathbf{B}$, and thus $\overline{\mathbf{S}}\mathbf{C} \approx \underline{\mathbf{S}}\mathbf{C}\mathbf{D}$, where $\mathbf{D}$ depends on the unknown fundamental frequency ω_0. We can now introduce a metric that measures the extent to which the right and left side resemble each other as $\|\overline{\mathbf{S}}\mathbf{C} - \underline{\mathbf{S}}\mathbf{C}\mathbf{D}\|_F^2$. This expression can be expanded as

$$\|\overline{\mathbf{S}}\mathbf{C} - \underline{\mathbf{S}}\mathbf{C}\mathbf{D}\|_F^2 = -2\,\mathrm{Re}\left(\mathrm{Tr}\left\{\overline{\mathbf{S}}\mathbf{C}\mathbf{D}^H\mathbf{C}^H\underline{\mathbf{S}}^H\right\}\right) \tag{51}$$

$$+\,\mathrm{Tr}\left\{\overline{\mathbf{S}}\mathbf{C}\mathbf{C}^H\overline{\mathbf{S}}^H\right\} + \mathrm{Tr}\left\{\underline{\mathbf{S}}\mathbf{C}\mathbf{C}^H\underline{\mathbf{S}}^H\right\}. \tag{52}$$

Noting that the last two terms do not depend on ω_0 and introducing $\delta_l = \left[\mathbf{C}^H\underline{\mathbf{S}}^H\overline{\mathbf{S}}\mathbf{C}\right]_{ll}$, we finally obtain the estimator

$$\hat{\omega}_0 = \arg\max_{\omega_0} 2\,\mathrm{Re}\left\{\sum_{l=1}^{L}\delta_{2l-1}e^{-j\omega_0 l} + \delta_{2l}e^{j\omega_0 l}\right\}. \tag{53}$$

As can be seen, the resulting estimator is extremely simple having complexity $\mathcal{O}(L)$ for each fundamental frequency candidate, albeit the initialization, i.e., the computation of δ_l, is somewhat complex. More specifically, it requires computations of complexity $\mathcal{O}(M^3) + \mathcal{O}(L^3) + \mathcal{O}(M^2L) + \mathcal{O}(L^2M)$. We also note that the involved cost function is generally smooth and well-behaved. Regarding the size of the covariance matrix, M should be chosen according to (44) for obtaining a rank $2L$ estimate of $\mathbf{S}$ and for $\mathbf{\Sigma}$ to be unique.

IV. Experimental Results

A. Exact vs. Asymptotic Bounds

We will start out the experimental part of this paper by exploring the difference between the exact and asymptotic CRLBs for the problem of estimating the fundamental frequency and the dependency of this difference on various parameters. This is interesting for a number of reasons. Many of the estimators derived based on complex models are based on

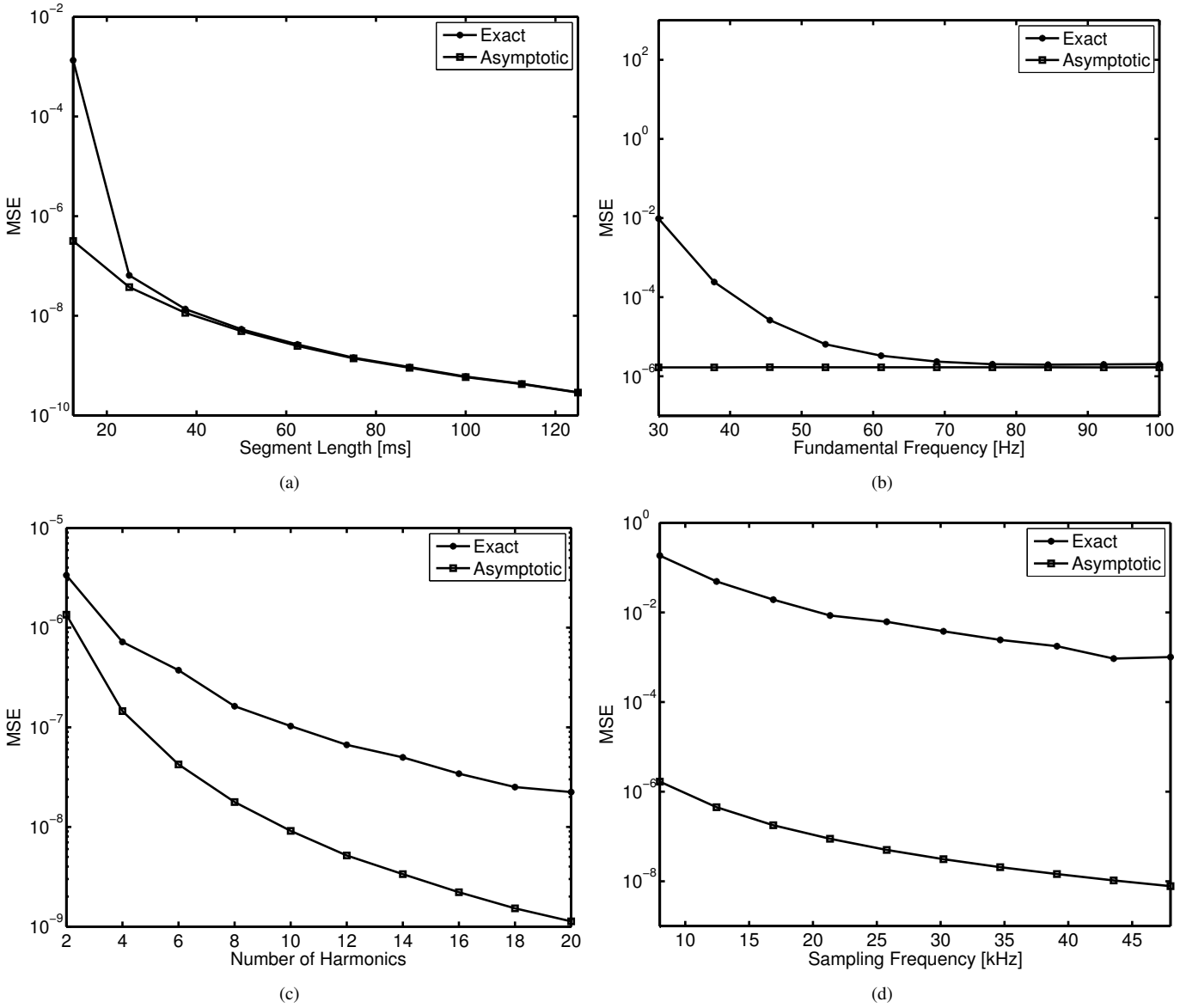


Fig. 1. Exact and asymptotic Cramér-Rao lower bounds as functions of various parameters, namely (a) the segment length (in ms), (b) the fundamental frequency (in Hz), (c) the number of harmonics L, and (d) the sampling frequency (in kHz). Each point on the curves is obtained over 1000 realizations of the involved parameters.

the same asymptotic approximation that the asymptotic CRLB is based on. Hence, if the asymptotic approximation is accurate for the CRLB, it is also likely to be accurate for the various estimators. Moreover, we can also learn something about the conditions under which the approximation will hold and learn if anything can be done about it. To make it easier to interpret the results, we will do this assuming typical physical values encountered in speech and audio applications. In the first experiment, a low fundamental frequency of 50 Hz is assumed along with a sampling frequency of 8 kHz. Moreover, the noise variance is kept fixed at one throughout these experiments. The remaining parameters were uniformly distributed phases and Rayleigh distributed amplitudes with five harmonics. Based on these values, the exact CRLB based on (9) and the asymptotic approximation in (10) were computed as a function of the segment length (in ms) for 1000 realizations of the parameters for each experimental condition. The results, in the form of the averages over these realizations, are shown in Figure 1(a). As can be seen, there is a huge discrepancy between the two bounds for short segments, and this discrepancy vanishes for long segments. This clearly shows that the claim that the problem of estimating low fundamental frequencies is difficult is indeed true. It also shows that it is entirely unrealistic to expect estimators to perform close to the asymptotic CRLB under these circumstances, and, hence, an estimator may be falsely deemed suboptimal if its performance is compared to the wrong bound.

In the next experiment, the segment length is kept fixed at 20 ms while the fundamental frequency is varied with the remaining parameters and experimental conditions being as before. The results are shown in Figure 1(b). The same observations as for the varying segment length can be observed

here, namely that as the fundamental frequency is lowered, relative to N, the discrepancy between the asymptotic and exact CRLBs grow. Beyond a certain frequency, here 80 Hz, there is basically no difference between the two bounds and asymptotic approximations must therefore be valid from this frequency and beyond. It should be noted that depending on the physics of the observed phenomenon, a low fundamental frequency may also have more harmonics, as they can in principle extend up to half the sampling frequency. This is not reflected in this experiment. It can be seen from (10) that, in theory, the more harmonics that are present, the more accurately the underlying fundamental frequency also can be estimated, at least for a sufficiently high N. For this reason, the next experiment focuses on the dependency on the number of harmonics, L. In this experiment, a fundamental frequency of 50 Hz is used for different L while other experimental settings were as before. The results can be seen in Figure 1(c). From the figure, it can be seen that the discrepancy between the two bounds actually increases as a function of L, meaning that the more harmonics are in the signal, it becomes relatively more difficult to determine the fundamental frequency, due to it being so low. On the other hand, the bound does decrease as a function of L even if the gap increases, so it is still beneficial to incorporate the additional harmonics in the model. Part of the reason that the bounds decrease as a function of L is that it effectively leads to an increase in SNR, as defined in (11) when the noise variance is kept fixed.

The final experiment involving the differences between the CRLBs is one where all the prior parameters are kept fixed while the sampling frequency is changed, and this is motivated as follows: since the highest possible segment length (in ms) is dictated by the stationarity of the observed signal, it is not possible to mitigate the problems associated with low fundamental frequencies by simply changing the segment length beyond a certain point. However, the sampling frequency can of course be changed in many situations, and raising the sampling frequency while keeping the segment length in ms fixed, of course leads to a higher number of samples N. Here, the behavior of the asymptotic and exact CRLBs is observed for a 20 ms segment and a fundamental frequency of 50 Hz with five harmonics. In Figure 1(d), the resulting curves can be seen. The figure shows that simply changing the sampling frequency does not alleviate the discrepancy between the two CRLBs, and the explanation is that while raising the fundamental does lead to a higher N, it also leads to a lower ω_0. But it is also interesting to note that both bounds do decrease as a function of the sampling rate, meaning that we are able to estimate the fundamental frequency more accurately by increasing the sampling frequency. An explanation for this is that while increasing the sampling frequency results in a proportionally higher N and lower ω_0 the effect of the noise on the ability to estimate the parameters is nonlinear. That this is the case can be seen from (10), from which it can be observed that the bound is inversely proportional to N^3.

B. Tested Methods

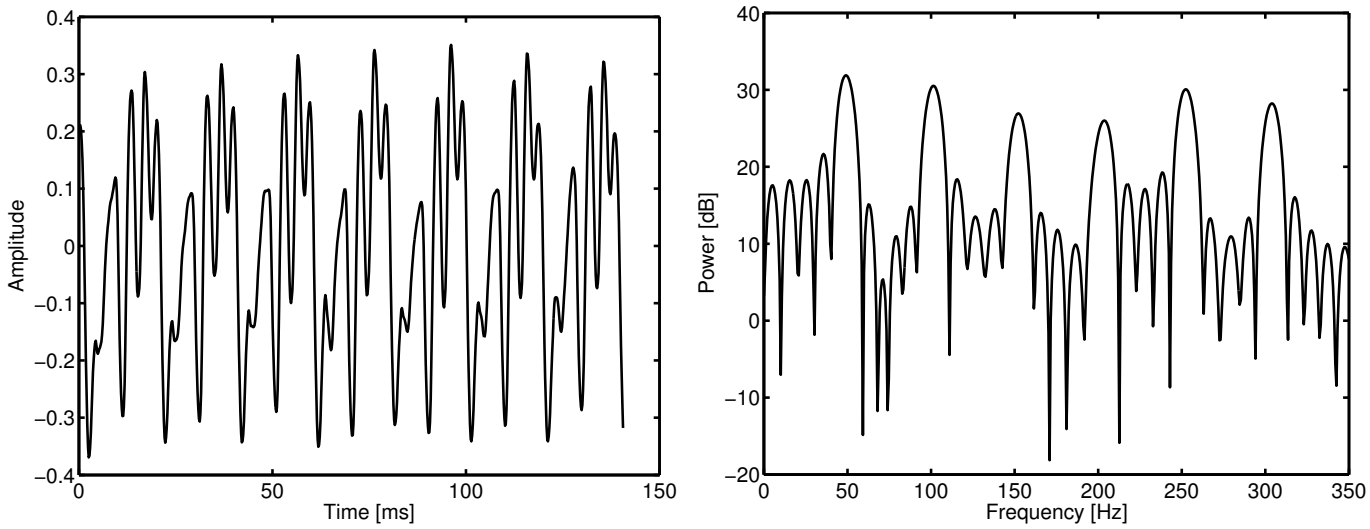
In the following experiments, we will compare the performance of the presented estimators to the previously published methods based on a complex signal model and/or asymptotic approximations. We will denote the methods for real signals by prefix "r" and their complex counterparts by prefix "c". To summarize, the following methods will be compared:

- rWLS is the harmonic fitting method based on WLS as presented in Section III-B. It requires that unconstrained frequencies and their amplitudes are found. This is done using ESPRIT and LS, respectively.
- rFILT is the optimal filtering method presented in Section III-C.
- rNLS is the NLS method of Section III-A.
- rABS is the subspace method based on measuring the angles between subspaces as described in Section III-D.
- rSHIFT is another subspace method, but based on the shift-invariance property, as presented in Section III-E.

We will compare the performance of these methods to a number of reference methods, namely the following:

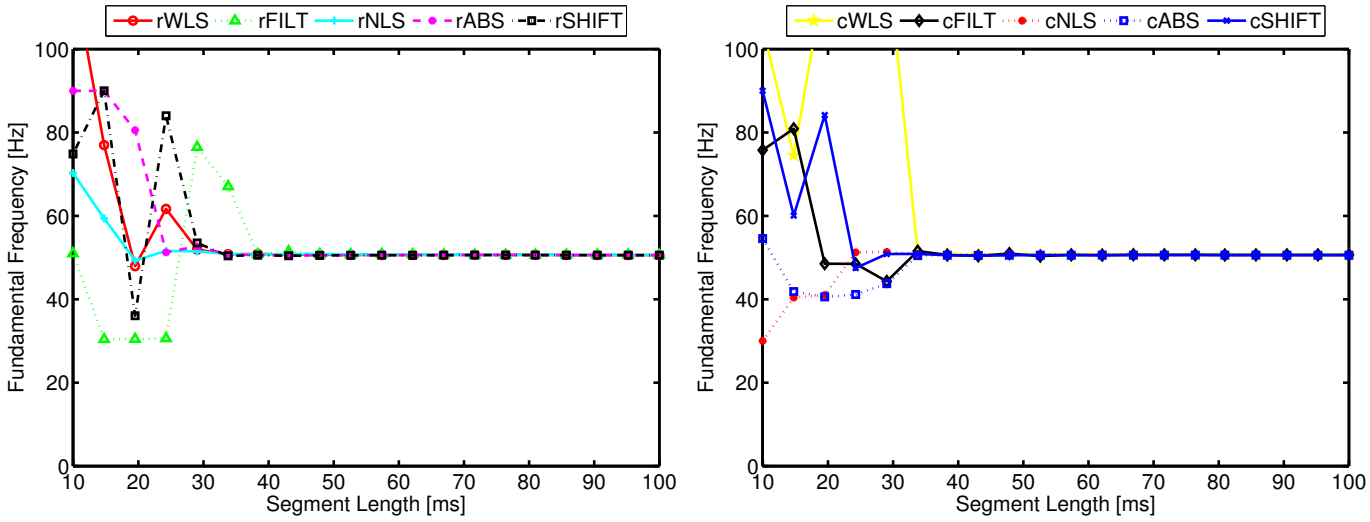
- cWLS is the harmonic fitting method as originally proposed in [6]. It uses asymptotic approximations of the weighting matrix to obtain a simple expression for the fundamental frequency. Like its real counterpart it requires unconstrained frequency and their amplitude estimates. Here, the same as for rWLS are used.
- cFILT is the optimal filtering method proposed in [8]. It differs from rFILT in that it does not take the existence of complex conjugate pairs of harmonics into account.
- cNLS is the approximate NLS method as described in [8]. It is similar to the methods of [4], [5]. It differs from rNLS in the following way: it is based on the asymptotic orthogonality of complex sinusoids and, hence, takes neither the existence of complex conjugate pairs nor the interaction between the harmonics into account.
- cABS is the MUSIC-based method of [18], except that the model order is assumed known. Unlike rABS, it uses an approximation of the angles between the subspaces.
- cSHIFT is the method proposed in [19], which is based on the shift-invariance property of the signal subspace. It differs from rSHIFT in that it does not take the existence of complex conjugate pairs of complex sinusoids into account. Unlike [19] it uses TLS rather than LS.

All estimators are implemented in a two-step fashion where a coarse fundamental frequency estimate is first found using a grid search after which a simple dichotomous search is used to obtain a refined estimate. The same grid size and dichotomous search algorithm is used for all the methods. For most of the methods, a covariance matrix size/filter length of $M = N/2$ is used, except for the optimal filtering methods where $M = N/4$ have been used (the reason for this will become clear later). For the estimators relying on a complex model, the real signal is mapped to a complex one via the Hilbert transform. The optimal filtering methods require an invertible covariance matrix for which reason the down-sampled analytic signal is used for cFILT. To address the numerical issue associated with very low fundamental frequencies, which may cause the involved matrices to be rank deficient numerically but not on paper, the Moore-Penrose pseudo-inverse [40] is used whenever appropriate.



(a) (b)

Fig. 2. Example of a signal having a low frequency, here a tone played by a contrabassoon. Shown are (a) the time-domain signal, and (b) part of its spectrum, namely the low frequencies, estimated using the periodogram.



(a) (b)

Fig. 3. Fundamental frequency estimates obtained for the signal in Figure 2 as a function of the segment length (in ms) for (a) the estimators for real-valued signals and (b) their complex counterparts.

C. A Signal Example

Next, we will illustrate the problems associated with low fundamental frequencies using a recorded signal, namely a tone played by a contrabassoon. The signal is shown in Figure 2(a) along with its spectrum in Figure 2(b), here estimated using the periodogram computed using a 8192 point FFT and a rectangular window. Note that a sampling frequency of 8820 Hz is used. In studying the effect of the low fundamental frequency on the ability to obtain accurate estimates, the segment length will be varied from 10 ms to 100 ms (with all segments beginning at the start of the signal shown in Figure 2(a)). The various estimators are then run on these segments. The number of harmonics was determined by visual inspection of the spectrum. The results are shown in Figure 3 for (a) the presented estimators, and (b) the estimators based on asymptotic approximations and complex signal models. A number of interesting observations can be made from the figures. Firstly, all estimators, both the real ones and their complex counterparts, converge to the same result when the segment length is increased. It can also be seen that all the methods break down when the segment length gets extremely short. Moreover, for this particular example, the methods for real signals generally outperform the complex ones, but it should also be noted that other factors may play a role due to the complex nature of real-life signals.

D. Monte Carlo Simulations

The methods are compared using Monte Carlo simulations by generating signals according to the model in (2) and then applying the various estimators to the resulting signal. The

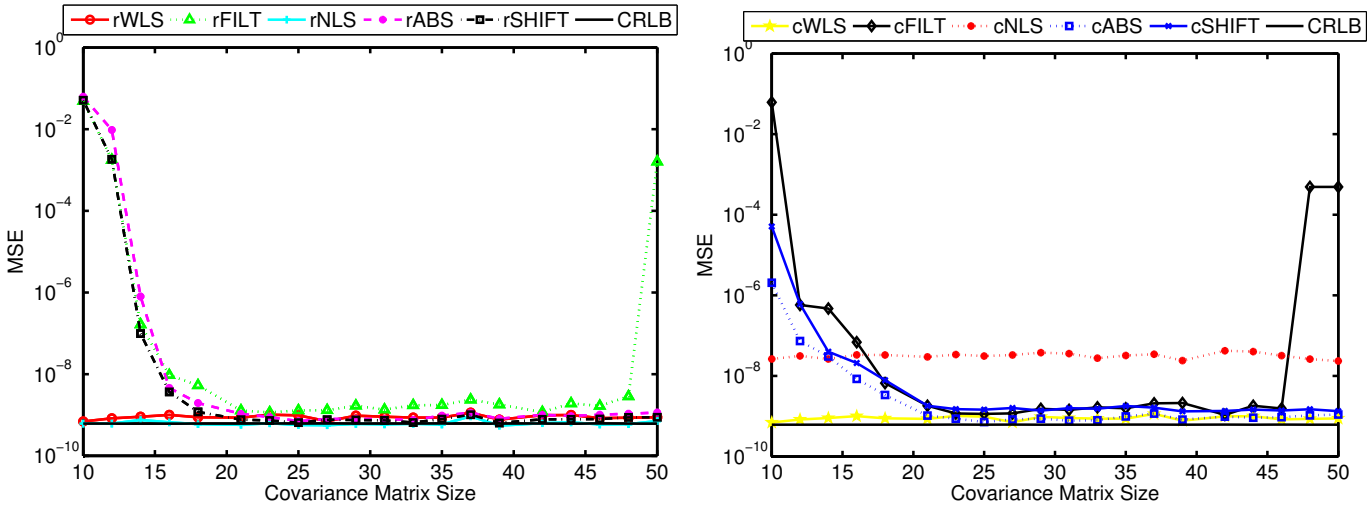


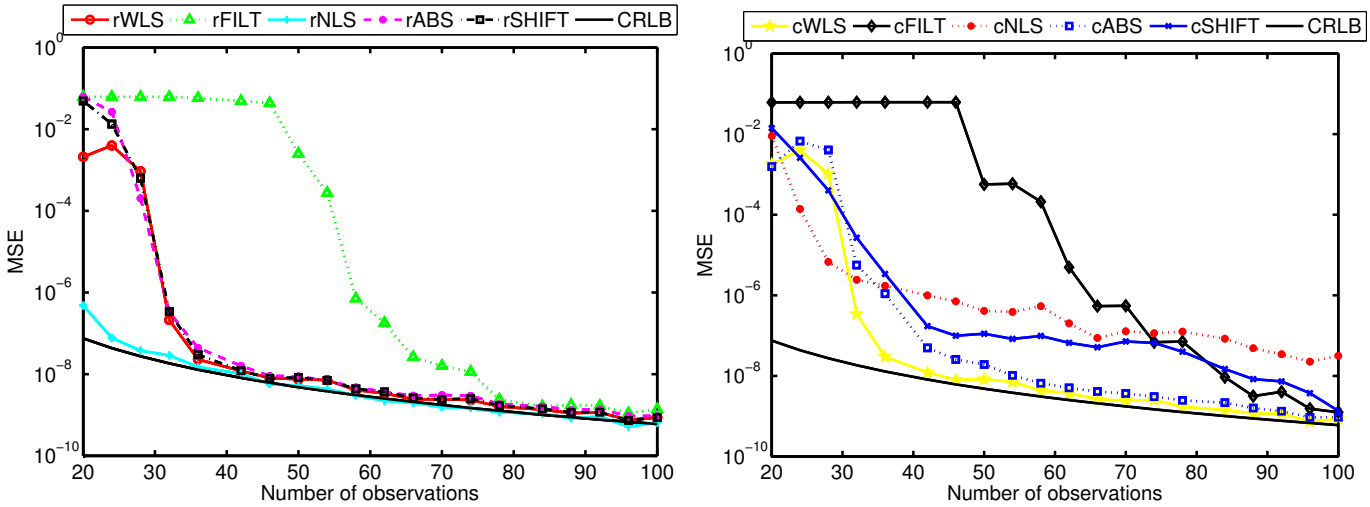
Fig. 4. Performance measured in terms of the Mean Square estimation Error (MSE) as a function of the covariance matrix size, M, for (a) the real estimators and (b) their complex counterparts based on asymptotic approximations.

so-obtained parameter estimates are then compared to the true parameters and the estimation error is measured in terms of the mean square error (MSE). For each set of experimental conditions, 100 realizations are used and the CRLB shown in the figures to follow is the average over the exact CRLB. The signals were generated with the following parameters, except when otherwise stated (e.g., when a certain parameter is varied): a fundamental frequency with $\omega_0 = 0.3129$ is used with five harmonics, each having unit amplitude and phases uniformly distributed between $-\pi$ and π. Segments of $N = 100$ samples were used with white Gaussian noise added at an SNR of 40 dB, according to the definition of the SNR in (11).

First, the influence of the covariance matrix size, which is also the filter length for the filtering methods, on the performance of the various estimators is investigated. This is done by simply varying M while keeping all other parameters fixed. The results are shown in Figure 4 for the real estimators (a) and the complex ones (b). Note that neither the NLS nor the WLS class methods make use of the covariance matrix an their performance hence does not depend on M. It can generally be observed that as long as the covariance matrix size is not chosen too low or too high, the methods perform well. In fact, the only class of methods that are sensitive to M being close to $M/2$ appears to be the optimal filtering methods (we remind the reader that $N = 100$ is used here). All methods, except one, perform close to the CRLB. For the cNLS method, a gap between its MSE and the CRLB can be seen. This demonstrates the clear sub-optimality of this method for the problem at hand and illustrates the importance of avoiding asymptotic approximations. It should be noted that the cNLS method performs extremely well for sufficiently high N and ω_0, being statistically efficient. Moreover, it has also been confirmed experimentally that the poor performance reported (and in the experiments to follow) here is not due to the suboptimality of the Hilbert transform used but rather, as stated, the asymptotic approximation.

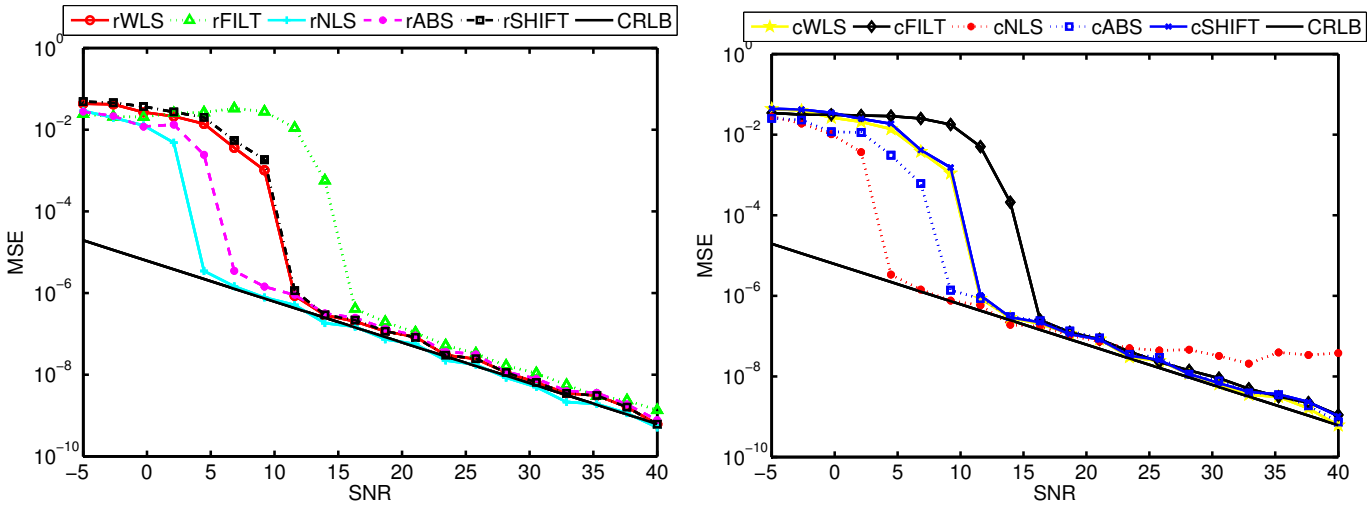
We will now proceed to investigate the dependency of the performance for the various estimators on the number of samples N. For the methods requiring a covariance matrix, it was stated that M should be chosen proportionally to N; otherwise, the estimator would not be consistent. So, in varying N, the covariance matrix size will also be varied with $M = N/2$ for all methods, except the optimal filtering methods for which $M = N/4$ is used. The results are shown in Figure 5(a) and Figure 5(b) for the two classes of methods. It can be seen that all the methods appear to be consistent in that the MSE decreases as a function of N. It can also be seen that the filtering methods, rFILT and cFILT perform poorly for low N, and that cNLS is clearly sup-optimal performing far from the CRLB, unlike rNLS, for the entire range of N shown here. Similarly, the cSHIFT methods perform poorly. Other than that, it appears that the remaining methods, aside from rNLS, break down below 40 samples.

In the next experiment, the performance of the various methods is investigated as a function of the SNR. From the asymptotic SNR in (10), one would perhaps expect this to be a trivial experiment as the noise variance is a linear parameter. However, due to the estimation problem being nonlinear, it is difficult to predict exactly how the performance of estimators will depend on the SNR. Moreover, it is well-known that, for nonlinear problems, estimators will exhibit so-called threshold behavior, which means that below a certain point, the estimators will break down producing what is essentially useless results. The MSE as a function of the SNR is depicted in Figures 6(a) and 6(b) for the real and complex estimators, respectively. A number of interesting observations can be made from these figures. For most of the methods, except cNLS, it can be seen that the performance increases as a function of the SNR, as can be expected from good estimators. The cNLS method can be seen to hit a floor for high SNRs. This is likely to be due to the approximations used in that method being inaccurate. For low, SNRs, however, this appear to not matter much as the error is dominated by the noise,



(a) (b)

Fig. 5. Performance measured in terms of the Mean Square estimation Error (MSE) as a function of the number of observations, N, for (a) the real estimators and (b) their complex counterparts based on asymptotic approximations.

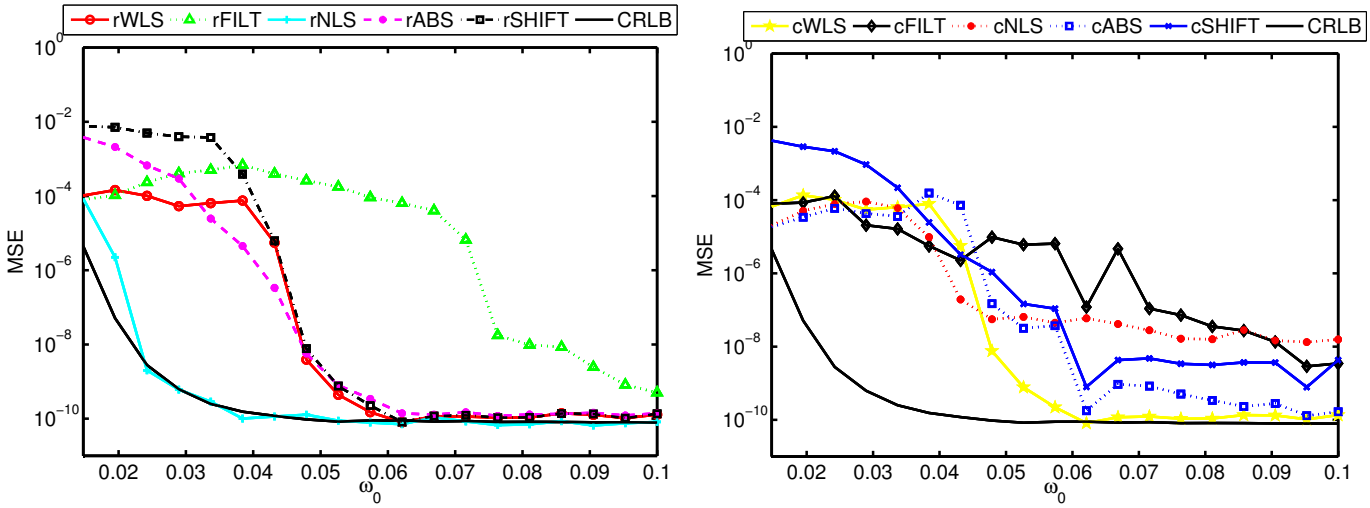


(a) (b)

Fig. 6. Performance measured in terms of the Mean Square estimation Error (MSE) as a function of the SNR for (a) the real estimators and (b) their complex counterparts based on asymptotic approximations.

with the MSE following the CRLB. It even appears that the cNLS method breaks down later than the cWLS, cSHIFT and cFILT methods with also the cABS method performing quite well for low SNRs. The rNLS can be observed to mitigate the problems of the cNLS as it follows the CRLB even for high SNRs. In fact, it can be seen to be statistically efficient above SNRs of 5 dB. Curiously, the rABS and cABS appear to perform almost equally well, being fairly robust against low SNRs, although it is not statistically efficient. The rWLS, rFILT and rSHIFT methods appear to perform similarly to their complex counterparts in this experiment, with the optimal filtering method performing the worst.

In the final and most important experiment, the role of the fundamental frequency will be investigated. More specifically, the fundamental frequency is varied from a value for which it is expected that all methods work to a low value close to zero, and it is expected they eventually will exhibit threshold behavior. The results are shown in Figures 7(a) and 7(b) for the two classes of methods. Starting with the complex methods, a number of interesting points can be made. Firstly, all except the cWLS perform poorly with the resulting MSEs differing substantially from the CRLB. The cWLS method performs well, following the CRLB, until about a fundamental frequency of 0.06. The cABS method also performs quite well, but performs further from the CRLB as the fundamental frequency is lowered. The cNLS, cFILT and cSHIFT methods can be seen to generally not perform well at all. For the real methods, it can be observed that the rNLS method performs the best, followed by the rWLS, rABS, and rSHIFT methods with the rFILT method performing quite poorly and worst of the methods. Comparing the two figures an important observation can be made: it can clearly be seen that all methods, except the



(a) (b)

Fig. 7. Performance measured in terms of the Mean Square estimation Error (MSE) as a function of the fundamental frequency, ω_0, for (a) the real estimators and (b) their complex counterparts based on asymptotic approximations.

rWLS method, are improved by the modifications presented in this paper. This clearly demonstrates that the commonly used approximations are not suitable for low fundamental frequencies and that it is possible to avoid them. Regarding the rWLS method, from the experiments, it appears that the approximations used in the weighting matrix in the cWLS method is not the reason for threshold behavior as the rWLS method behaves in the same way, rather the dominant error source is most likely the unconstrained frequencies. The reader should be aware that the rWLS method, like the cWLS method, is dependent on the unconstrained frequencies being accurate, and it can of course be expected that this will not be the case when the fundamental frequency is low. Note that the high sensitivity of this method to spurious frequency estimates was also demonstrated in [18], albeit under different circumstances.

V. Conclusion

In this paper, the problem of estimating low fundamental frequencies from real-valued measurements has been considered. The problem has been analyzed via comparisons of the asymptotic and approximate Cramér-Rao lower bounds. These comparisons show that the asymptotic approximations frequently used in estimators and in the computation of estimation bounds are not accurate under these circumstances. To mitigate this, a number of estimators have been presented in which such approximations are avoided, and these estimators can therefore be said to be exact. The estimators are based on the methodologies of maximum likelihood, leading to a nonlinear least-squares method and a harmonic fitting algorithm that fits individual frequencies to a fundamental frequency estimate, optimal filtering as known from Capon's classical beam-former, and subspace methods, herein one based on subspace orthogonality and one based on subspace shift-invariance. All of the methods, except the harmonic fitting one, which makes use of an set of intermediate parameters, have cubic complexity in the number of samples and/or the number of harmonics. In Monte Carlo simulations, the performance of the various estimators has been investigated and compared to methods employing asymptotic approximations. These simulations showed that, among the considered methods, the nonlinear least-squares method performed the best, the optimal filtering method performed the worst, and the remaining methods in-between. More importantly, however, the simulations showed that for all the considered methods, except the harmonic fitting one, it is possible to achieve improved performance by using the exact estimators. Moreover, it can be seen that not only do the proposed methods perform closer to the Cramér-Rao lower bound, but their threshold behavior is also improved for low fundamental frequencies.

References

[1] E. Conte, A. Filippi, and S. Tomasin, "ML period estimation with application to vital sign monitoring," *IEEE Signal Process. Lett.*, vol. 17, no. 11, pp. 905–908, 2010.

[2] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34(5), pp. 1124–1138, Oct. 1986.

[3] G. Ogden, L. Zurk, M. Siderius, E. Sorensen, J. Meyers ad S. Matzner, and M. Jones, "Frequency domain trackin of passive vessel harmonics," *J. Acoust. Soc. Am.*, vol. 126, pp. 2249, 2009.

[4] M. Noll, "Pitch determination of human speech by harmonic product spectrum, the harmonic sum, and a maximum likelihood estimate," in *Proc. Symposium on Computer Processing Communications*, 1969, pp. 779–797.

[5] B. G. Quinn and P. J. Thomson, "Estimating the frequency of a periodic function," *Biometrika*, vol. 78(1), pp. 65–74, 1991.

[6] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Signal Processing*, vol. 80, pp. 1937–1944, 2000.

[7] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 12(1), pp. 76–87, 2004.

[8] M. G. Christensen, P. Stoica, A. Jakobsson and S. H. Jensen, "Multi-pitch estimation," *Signal Processing*, vol. 88(4), pp. 972–983, Apr. 2008.

[9] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 22, no. 5, pp. 353–362, Oct. 1974.

[10] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, vol. 41(2), pp. 293–309, 1967.

[11] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Trans. Signal Process.*, vol. 39, no. 1, pp. 40–48, Jan. 1991.

[12] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111(4), pp. 1917–1930, Apr. 2002.

[13] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., chapter 5, pp. 495–518. Elsevier Science B.V., 1995.

[14] K. W. Chan and H. C. So, "Accurate frequency estimation for real harmonic sinusoids," *IEEE Signal Process. Lett.*, vol. 11(7), pp. 609–612, July 2004.

[15] J. Moorer, "The optimum comb method of pitch period analysis of continuous digitized speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 22, no. 5, pp. 330–338, Oct 1974.

[16] M. G. Christensen and A. Jakobsson, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Trans. Signal Processing*, vol. 58(12), pp. 5969–5983, Dec. 2010.

[17] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 27–30 April 1993, vol. 2, pp. 728–731.

[18] M. G. Christensen, A. Jakobsson and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15(5), pp. 1635–1644, July 2007.

[19] M. G. Christensen, A. Jakobsson and S. H. Jensen, "Fundamental frequency estimation using the shift-invariance property," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2007, pp. 631–635.

[20] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, vol. 5 of *Synthesis Lectures on Speech & Audio Processing*, Morgan & Claypool Publishers, 2009.

[21] E. J. Hannan and B. G. Quinn, "The resolution of closely adjacent spectral lines," *J. of Time Series Analysis*, vol. 10, pp. 13–31, 1989.

[22] D. Huang, "On low and high frequency estimation," *J. of Time Series Analysis*, vol. 17(4), pp. 351–365, 1996.

[23] P. Stoica and A. Eriksson, "MUSIC estimation of real-valued sine-wave frequencies," *Signal Processing*, vol. 42, pp. 139–146, 1995.

[24] K. Mahata, "Subspace fitting approaches for frequency estimation using real-valued data," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3099–3110, Aug. 2005.

[25] A. Jakobsson, T. Ekman, and P. Stoica, "Capon and APES spectrum estimation for real-valued signals," in *Eighth IEEE Digital Signal Processing Workshop*, 1998.

[26] H.C. So, K. W. Chan, Y.T. Chan, and K.C. Ho, "Linear prediction approach for efficient frequency estimation of multiple real sinusoids: algorithms and analyses," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2290–2305, July 2005.

[27] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34(3), pp. 276–280, Mar. 1986.

[28] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Sinusoidal order estimation using angles between subspaces," *EURASIP J. on Advances in Signal Processing*, vol. 2009, pp. 1–11, 2009, Article ID 948756.

[29] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: Survey, new results and an application," *IEEE Trans. Signal Process.*, vol. 48(2), pp. 338–352, Feb. 2000.

[30] S. Godsill and M. Davy, "Bayesian computational models for inharmonicity in musical instruments," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 2005, pp. 283–286.

[31] B. G. Quinn and E. J. Hannan, *The Estimation and Tracking of Frequency*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2001.

[32] G. L. Bretthorst, "An introduction to parameter estimation using Bayesian probability theory," in *Max. Entropy and Bayesian Methods*, P. Fougere, Ed., pp. 53–79. 1990.

[33] G. Bienvenu and L. Kopp, "Optimality of high resolution array processing using the eigensystem approach," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31(5), pp. 1235–1248, Oct. 1983.

[34] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, 1993.

[35] P. Stoica, A. Jakobsson, and J. Li, "Cisoid Parameter Estimation in the Colored Noise Case: Asymptotic Cramér-Rao Bound, Maximum Likelihood and Nonlinear Least-Squares," *IEEE Trans. Signal Process.*, vol. 45, pp. 2048–2059, August 1997.

[36] P. Stoica and T. Söderström, "On reparameterization of loss functions used in estimation and the invariance principle," *Elsevier Signal Processing*, vol. 17, pp. 383–387, 1989.

[37] A.L. Swindlehurst and P. Stoica, "Maximum likelihood methods in radar array signal processing," *Proc. IEEE*, vol. 86, no. 2, pp. 421–441, 1998.

[38] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21(4), pp. 36–47, July 2004.

[39] G. Bienvenu, "Influence of the spatial coherence of the background noise on high resolution passive methods," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1979, pp. 306–309.

[40] G. H. Golub and C. F. V. Loan, *Matrix Computations*, The Johns Hopkins University Press, 3rd edition, 1996.

[41] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Pearson Prentice Hall, 2005.

Mads Græsbøll Christensen (S'00–M'05–SM'11)) was born in Copenhagen, Denmark, in March 1977. He received the M.Sc. and Ph.D. degrees in 2002 and 2005, respectively, from Aalborg University (AAU) in Denmark, where he is also currently employed at the Dept. of Architecture, Design & Media Technology as Associate Professor. At AAU, he is head of the Audio Analysis Lab which conducts research in audio signal processing.

He was formerly with the Dept. of Electronic Systems, Aalborg University and has been a Visiting Researcher at Philips Research Labs, ENST, UCSB, and Columbia University. He has published more than 100 papers in peer-reviewed conference proceedings and journals as well as 1 research monograph. His research interests include digital signal processing theory and methods with application to speech and audio, in particular parametric analysis, modeling, enhancement, separation, and coding.

Dr. Christensen has received several awards, including an ICASSP Student Paper Award, the Spar Nord Foundation's Research Prize for his Ph.D. thesis, a Danish Independent Research Council Young Researcher's Award, and the Statoil Prize 2013, as well as prestigious grants from the Danish Independent Research Council and the Villum Foundation's Young Investigator Programme. He has served as an Associate Editor for the IEEE Signal Processing Letters.

Paper D

Instantaneous Pitch Estimation with Optimal Segmentation for Non-Stationary Voiced Speech

S. M. Nørholm, J. R. Jensen, and **M. G. Christensen**

The paper has been published in the
IEEE/ACM Transactions on Audio, Speech, Language Processing, vol. 24, no. 12, pp. 2354—2367, December 2016.

Instantaneous Fundamental Frequency Estimation with Optimal Segmentation for Non-Stationary Voiced Speech

Sidsel Marie Nørholm, Jesper Rindom Jensen, *Member, IEEE,*
and Mads Græsbøll Christensen, *Senior Member, IEEE*

***Abstract*—In speech processing, the speech is often considered stationary within segments of 20–30 ms even though it is well known not to be true. In this paper, we take the non-stationarity of voiced speech into account by using a linear chirp model to describe the speech signal. We propose a maximum likelihood estimator of the fundamental frequency and chirp rate of this model, and show that it reaches the Cramer-Rao lower bound. Since the speech varies over time, a fixed segment length is not optimal, and we propose making a segmentation of the signal based on the maximum a posteriori (MAP) criterion. Using this segmentation method, the segments are on average longer for the chirp model compared to the traditional harmonic model. For the signal under test, the average segment length is 24.4 ms and 17.1 ms for the chirp model and traditional harmonic model, respectively. This suggests a better fit of the chirp model than the harmonic model to the speech signal. The methods are based on an assumption of white Gaussian noise, and, therefore, two prewhitening filters are also proposed.**

***Index Terms*—Harmonic chirp model, parameter estimation, segmentation, prewhitening.**

I. Introduction

PARAMETER estimation of harmonic signals is relevant to the fields of speech processing and communication. In speech models, the speech signal is often split into a voiced part and an unvoiced part. The voiced part of the speech signal is produced by the vibration of the vocal cords, and, therefore, has a structure with a fundamental frequency and a set of overtones given by integer multiples of the fundamental. Over the years, several fundamental frequency estimators have been proposed based on different methods, such as autocorrelation [2], statistical [3]–[5], optimal filtering [6], or eigenvalue decomposition [7], [8]. Some methods work directly in the time domain [8], [9] whereas others use the spectrum or cepstrum [10], [11]. Comparisons of various fundamental frequency estimators have shown that different domains offer different advantages in e.g., the two genders [12]. Most of these fundamental frequency estimators split the signal into segments of 20–30 ms [13], make a voiced/unvoiced decision [14], [15], and estimate the parameters of each voiced segment separately. In most models, the signal is assumed stationary within each segment, even though it is well known that this assumption of stationarity does not hold [13], [16]. Some estimators overcome this problem of non-stationarity by looking at shorter segments, as, e.g., in [17], [18] where the fundamental frequency is estimated based on a single period of voiced speech. This overcomes the problem of non-stationarity, however, the lack of data points, that each estimate is based on, gives a greater uncertainty of the estimates. This is also seen in [18] where the method has a poor performance with respect to fine pitch error (FPE). Another approach, giving higher estimation accuracy, is to model the change in fundamental frequency within each segment. This can be done by extending the harmonic model [19]–[22] to a harmonic chirp model, which has also been suggested in [23]–[25]. Here, the harmonic structure remains the foundation of the model, but the fundamental frequency is allowed to change linearly within each segment. This introduces an extra parameter to estimate, but with the benefit that the model fits the speech signal better. Using the harmonic chirp model instead of the traditional harmonic model can, therefore, lead to better speech enhancement [26], but with a better fit of the model it is also possible to work with longer segments. In general, longer segments lead to better performance of the estimators, and so a smaller error on the estimated parameters can be obtained. However, the optimal segment length depends on the features of the signal, which are varying over time in the case of speech signals. At some time instances, the parameters are almost constant, and, in such periods, long segments can be used whereas at other points in time, the parameters will change fast and shorter segments should be used. Instead of using a fixed segment length, it is, therefore, better to have a varying segment length that depends on the signal characteristics at the given point in time. In [27], [28], the signal is modelled based on linear prediction (LP), and the segment length is chosen according to a trade-off between bit rate and distortion. The principle can, however, be used with other criteria for choosing the segment length, depending on what is most relevant in the given situation. The noise characteristics also have an impact on the performance of parameter estimators and optimal segmentation. Most methods make an assumption of white Gaussian noise, which is rarely experienced in real life scenarios. One way to address this problem is to preprocess the signal in a way that makes the noise resemble white Gaussian noise, as is, e.g., done through Cholesky factorisation [29].

The contribution in this paper is three-fold. First, we pro-

This work was funded by the Villum Foundation and the Danish Council for Independent Research, grant ID: DFF 1337-00084. Part of this material was published at Asilomar 2014 [1].

S. M. Nørholm, J. R. Jensen and M. G. Christensen are with the Audio Analysis Lab, AD:MT, Aalborg University, DK-9000 Aalborg, Denmark, e-mail: {smn, jrj, mgc}@create.aau.dk

pose estimating the fundamental frequency and fundamental chirp rate by maximising the likelihood. Since maximising with respect to two parameters leads to a search in a two-dimensional space, we suggest an iterative procedure where first a one dimensional optimisation of the chirp parameter is performed followed by a one dimensional optimisation of the fundamental frequency based on the newly found estimate of the chirp rate. The estimation process is ended by convergence of the two-dimensional cost function. The proposed parameter estimator is a continuation of [1]. Our iterative procedure offers some benefits over the method suggested in [25], where an approximate cost function is introduced in order to decrease the computational load. The approximate cost function in [25] is evaluated over a two-dimensional grid, which means that fundamental frequency and chirp rate have to be found for each point in the grid before the optimum is found. In this paper, the original cost function is evaluated iteratively, giving fewer points for evaluation thus making the procedure suggested in this paper faster. Second, we suggest a maximum a posteriori (MAP) criterion to either make model selection between the traditional harmonic model and the harmonic chirp model, or make optimal segmentation of the signal based on one of the models. The optimal segmentation is based on the principle suggested in [27], [28]. The principle is adapted to the harmonic chirp model by using the maximum a posteriori (MAP) criterion for choosing the segment length. The model selection and optimal segmentation are introduced to give better representations of the signal. With the model selection, the more complex harmonic chirp model is favoured over the traditional harmonic model whenever it is beneficial according to the MAP principle. This reduces the error in, e.g., reconstruction or filtering [26] of the signal while keeping complexity low by choosing the traditional harmonic model whenever this is sufficient. With optimal segmentation, the segment length differs over time, optimising the fit of the model to the signal in each segment. This results in parameters that better describe the signal in the segment, and so also a lower error on, e.g., reconstruction or filtering. Third, we suggest two different methods to prewhiten the noise. Both the maximum likelihood estimator of the fundamental frequency and chirp rate and the MAP criterion are based on an assumption of white Gaussian noise, and, therefore, a prewhitening step is necessary if the noise is not white Gaussian. Both methods are based on noise power spectral density (PSD) estimation [30]–[33] and generate a filter to counteract the spectral shape of the noise. The filter is either based directly on the estimated spectrum of the noise or linear prediction of the noise.

The paper is organised as follows. In Section II, the harmonic chirp model is introduced. In Section III, the maximum likelihood estimator of the fundamental frequency and fundamental chirp rate is derived. In Section IV, the general MAP criterion is introduced for the harmonic chirp model along with the MAP model selection criterion between the traditional harmonic model, the harmonic chirp model and the noise only model. This is followed by the segmentation principle based on the MAP criterion in Section V. In Section VI, the two prewhitening methods are described. In Section VII, the proposed methods are tested through simulations on synthetic chirp signals and speech, and the paper is concluded in Section VIII.

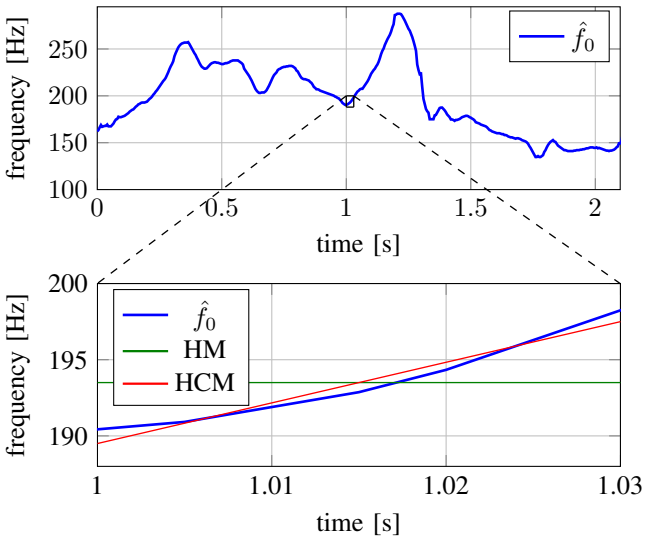


Fig. 1: Sketch of the difference between the harmonic model (HM) and the harmonic chirp model (HCM). The top plot shows a fundamental frequency track ($\hat{f}_0$) of a speech signal. The bottom plot is an enlargement of the small black square in the top plot.

II. Harmonic chirp model

In order to illustrate the difference between the harmonic model and the harmonic chirp model, a fundamental frequency track of a speech signal is plotted in the top of Fig. 1. The figure shows that the fundamental frequency changes continuously over time. This is also illustrated in the bottom figure with an enlargement of the 30 ms segment marked by the black square in the top figure. In this 30 ms segment, the fundamental frequency changes by approximately 8 Hz, whereas the harmonic model (HM), and most other fundamental frequency estimators, would assume the instantaneous fundamental frequency to be stationary within the segment. The harmonic chirp model (HCM) does not assume stationarity, but assumes a linear change of the fundamental frequency within a segment. As shown in the bottom figure, this model better describes the instantaneous fundamental frequency in the segment. With a better model, it is possible to work with longer segments, which will give higher accuracy on the estimated parameters. Further, it can lead to more efficient coding and signal reconstruction.

The harmonic chirp model is an extension of the traditional harmonic model. Therefore, the frequencies of the harmonics are still given by integer multiples of a fundamental frequency. However, in the chirp model, the instantaneous frequency of the l'th harmonic, $\omega_l(n)$, varies with the time index $n = n_0, ..., n_0 + N - 1$ in a linear way:

$$\omega_l(n) = l(\omega_0 + kn), \tag{1}$$

where $\omega_0 = 2\pi f_0/f_s$, with f_s the sampling frequency, is the normalised fundamental frequency, and k is the normalised

fundamental chirp rate. This means that in order to obtain the instantaneous frequency, both the fundamental frequency and the chirp rate are needed. The instantaneous phase, $\varphi_l(n)$, of the sinusoids are given by the integral of the instantaneous frequency as

$$\varphi_l(n) = l\left(\omega_0 n + \frac{1}{2}kn^2\right) + \phi_l, \tag{2}$$

where $\phi_l \in [0, 2\pi]$ is the initial phase of the l'th harmonic. This leads to the complex harmonic chirp model for a voiced speech signal, $s(n)$:

$$s(n) = \sum_{l=1}^{L} A_l e^{j\varphi_l(n)} \tag{3}$$

$$= \sum_{l=1}^{L} \alpha_l e^{jl\left(\omega_0 n + k/2n^2\right)}, \tag{4}$$

where L is the number of harmonics and $\alpha_l = A_l e^{j\phi_l}$, $A_l > 0$ is the complex amplitude of the l'th harmonic. For speech signals the model order has to be estimated, which can be done, e.g., by use of the MAP criterion introduced in Section IV (see also [8]). The complex signal model is used instead of the real because it can ease both notation and computation. A real signal can be easily converted to a complex signal by use of the Hilbert transform [34] and without loss of information, downsampled by a factor of two.

A special case of the harmonic chirp model for $k = 0$ is the traditional harmonic model:

$$s(n) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n}. \tag{5}$$

Defining a vector of samples

$$\mathbf{s} = [s(n_0)\ s(n_0+1)\ \ldots\ s(n_0+N-1)]^T, \tag{6}$$

where $(\cdot)^T$ denotes the transpose. Note that the dependency on the index n_0 is left out for ease of notation. The signal model is then written as

$$\mathbf{s} = \mathbf{Z}\mathbf{a}, \tag{7}$$

where $\mathbf{Z}$ is a matrix constructed from a set of L modified Fourier vectors matching the harmonics of the signal,

$$\mathbf{Z} = [\mathbf{z}(\omega_0, k)\ \ \mathbf{z}(2\omega_0, 2k)\ \ \ldots\ \ \mathbf{z}(L\omega_0, Lk)], \tag{8}$$

with

$$\mathbf{z}(l\omega_0, lk) = \begin{bmatrix} e^{jl(\omega_0 n_0 + k/2n_0^2)} \\ e^{jl(\omega_0(n_0+1)+k/2(n_0+1)^2)} \\ \vdots \\ e^{jl(\omega_0(n_0+N-1)+k/2(n_0+N-1)^2)} \end{bmatrix}. \tag{9}$$

The vector $\mathbf{a}$ contains the complex amplitudes of the harmonics, $\mathbf{a} = [\alpha_1\ \alpha_2\ \ldots\ \alpha_L]^T$.

The signal we want to make parameter estimation on, is often buried in noise, $v(n)$, to give the observed signal, $x(n)$,

$$x(n) = s(n) + v(n), \tag{10}$$

which can also be put into a vector of observed samples

$$\mathbf{x} = \mathbf{s} + \mathbf{v}, \tag{11}$$

where $\mathbf{x}$ and $\mathbf{v}$ are defined similarly to $\mathbf{s}$ in (6). For real signals as speech, the signal model will not fit the desired signal perfectly, and so $\mathbf{v}$ will also cover the part of the speech signal that does not align with the given model as, e.g., unvoiced speech during mixed excitations.

III. Estimation of frequency and chirp rate

The fundamental frequency and chirp rate are estimated by maximising the likelihood. The maximum likelihood estimates are the parameters of the model that describe the observed signal the best, i.e., the parameters that maximise the probability of the observed data, $\mathbf{x}$, given the parameters:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta}), \tag{12}$$

where $\boldsymbol{\theta}$ is a vector containing the parameters of the model. Under the assumption of circularly symmetric Gaussian noise, the likelihood function can be written as [8]:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{\pi^N \det(\mathbf{R}_v)} e^{-(\mathbf{x}-\mathbf{s})^H \mathbf{R}_v^{-1}(\mathbf{x}-\mathbf{s})} \tag{13}$$

$$= \frac{1}{\pi^N \det(\mathbf{R}_v)} e^{-\mathbf{v}^H \mathbf{R}_v^{-1}\mathbf{v}}, \tag{14}$$

where $\det(\cdot)$ denotes the determinant of the argument, $(\cdot)^H$ the Hermitian transpose and $\mathbf{R}_v = \mathbb{E}[\mathbf{v}\mathbf{v}^H]$ the noise covariance matrix, with $\mathbb{E}(\cdot)$ the mathematical expectation. Often the log likelihood is maximised instead of the likelihood

$$\ln \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = -N \ln \pi - \ln \det(\mathbf{R}_v) - \mathbf{v}^H \mathbf{R}_v^{-1}\mathbf{v}. \tag{15}$$

In the case of white noise, the noise covariance matrix reduces to a diagonal matrix, $\mathbf{R}_v = \sigma_v^2 \mathbf{I}_N$, where σ_v^2 is the variance of the noise signal and $\mathbf{I}_N$ is an $N \times N$ identity matrix. The log likelihood can, therefore, be reduced to

$$\ln \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = -N \ln \pi - N \ln \sigma_v^2 - \frac{1}{\sigma_v^2}||\mathbf{v}||_2^2. \tag{16}$$

The noise and its variance can be found using the signal model in (7)

$$\mathbf{v} = \mathbf{x} - \mathbf{s} = \mathbf{x} - \mathbf{Z}\mathbf{a} \Rightarrow \tag{17}$$

$$||\mathbf{v}||_2^2 = ||\mathbf{x} - \mathbf{Z}\mathbf{a}||_2^2, \tag{18}$$

$$\sigma_v^2 = \frac{1}{N}||\mathbf{x} - \mathbf{Z}\mathbf{a}||_2^2, \tag{19}$$

which turns the log likelihood into

$$\ln \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = -N \ln \pi - N \ln \frac{1}{N}||\mathbf{x} - \mathbf{Z}\mathbf{a}||_2^2 - N. \tag{20}$$

In the estimation of the fundamental frequency and chirp rate, it is only necessary to consider terms dependent on these two parameters, and the log likelihood function can be reduced to the nonlinear least squares (NLS) estimator that minimises the error between the observed signal and the signal model:

$$\{\widehat{\mathbf{a}}, \widehat{\omega}_0, \widehat{k}\} = \arg\min_{\mathbf{a},\omega_0,k} ||\mathbf{x} - \mathbf{s}||_2^2 \tag{21}$$

$$= \arg\min_{\mathbf{a},\omega_0,k} ||\mathbf{x} - \mathbf{Z}\mathbf{a}||_2^2. \tag{22}$$

Here, we are interested in the joint estimation of the fundamental frequency and chirp rate, and, therefore, the amplitudes are substituted with their least squares estimate [9],

$$\widehat{\mathbf{a}} = (\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{x}, \tag{23}$$

to give the estimator:

$$\{\widehat{\omega}_0, \widehat{k}\} = \arg\min_{\omega_0,k} ||\mathbf{x} - \mathbf{Z}(\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{x}||_2^2 \tag{24}$$

$$= \arg\min_{\omega_0,k} \left(\mathbf{x}^H(\mathbf{I}_N - \mathbf{Z}(\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H)\mathbf{x}\right) \tag{25}$$

$$= \arg\min_{\omega_0,k} \left(\mathbf{x}^H\Pi^{\perp}\left(\omega_0, k\right)\mathbf{x}\right), \tag{26}$$

where Π is an orthogonal projection matrix

$$\Pi\left(\omega_0, k\right) = \mathbf{Z}(\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H \tag{27}$$

and $\Pi^{\perp}$ its orthogonal complement

$$\Pi^{\perp}\left(\omega_0, k\right) = \mathbf{I}_N - \Pi\left(\omega_0, k\right). \tag{28}$$

This process includes a two-dimensional optimisation over ω_0 and k. To solve the problem in a computationally efficient manner, we propose iterating between two one-dimensional searches [1]. First, the chirp rate in step i, k^i, is estimated using the fundamental frequency estimate from the previous iteration, $\omega_0^{(i-1)}$, $i = 1, 2, ...$

$$k^{(i)} = \arg\min_{k} \left(\mathbf{x}^H\Pi^{\perp}(\omega_0^{(i-1)}, k)\mathbf{x}\right). \tag{29}$$

This estimate of the chirp rate is used to find a new estimate of the fundamental frequency

$$\omega_0^{(i)} = \arg\min_{\omega_0} \left(\mathbf{x}^H\Pi^{\perp}(\omega_0, k^{(i)})\mathbf{x}\right). \tag{30}$$

The estimates of ω_0 and k are found by iterating between (29) and (30) until convergence of the cost function in (26), but could alternatively be ended by the convergence of the estimated parameters. The fundamental frequency and chirp rate minimising the cost function in (26) are found by searching among candidates in a grid centred at the value of the parameter from the previous iteration, $i - 1$. The grid search is followed by a Dichotomous search [35] to get a refined estimate of the minimum. It is expected that the fundamental frequency estimate is close to the estimate found under the assumption of stationarity within the analysis frame. Therefore, a fundamental frequency estimate found under the traditional harmonic assumption, e.g., by using one of the methods in [8], will be a good choice as an initialisation of the iterations, i.e., $\omega_0^{(0)} = \omega_{0,h}$. The chirp rate is expected to be small and the first grid search is, therefore, centred around zero, i.e., $k^{(0)} = 0$. The estimation process is summarised in Table I.

The best obtainable performance of an unbiased estimator is given by the Cramer-Rao lower bound (CRLB). The CRLB sets a lower limit to the variance of the parameter estimate

$$\text{var}(\widehat{\theta}_g) \geq [\mathcal{I}(\boldsymbol{\theta})^{-1}]_{gg}, \tag{31}$$

where θ_g is the g'th parameter of the parameter vector $\boldsymbol{\theta}$ of length G, $[\cdot]_{gg}$ denotes the matrix element of row g and column g, and $\mathcal{I}(\boldsymbol{\theta})$ is the Fisher information matrix (FIM) [36] of size $G \times G$:

$$[\mathcal{I}(\boldsymbol{\theta})]_{gh} = -\mathbb{E}\left\{\frac{\partial^2 \ln(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial\theta_g\partial\theta_h}\right\}. \tag{32}$$

Under the assumptions of white Gaussian noise and a noise covariance matrix independent of the parameters, the FIM reduces to:

$$\mathcal{I}(\boldsymbol{\theta}) = \frac{2}{\sigma_v^2}\text{Re}\left\{\frac{\partial\mathbf{s}^H}{\partial\boldsymbol{\theta}}\frac{\partial\mathbf{s}}{\partial\boldsymbol{\theta}^T}\right\} \tag{33}$$

$$= \frac{2}{\sigma_v^2}\text{Re}\left\{\mathbf{D}^H(\boldsymbol{\theta})\mathbf{D}(\boldsymbol{\theta})\right\} \tag{34}$$

with

$$\mathbf{D}(\boldsymbol{\theta}) = [\mathbf{d}(\omega_0)\,\mathbf{d}(k)\,\mathbf{d}(A_1)\,\mathbf{d}(\phi_1)\,\ldots\,\mathbf{d}(A_L)\,\mathbf{d}(\phi_L)], \tag{35}$$

$$\mathbf{d}(y) = \frac{\partial\mathbf{s}}{\partial y}. \tag{36}$$

For the signal model at hand, the elements of the $\mathbf{d}$ vectors are:

$$[\mathbf{d}(\omega_0)]_n = \sum_{l=1}^{L} jlnA_l e^{jl(\omega_0 n + k/2n^2) + j\phi_l}, \tag{37}$$

$$[\mathbf{d}(k)]_n = \sum_{l=1}^{L} \frac{1}{2}jln^2A_l e^{jl(\omega_0 n + k/2n^2) + j\phi_l}, \tag{38}$$

$$[\mathbf{d}(A_l)]_n = e^{jl(\omega_0 n + k/2n^2) + j\phi_l}, \tag{39}$$

$$[\mathbf{d}(\phi_l)]_n = jA_l e^{jl(\omega_0 n + k/2n^2) + j\phi_l}. \tag{40}$$

The CRLB depends on the choice of n_0. The best estimates are obtained if the segment is centred around $n = 0$ [23], and, therefore, n_0 should be chosen as $n_0 = -(N-1)/2$ for N odd and $n_0 = -N/2$ for N even. The CRLB also depends on the number of harmonics and the amplitude of the l'th harmonic A_l. The CRLB for a harmonic signal [8] decreases with $A_l^2 l^2$, which means that the more harmonics included in the estimate of fundamental frequency and chirp rate, the better the estimate.

TABLE I: Estimation of fundamental frequency and chirp rate.

for each sample
initialisation
$\omega_0^{(0)} = \omega_{0,h}$
$k^{(0)} = 0$
$\Delta k = 2\alpha_k/(K-1)$
$\Delta\omega = 2\alpha_\omega/(K-1)$
repeat
$K = \{k^{(i-1)} - \alpha_k, \Delta k, ..., k^{i-1} + \alpha_k\}$
$\Omega = \{\omega_0^{(i-1)} - \alpha_\omega, \Delta\omega, ..., \omega_0^{i-1} + \alpha_\omega\}$
$k^{(i)} = \arg\min_{k\in K}\left(\mathbf{x}^H\Pi^{\perp}(\omega_0^{(i-1)}, k)\mathbf{x}\right)$
$\omega_0^{(i)} = \arg\min_{\omega_0\in\Omega}\left(\mathbf{x}^H\Pi^{\perp}(\omega_0, k^{(i)})\mathbf{x}\right)$
until (convergence)

IV. MAP CRITERION AND MODEL SELECTION

Model selection and segmentation can be done with a maximum a posteriori (MAP) model selection criterion. The principle behind the MAP criterion is to choose the model, $\mathcal{M}$,

that maximises the posterior probability given the observed data, $\mathbf{x}$:

$$\widehat{\mathcal{M}} = \arg\max_{\mathcal{M}} p(\mathcal{M}|\mathbf{x}). \tag{41}$$

Using Bayes' theorem [37] this can be rewritten as:

$$\widehat{\mathcal{M}} = \arg\max_{\mathcal{M}} \frac{p(\mathbf{x}|\mathcal{M})p(\mathcal{M})}{p(\mathbf{x})}. \tag{42}$$

Choosing the same prior probability, $p(\mathcal{M})$, for every model to avoid favouring any model beforehand, and noting that the probability of a given data vector, $p(\mathbf{x})$, is constant once it has been observed, the MAP estimate can be reduced to:

$$\widehat{\mathcal{M}} = \arg\max_{\mathcal{M}} p(\mathbf{x}|\mathcal{M}), \tag{43}$$

which is the likelihood of the observed data given the model. The likelihood is also dependent on other parameters like the fundamental frequency and the model order. As opposed to the maximum likelihood approach, these have to be integrated out in the Bayesian framework to give the marginal density of the data given the model [8]:

$$p(\mathbf{x}|\mathcal{M}) = \int_{\boldsymbol{\Theta}} p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})d\boldsymbol{\theta}. \tag{44}$$

An approximation to this integral can be found assuming high amounts of data and a likelihood that is highly peaked around the maximum likelihood estimates of $\boldsymbol{\theta}$ [8], [38], [39]

$$p(\mathbf{x}|\mathcal{M}) = \pi^{G/2} \det(\widehat{\mathbf{H}})^{-1/2} p(\mathbf{x}|\widehat{\boldsymbol{\theta}}, \mathcal{M})p(\widehat{\boldsymbol{\theta}}|\mathcal{M}), \tag{45}$$

where $\widehat{\mathbf{H}}$ is the Hessian of the log-likelihood function evaluated at $\widehat{\boldsymbol{\theta}}$:

$$\widehat{\mathbf{H}} = -\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}. \tag{46}$$

Now an expression for the MAP estimator can be found by taking the negative logarithm of (45). The term $\pi^{G/2}$ can be assumed constant for large N and is, therefore, neglected, while a weak prior on $p(\boldsymbol{\theta}|\mathcal{M})$ has been used [38] to obtain the expression [8]:

$$\widehat{\mathcal{M}} = \arg\min_{\mathcal{M}} -\ln \mathcal{L}(\widehat{\boldsymbol{\theta}}|\mathbf{x}) + \frac{1}{2}\ln\det(\widehat{\mathbf{H}}). \tag{47}$$

This corresponds to minimising a cost function, where the first part is the likelihood from (16), and the second part is a model-dependent penalty term.

The penalty term is found by noting that the Hessian is related to the Fisher information matrix in (32). Evaluating the Fisher information matrix at $\theta = \widehat{\theta}$ gives the expected value of the Hessian, and, therefore, the elements in the Hessian can be found by using (35)-(40). To ease complexity, an asymptotic expression for the Hessian can be found by looking at the elements of the matrix. The diagonal elements of the Hessian for the harmonic chirp model are given by:

$$\widehat{\mathbf{H}}_{\omega_0\omega_0} = \sum_{l=1}^{L} \frac{1}{12}(N^3 - N)l^2\widehat{A}_l^2, \tag{48}$$

$$\widehat{\mathbf{H}}_{kk} = \sum_{l=1}^{L} \frac{1}{960}(3N^5 - 10N^3 + 7N)l^2\widehat{A}_l^2, \tag{49}$$

$$\widehat{\mathbf{H}}_{A_lA_l} = N, \tag{50}$$

$$\widehat{\mathbf{H}}_{\phi_l\phi_l} = N\widehat{A}_l^2, \tag{51}$$

for N odd and $n_0 = -(N-1)/2$. From this, when the Hessian is evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, the model order and amplitudes can be considered constant, and the Hessian is then only dependent on N. To make this dependency negligible, a diagonal normalisation matrix, $\mathbf{K}$, is introduced [8], [40]

$$\mathbf{K} = \begin{bmatrix} N^{-3/2} & & \mathbf{0} \\ & N^{-5/2} & \\ \mathbf{0} & & N^{-1/2}\mathbf{I}_{2L} \end{bmatrix}, \tag{52}$$

resulting in

$$\widehat{\mathbf{H}} = \mathbf{K}^{-1}\mathbf{K}\widehat{\mathbf{H}}\mathbf{K}\mathbf{K}^{-1}. \tag{53}$$

The definition of the elements in $\mathbf{K}$ as $N^{-x/2}$ instead of N^{-x}, where $x = 1, 3, 5$, and multiplication with $\mathbf{K}$ from both sides is done to ensure that also the off-diagonal elements of $\widehat{\mathbf{H}}$ are compensated for in the right way. The determinant of the Hessian is then given by:

$$\det(\widehat{\mathbf{H}}) = \det(\mathbf{K}^{-2})\det(\mathbf{K}\widehat{\mathbf{H}}\mathbf{K}), \tag{54}$$

where the main dependency on N is now moved to the term $\mathbf{K}^{-2}$ whereas $\mathbf{K}\widehat{\mathbf{H}}\mathbf{K}$ is assumed small and constant for large N. Taking the natural logarithm of the determinant gives:

$$\ln\det(\widehat{\mathbf{H}}) = \ln\det(\mathbf{K}^{-2}) + \ln\det(\mathbf{K}\widehat{\mathbf{H}}\mathbf{K}) \tag{55}$$

$$= 3\ln N + 5\ln N + 2L\ln N + \mathcal{O}(1). \tag{56}$$

An expression for the cost associated with the harmonic chirp model can now be found by combining the log likelihood for the harmonic chirp model in (20) with the penalty term in (56) where the term $\mathcal{O}(1)$ is ignored:

$$J_c = N\ln\pi + N\ln\frac{1}{N}||\mathbf{x} - \mathbf{Z}\mathbf{a}||_2^2 + N + \frac{3}{2}\ln N + \frac{5}{2}\ln N + L\ln N. \tag{57}$$

For the traditional harmonic model, the Hessian will not contain a term related to the chirp rate, k, and the penalty for the MAP estimator will, therefore, also be short of this term:

$$J_h = N\ln\pi + N\ln\frac{1}{N}||\mathbf{x} - \mathbf{Z_0}\mathbf{a}||_2^2 + N + \frac{3}{2}\ln N + L\ln N, \tag{58}$$

where $\mathbf{Z}_0$ equals $\mathbf{Z}$ for $k = 0$. The MAP expressions for the harmonic chirp model and the traditional harmonic model can be used to choose between them by choosing the one with the smallest cost. Due to Occam's razor [41], the simplest model is always preferred if the models describe the signal equally well.

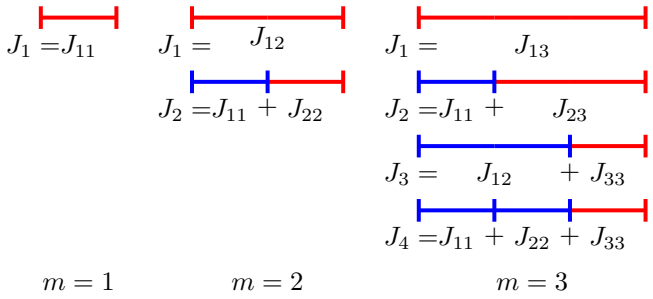


Fig. 2: Principle of segmentation. $M = 3$. Modified from [27].

This is assured by the extra penalty that naturally appears in the MAP expression for the chirp model. The error between the chirp model and the observed signal has to decrease enough relative to the traditional harmonic model to outweigh this penalty term before the chirp model is favoured over the traditional harmonic model. Aside from choosing between the two different harmonic models, the MAP estimator can also be used for voiced/unvoiced detection by determining whether a harmonic signal is present or not by comparing the two models with a zero order model,

$$J_0 = N \ln \pi + N \ln \sigma_x^2 + N, \tag{59}$$

where σ_x^2 is the variance of the observed signal. The voiced/unvoiced detection can also be done by using the generalised likelihood ratio test (GLRT) [42], [43]. In this method, the ratio of the likelihood of the presence of voiced speech found based on the harmonic model to the likelihood of a noise-only signal is calculated and compared to a threshold. The method has a constant false alarm ratio (CFAR) and so the threshold is set to ensure a given CFAR that is independent of the signal-to-noise ratio (SNR). Other methods as, e.g, described in [14], [15] can also be used.

V. Segmentation

The characteristics of the observed signal are varying over time and sometimes faster than others, meaning that a fixed segment length is not optimal. Using the MAP criterion, the cost associated with different segment lengths can be compared and the optimal chosen being the one minimising (57). The segmentation assures that the optimal trade-off between segment length and fit of the model is found, and so the segment length is chosen as long as possible without introducing too large modelling errors. It follows from the CRLB that long segments are desired and gives higher estimation accuracy. The segmentation is based on the principle in [27], [28] which is sketched in Fig. 2. In the figure, J_{xy} is the cost of a segment starting at block x and ending at block y, with both block x and y included in the segment.

A minimal segment length, N_{min}, is chosen, generating a block of N_{min} samples and dividing the signal into M blocks. Since this will give 2^{M-1} ways of segmenting the signal, a maximum number of blocks in one segment, K_{max}, is also set since very long segments are highly unlikely, and setting a maximum will bound the computational complexity. The maximum number of samples in one segment is, therefore, $N_{\text{max}} = K_{\text{max}} N_{\text{min}}$. Using a dynamic programming algorithm, the optimal number of blocks in a segment, k_{opt}, is found for all blocks, $m = 1, ..., M$, starting at $m = 1$ moving continuously to $m = M$. For each block, the cost of all new block combinations is calculated while old combinations are reused from earlier blocks. Relating to Fig. 2, the red segments are calculated whereas the blue segments are reused from earlier. To decrease the number of calculations further, only a block combination minimising the cost is used in a later step, which in Fig. 2 means that only one of J_3 and J_4 is considered for $m = 3$, corresponding to the block combination that minimised the cost at $m = 2$. When the end of the signal is reached, backtracking is used to find the optimal segmentation of the signal, starting at the last block, and jumping through the signal to the beginning. This is done by starting at $m = M$ and setting the number of blocks in the last segment of the signal to $k_{\text{opt}}(M)$. In this way, the next segment ends at block $m = M - k_{\text{opt}}(M)$ and includes $k_{\text{opt}}(M - k_{\text{opt}}(M))$ blocks. This is continued until $m = 0$. The segmentation is summarised in Table II.

TABLE II: Segmentation.

while $\boldsymbol{m \times N_{\text{min}} \leq}$ **length(signal)**
  $K = \min([m, K_{\text{max}}])$
  for $\boldsymbol{k = 1 : K}$
    blocks of signal to use is $m - k + 1, ..., m$
    find analytic signal and downsample
    estimate ω_0 and k using Table I
    estimate $\mathbf{a}$ and $\mathbf{Z}$ from (23), (8) and (9)
    calculate $J_{(m-k+1)m}$ from (57)
    $J(k) = \begin{cases} J_{(m-k+1)m} + J_{1(m-k)} & \text{if } m-k>0, \\ J_{(m-k+1)m} & \text{otherwise.} \end{cases}$
  end for
  $k_{\text{opt}}(m) = \arg\min J(k)$
  $m = m + 1$
end while

backtrack
$m = M$
while $\boldsymbol{m > 0}$
  number of blocks in segment is $k_{\text{opt}}(m)$
  $m = m - k_{\text{opt}}(m)$
end while

VI. Prewhitening

The maximum likelihood estimates of the fundamental frequency and chirp rate and the MAP model selection and segmentation criterion were found under the assumption of white Gaussian noise. However, in real life scenarios the noise is not always white. A prewhitening step is therefore required. The observed signal can be prewhitened by passing it through a filter that changes the noise from coloured to white. This is illustrated in Fig. 3. In the figure, $H(z)$ is a filter with a frequency response similar to the spectrum of the noise. The coloured noise can be seen as white noise filtered using a filter with coefficients given by $H(z)$. Therefore, to obtain a flat frequency spectrum of the noise, the action is reversed by dividing by $H(z)$, here denoted by $A(z)$. Naturally, the desired signal will also be altered by the passage through the filter. This may have an influence on the results depending on how much the signal is changed, and what the prewhitened signal

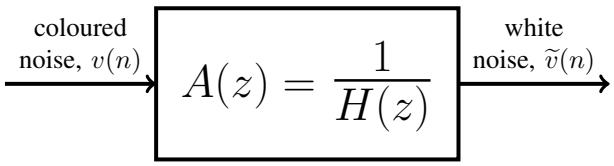


Fig. 3: Prewhitening of noise by passing it through the filter $A(z)$.

is used for. At the very best, the linear transformation of the signal will not affect the CRLB of the parameter estimation.

To obtain $H(z)$, information about the noise spectrum is needed. Different methods exist to estimate the power spectral density (PSD) of the noise given a mixture of desired signal and noise [30]–[33]. The PSD can be used directly to generate a simple finite impulse response (FIR) filter based on the frequency coefficients of the PSD. Alternatively, also based on the PSD, linear prediction (LP) can be used to find the characteristic parts of the noise spectrum and filter the observed signal based on this. In linear prediction, the present sample is estimated based on P prior samples:

$$\widehat{v}(n) = -\sum_{p=1}^{P} a_p v(n-p), \tag{60}$$

leading to a filter of the form:

$$A(z) = 1 + \sum_{p=1}^{P} a_p z^{-p}. \tag{61}$$

After filtering, the signal is normalised to have the same standard deviation before and after the filtering. To ensure that the desired signal has a smooth evolution over time after filtering, i.e., no drastic changes in amplitude or phase, it is important that the PSD is smooth. This is ensured by most recent PSD methods where the value in one time frame is a weighted combination of the preceding time frame and an estimate from the current time frame.

VII. Simulations

In the following, the different proposed methods are tested through simulations on synthetic signals and speech. The synthetic signals are made according to (7). Unless otherwise stated in the specific subsections, the signals were generated with $L = 10$, $A_l = 1 \forall l$, random phase, fundamental frequency, and fundamental chirp rate, in the intervals $\phi_l \in [0, 2\pi]$, $f_0 \in [100, 300]$ Hz, $k \in [-500, 500]$ Hz/s and the sampling frequency, f_s, was set to 8000 Hz.

The speech signal, "Why were you away a year, Roy?", was used in some simulations and to illustrate the function of some methods. The sentence is uttered by a female speaker and sampled at 8000 Hz. Additionally, the five male and five female speech signals from the Keele database [44] are used. The signals have a duration of approximately 30 seconds each. The signals are downsampled to 8000 Hz. With these signals, follow the corresponding laryngograph signals and an annotated fundamental frequency that can be used for evaluation of the proposed method. However, it should be noted that the annotated fundamental frequency is also only an estimate and not the ground truth.

In most experiments, it is desirable to evaluate the methods at different SNRs, e.g., in an interval from -10 to 10 dB to simulate situations with different levels of background noise. Therefore, noise was added to the signals with a variance calculated to fit the desired input SNR defined as

$$\text{iSNR} = \frac{\sigma_s^2}{\sigma_v^2}, \tag{62}$$

where σ_s^2 is the variance of the desired signal. The noise signals used are white Gaussian noise, as well as different types of noise from the AURORA database [45].

For each segment of noisy speech, the discrete-time analytic signal [34] is computed, and the parameter estimation is performed on this complex, downsampled version of the signal.

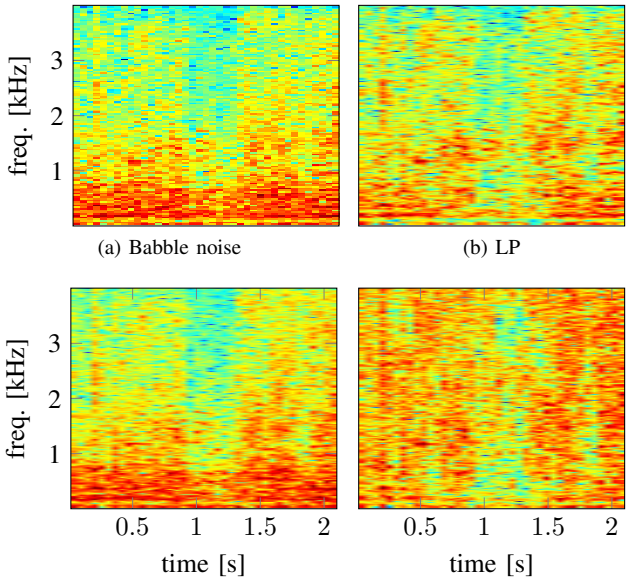
A. Prewhitening

The prewhitening using the FIR filter and LP is tested on "Why were you away a year, Roy?" and compared to prewhitening using Cholesky factorisation [46]. The signal is added noise at input SNRs of 0 and 10 dB, and the prewhitening filters are generated based on the noisy signal. The PSD is found using an implementation of [31] given in [30]. The PSD is obtained using 256 frequency points which equal the number of coefficients in the FIR filter, whereas the LP filter is made with five coefficients. The spectrum of babble noise at an input SNR of 10 dB before and after prewhitening is shown in Fig. 4. Here, it seems that the whitest noise signal is obtained using the Cholesky factorisation, followed by LP, while the FIR filter seems to make a minor change to the original noise.

The prewhitening methods are compared by means of the spectral flatness, $\mathcal{F}$, which is the ratio of the geometric mean to the arithmetic mean of the power spectrum, $S(k)$, [47]:

$$\mathcal{F} = \frac{\left(\prod_{k=0}^{K-1} S(k)\right)^{1/K}}{\frac{1}{K}\sum_{k=0}^{K-1} S(k)}. \tag{63}$$

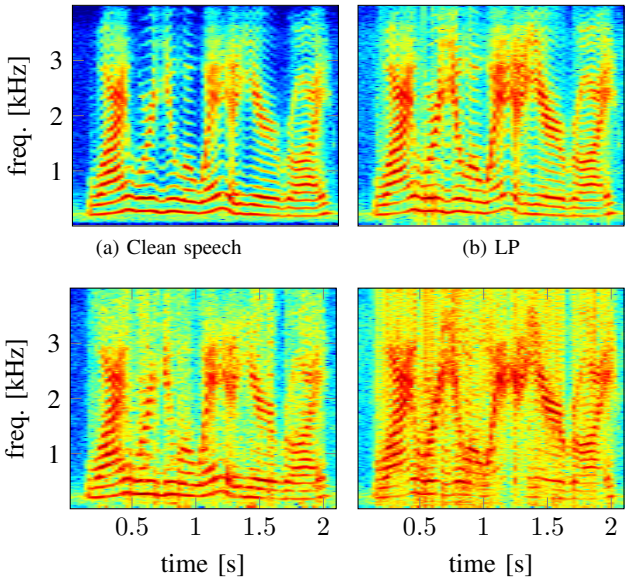
The spectral flatness gives a number between zero and one, where perfect white noise has a value of one. The spectral flatness for four different noise types at 0 and 10 dB is shown in Fig. 5, where the spectral flatness of the original noise and a white noise signal generated with MATLAB's `randn` are also shown for comparison. The spectral flatness is very similar at 0 and 10 dB for all noise types using a given prewhitening method. The results confirm the tendencies observed in Fig. 4. The Cholesky factorisation leads to the highest spectral flatness for all noise types, followed by linear prediction in the case of babble, car and street noise, while the FIR filter is better than linear prediction for exhibition noise. There is, however, large differences between the different noise types in how significant the advantage is of using one prewhitening method over another. The Cholesky factorisation is clearly best in terms of whitening the noise, but as is shown in Fig. 6, it is also the method that has the largest influence on the



(a) Babble noise (b) LP

(c) FIR (d) Cholesky

Fig. 4: Spectrograms of babble noise before (a) and after prewhitening with (b) LP filter, (c) FIR filter and (d) Cholesky factorisation. The four spectrograms are plotted with the same limits in dB.



(a) Clean speech (b) LP

(c) FIR (d) Cholesky

Fig. 6: Spectrograms of speech signal before (a) and after prewhitening with (b) LP filter, (c) FIR filter and (d) Cholesky factorisation. The four spectrograms are plotted with the same limits in dB.

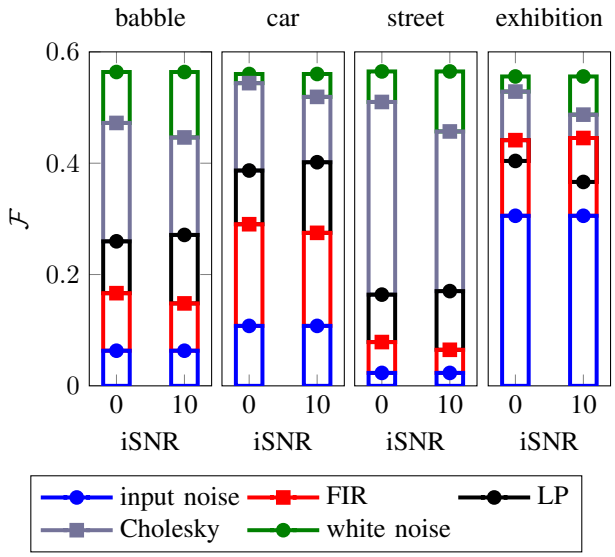


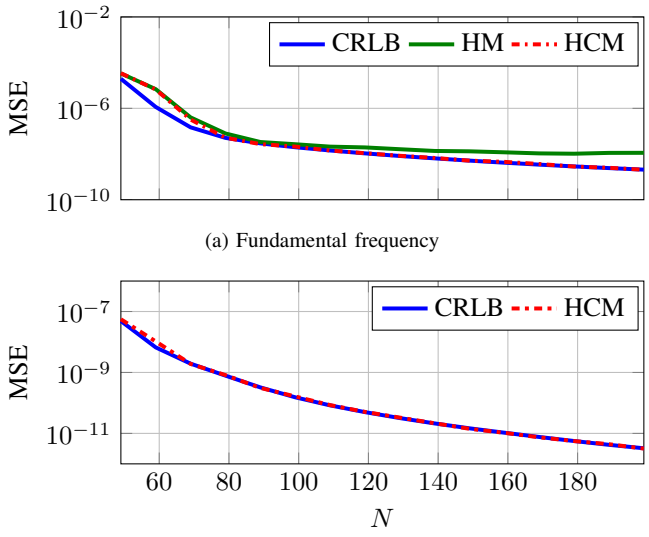
Fig. 5: Spectral flatness, $\mathcal{F}$, at 0 and 10 dB input SNR for original noise, prewhitened noise using FIR, LP and Cholesky factorisation. The spectral flatness for white noise is added for comparison.

desired signal. Here, it appears the LP filtering best preserves the desired signal with the FIR filter nearly as good, whereas the Cholesky factorisation clearly changes the appearance of the desired signal. Using the Cholesky factorisation for prewhitening, the signal model must be redefined to include the Cholesky matrix, as was done in [5]. Thus, it cannot be applied directly with the proposed model, and has been excluded from the following simulations. The FIR and LP filters only change the amplitude and phase, and, therefore, they only change the complex amplitude vector $\mathbf{a}$.

B. Fundamental frequency and chirp rate

The proposed estimator of fundamental frequency and chirp rate is first evaluated on synthetic chirp signals. Two experiments were made. In the first, the segment length, N, was varied from 49 to 199 samples with a fixed input SNR of 10 dB. In the second, the input SNR was varied from -10 to 10 dB with a fixed segment length of 199 samples. For each generated signal, noise was added, and an initial fundamental frequency estimate was found using a harmonic NLS estimator [8] with lower and upper limits of the search interval of 80 and 320 Hz. The model order is assumed known, i.e., $L = 10$. From here, the fundamental frequency and chirp rate were estimated, and the squared error was found. This was repeated 2000 times and the mean was taken to give the mean squared error (MSE). In Figs. 7 and 8, the MSE as a function of N and the input SNR is shown and compared to the CRLB and estimates obtained using a harmonic NLS estimator [8]. The chirp estimates reach the CRLB around a segment length of 110 and at an input SNR of around -5 dB under the given settings. The harmonic estimates are close to reaching the bound as well, but as the CRLB decreases for higher segment lengths and input SNRs, the error on the harmonic estimates do not decrease with the same rate resulting in a gap between the CRLB and the estimates.

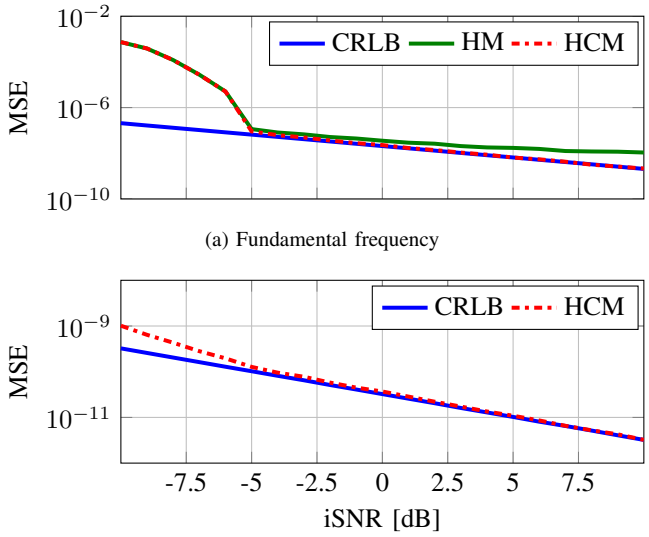
The estimator was used to estimate the fundamental frequency and chirp rate of "Why were you away a year, Roy?" with the spectrum shown in Fig. 6a. Here, the parameters are estimated directly from the clean signal in segments with



(a) Fundamental frequency

(b) Chirp rate

Fig. 7: Mean squared error (MSE) of the fundamental frequency and chirp rate as a function of N.



(a) Fundamental frequency

(b) Chirp rate

Fig. 8: Mean squared error for the fundamental frequency and chirp rate as a function of the input SNR.

a length of 198 samples (24.8 ms). The initial fundamental frequency estimate and model order were found jointly by using a harmonic NLS estimator and a MAP estimator [8], respectively. The limits on the harmonic fundamental frequency are set to 80 and 300 Hz. To confirm that the combination of the harmonic fundamental frequency and a chirp rate of zero is a good initialisation, an example of a two-dimensional cost function for a segment of a speech signal is shown in Fig. 9. The initialisation is marked by a yellow cross while the final estimate of fundamental frequency and chirp rate is marked by a red cross. As seen, the function is locally convex around the initial and true fundamental frequency and chirp rate. The figure also shows that the change in fundamental

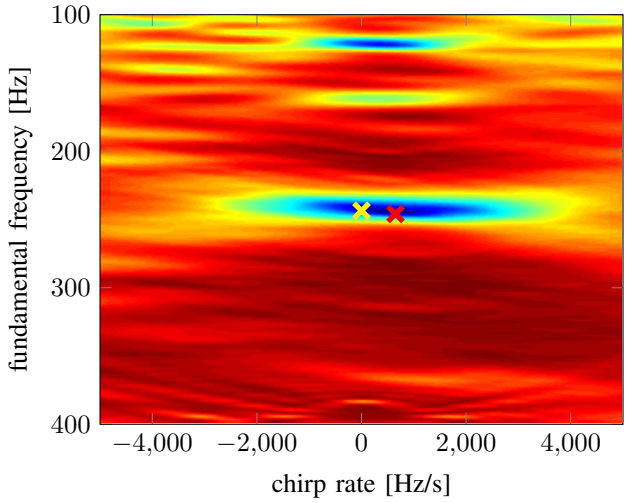


Fig. 9: Example of a cost function for a speech signal as a function of fundamental frequency and chirp rate.

frequency is rather small so if the fundamental frequency for some reason changes a lot, $\omega_0 < 0.6\omega_{0,\text{HM}}$ or $\omega_0 > 1.5\omega_{0,\text{HM}}$, the fundamental frequency is set to the harmonic estimate and the chirp rate is set to zero. However, it is important to note that the instantaneous fundamental frequency is not the same as the one found by the harmonic model. Now, the parameters are estimated in steps of 5 samples. The resulting estimates are shown in Fig. 10. The chirp rate can be interpreted as the tangent to the fundamental frequency curve at a given point. This means that the chirp rate should be negative when the fundamental frequency is decreasing, positive when it is increasing, and zero at a local maximum or minimum. To illustrate this, some maxima and minima of the fundamental frequency are marked by red stars in the upper plot and the chirp rates at the same points in time are marked in the bottom plot.

The estimation is repeated after the addition of noise to give an input SNR of 0 and 10 dB, but this time the parameters are only estimated once per segment of 198 samples. The estimation is done both for white Gaussian noise and babble noise as well as after prewhitening of the signal with babble noise using the FIR and LP filter. The sum of the absolute error between noisy and clean estimates is given in Table III at 0 and 10 dB. Here, only the time interval shown in Fig. 10 is considered since the beginning and end of the signal contain no speech. The white noise gives the best estimate at both 0 and 10 dB. At 0 dB, the LP prewhitened signal gives a lower error than the FIR filtered and clean babble noise whereas at 10 dB, the babble noise gives the lowest error followed by the FIR and LP filtered noise. This suggests that for the proposed ML estimator, the dominance of the desired signal at 10 dB decreases the importance of the noise shape relative to the effects of prewhitening on the signal. However, at 0 dB the noise is more dominant, and so the importance of prewhitening increases.

The fundamental frequency and chirp rate are also estimated from the signals in the Keele database. The fundamental

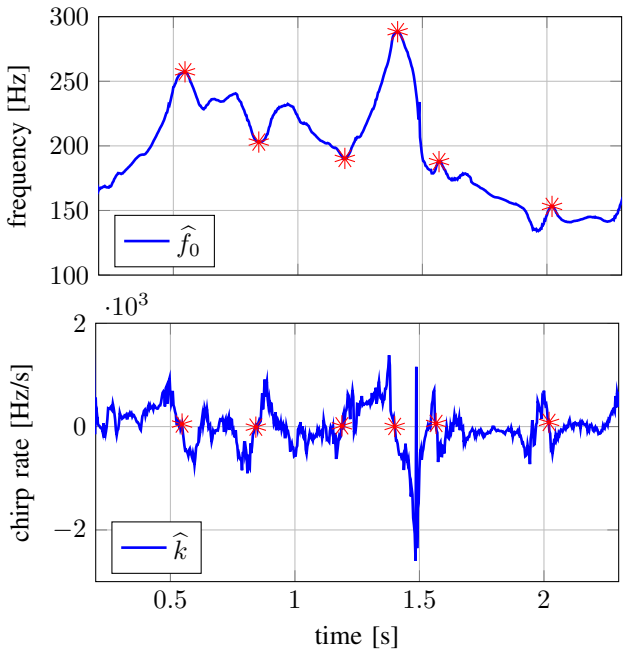


Fig. 10: Fundamental frequency and chirp rate estimation.

TABLE III: Sum of absolute error between noisy estimate and clean estimate of fundamental frequency in Hz at input SNRs of 0 and 10 dB.

	white noise	babble	FIR	LP
0 dB	585	2653	2483	1201
10 dB	167	408	714	787

frequency estimates are compared to YIN [2] and SWIPE [48] by means of the gross pitch error (GPE), the fine pitch error (FPE) and the reconstruction SNR. The GPE is defined as an estimate that deviates from the annotated fundamental frequency by more than 20 % [18]. The GPEs are not considered in the calculation of the FPE. The FPE is divided into two parts, the mean, μ, and the standard deviation, σ, of the errors on the estimates [12], [18]. Both are calculated from the difference between the estimated fundamental frequency and the annotated fundamental frequency. The annotated fundamental frequency is estimated in steps of 10 ms based on segments of 26.5 ms of data. This is also done for HM, HCM and YIN, however, it is not possible to choose the segment length in SWIPE. The lower and upper limit on the estimate are set to 50 and 300 Hz. The reconstruction SNR is calculated from the reconstructed signal based on (7). For YIN, SWIPE and the traditional harmonic model, the chirp rate is approximated by $\Delta f = (f_0(n+1) - f_0(n))/\Delta t$ where Δt is the time between two consecutive estimates of the fundamental frequency. Note that this will cause a delay in real-time applications. However, using past samples does not result in Δf for the correct segment and will degrade the reconstruction compared to only using the harmonic model. The estimated fundamental frequencies, chirp rates and Δf's are used in $\mathbf{Z}$ in (8). Since we are here considering non-stationary signals it makes a difference from where in the signal the reference point is set. From experiments on synthetic chirp signals it was found that YIN and SWIPE have the reference point towards the beginning of the signal whereas HM has its reference point around the middle. Therefore, we set $n_0 = 0$ for YIN and SWIPE and $n_0 = -N/2$ for HM. The mid-segment reference point for HM means that Δf is estimated incorrectly. The proper estimate would be $\Delta f = (f_0(n+1/2) - f_0(n-1/2))/\Delta t$, but this information is not available. The wrong estimate of Δf leads to a worse performance compared to using the harmonic model on its own. The result for HM without Δf is therefore also included in the comparison. The fundamental frequency, chirp rate and Δf are estimated for each 25 ms based on 25 ms of data, and the entire block of samples is reconstructed based on this estimate. The model order is estimated using a MAP estimator [8]. The amplitude vector, $\mathbf{a}$, is estimated using (23). The reconstruction SNR (rSNR) is then given by:

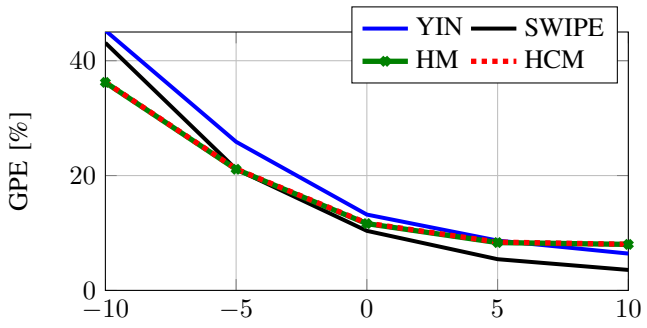
$$\text{rSNR} = \frac{\sigma_s^2}{\sigma_{(s-\hat{s})}^2}, \tag{64}$$

where $\hat{s}$ is the reconstructed signal, and $\sigma_{(s-\hat{s})}^2$ is the variance of the error signal between the original speech signal and the reconstructed signal.

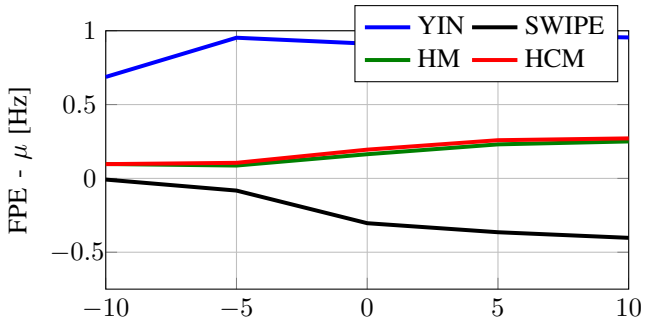
The results are shown in Fig. 11. In terms of GPE, the proposed method performs better than YIN and SWIPE at low input SNRs, while SWIPE is better at high input SNRs. The harmonic models perform equally. The bias, seen as the mean, μ, is small for all methods. It is approximately 1 Hz for YIN and within $\pm$ 0.5 Hz for the other methods. The proposed method does not perform as well as the traditional harmonic model in terms of standard deviation, σ. As mentioned earlier, the annotated fundamental frequency is not the ground truth, but a fundamental frequency estimate found from the laryngograph signal using an autocorrelation method which is also based on the harmonic assumption. In Fig. 9 it was seen that the instantaneous fundamental frequency found by the proposed method is not the same as the harmonic frequency. Therefore, it is not surprising that the method does not perform well when it is compared to the fundamental frequency estimated based on the harmonic assumption. Looking at the reconstruction SNR, the chirp model outperforms all other methods. The reconstruction SNR is the only of the four error measures that takes both fundamental frequency and chirp rate into account. Further, the reconstruction SNR does not depend on another estimate of the fundamental frequency as do the FPE and GPE, it compares to the original speech signal.

C. Model selection

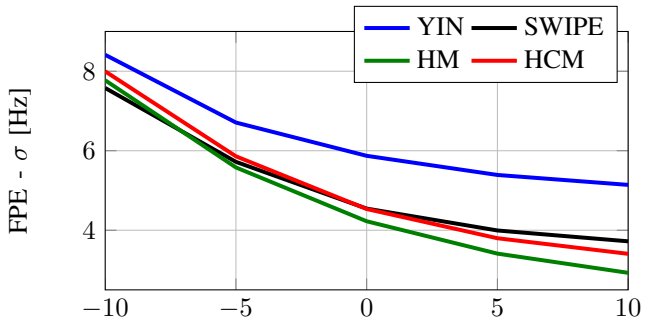
The model selection was first tested on synthetic signals degraded with white Gaussian noise to give an input SNR of 10 dB. In this part, the possible models included in the test are the traditional harmonic model and the harmonic chirp model. The model selection was tested for different chirp rates and different segment lengths. For each combination of chirp rate and segment length, 2000 signals were generated and the selected model was noted for each signal. The percent of the chirp model chosen is shown in Fig. 12. Even though



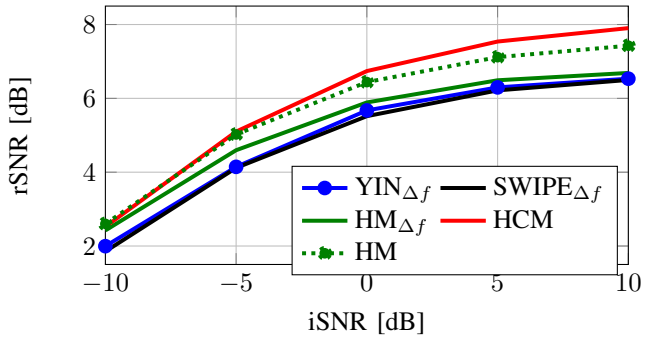
(a) Gross pitch error (GPE)



(b) Fine pitch error (FPE) - mean, μ



(c) Fine pitch error (FPE) - standard deviation, σ



(d) Reconstruction SNR

Fig. 11: Evaluation of the instantaneous fundamental frequency estimation by means of gross pitch error (GPE), fine pitch error (FPE) and reconstruction SNR (rSNR).

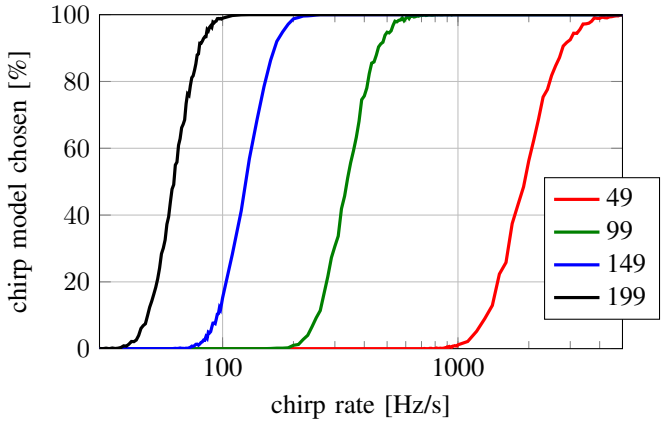


Fig. 12: Model selection for synthetic signals as a function of the chirp rate for different segment lengths from 49 to 199.

all generated signals, except for the ones with a chirp rate of zero, are chirp signals, the chirp model is not chosen in all cases. As mentioned in Section IV, this is due to the extra penalty term introduced to the chirp model and not to the harmonic model. The longer the signal is, the more prone it is to be denoted as a chirp signal since the error term $||\mathbf{x}-\mathbf{Za}||_2^2$ will increase with signal length when the model does not fit, making the cost of the harmonic model greater than that of the chirp model, despite the extra penalty to the chirp model.

Model selection was also performed on the speech signals from the Keele database in white Gaussian noise at different segment lengths. Here, the noise model is also included. The percentage of each chosen model is found by taking the number of segments labelled according to a given model out of the total number of segments in the signal. The result is shown in Fig. 13. The percentage of the chosen noise model is fairly independent of the segment length since the amount of unvoiced speech is independent of the segment length. For short segment lengths, the harmonic model is chosen approximately 55% of the time and the chirp model is never chosen, but as the segment length is increased, the two models are almost equally preferred. It should again be kept in mind that the chirp model has an extra penalty for being a more complex model so even though the error on the signal, $||\mathbf{x}-\mathbf{Za}||_2^2$, is smaller for the chirp model, it has to overcome the penalty as well before it is selected.

D. Segmentation

The segmentation is tested on the signal "Why were you away a year, Roy?". White Gaussian noise is added to the signal to give an input SNR of 10 dB. The signal is segmented according to the harmonic chirp model and the traditional harmonic model where, in both cases, the minimum segment length $N_{\text{min}} = 40$ and the maximum number of blocks $K_{\text{max}} = 10$, meaning that the minimum length of a segment is 40 samples (5 ms) and the maximum length of a segment, N_{max}, is 400 samples (50 ms). A representative example of the chosen segment length as a function of time is shown in Fig. 14. For comparison, the fundamental frequency estimate

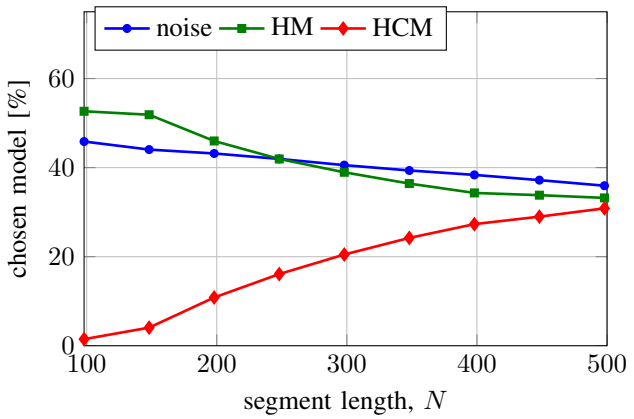


Fig. 13: Model selection as a function of the segment length (12.5 ms - 62.5 ms).

is plotted as well. In general, the chirp model gives rise to longer segment lengths than the traditional harmonic model. For this example, the average segment length is 195 samples (24.4 ms) using the chirp model and 137 samples (17.1 ms) using the traditional harmonic model. A typical choice of fixed segment length is 20–30 ms [13]. On average, this is a good choice when using the harmonic chirp model, however, shorter segments are better if the traditional harmonic model is used. The longer segments of the chirp model, of course, mean that the total number of segments is lower than for the harmonic model. The chirp model divides the signal into 105 segments and with the harmonic model, the number of segments is 150. Three areas in Fig. 14 are marked with circles as examples of the longer segments obtained with the chirp model. In the light blue circle, the fundamental frequency is decreasing quite fast, but the change is constant over time. Thus a long segment is obtained using the chirp model while shorter segments are obtained when the harmonic model is used. In the purple circle, the piece of speech is divided into four segments with the chirp model: two segments of maximum length, where the fundamental frequency is almost constant, and two shorter but still fairly long segments, where the fundamental frequency is increasing and decreasing, respectively. For the harmonic model, there are two long segments where the fundamental frequency is close to constant, but the rest of the piece is divided into shorter segments. In the brown circle, the piece is divided into two segments using the chirp model: one piece where the fundamental frequency is decreasing and one where it is increasing. The harmonic model covers the area in the middle, where the fundamental frequency is fairly constant, with two somewhat long segments, but in order to cover the whole area, shorter segments are added on both sides of the segments in the middle. The longer segments chosen for the chirp model suggests that the chirp model describes the signal in a better way than the traditional harmonic model since it to some extent takes the non-stationarity of the speech into account.

The signal is reconstructed using (7), as was done in the evaluation of the fundamental frequency estimate. The signal is reconstructed from the estimates in the optimal segments,

TABLE IV: Reconstruction SNR for chirp and harmonic signal using either optimal segmentation or a fixed segment length matching the mean segment length of the optimal segmentation, in this case $\bar{N}_{\text{HM}} = 140$ (17.5 ms) and $\bar{N}_{\text{HCM}} = 188$ (23.5 ms). The input SNR is 10 dB.

	chirp	harmonic
opt. segm.	12.49	12.38
fixed	10.88	11.29

TABLE V: Average segment length, $\bar{N}$, for chirp and harmonic signal for different noise types at 10 dB.

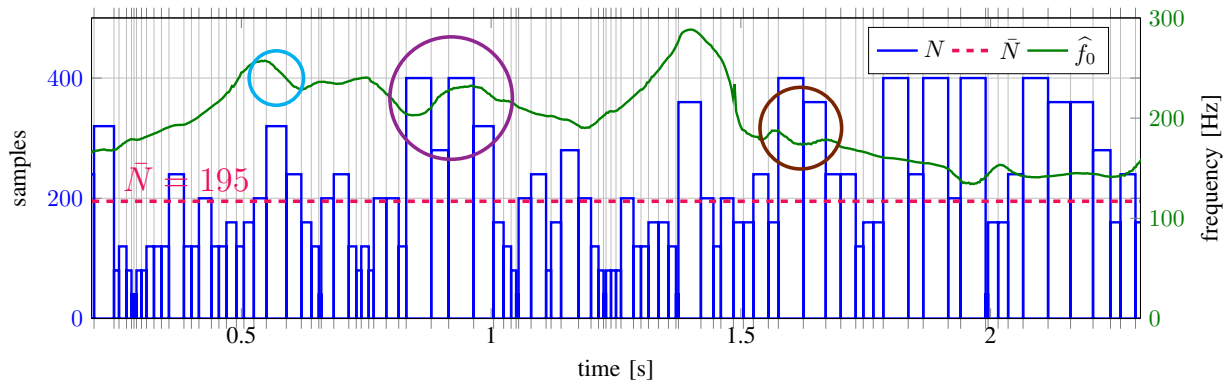
	chirp	harmonic
babble	69 (8.6 ms)	62 (7.7 ms)
FIR	73 (9.1 ms)	65 (8.1 ms)
LP	119 (14.9 ms)	91 (11.4 ms)

meaning that in some cases 40 samples (5 ms) are reconstructed based on one estimate of fundamental frequency and chirp rate, whereas in other cases, 400 samples (50 ms) are estimated based on one estimate. This is compared to estimates from segments with a fixed length where the length of the segments is set to the mean length of the segments from the optimal segmentation. In this case, $\bar{N}_{\text{HM}} = 140$ (17.5 ms) and $\bar{N}_{\text{HCM}} = 188$ (23.5 ms). This means that the reconstructions based on optimal segmentation and fixed segment length use the same number of segments to represent the signal. The reconstruction SNR is shown in Table IV. The table shows that with the same number of segments used for the reconstruction, a better reconstruction SNR can be obtained when optimal segmentation is used instead of using a fixed segment length. The reconstruction SNR is more than 1.5 dB better for the chirp model and more than 1 dB better for the traditional harmonic model when comparing optimal segmentation to a fixed segment length. By comparing the harmonic chirp model to the traditional harmonic model, a better reconstruction SNR is obtained with the harmonic chirp model when optimal segmentation is used, even though the chirp model uses only 109 segments and the traditional harmonic model uses 147 segments to represent the entire signal.

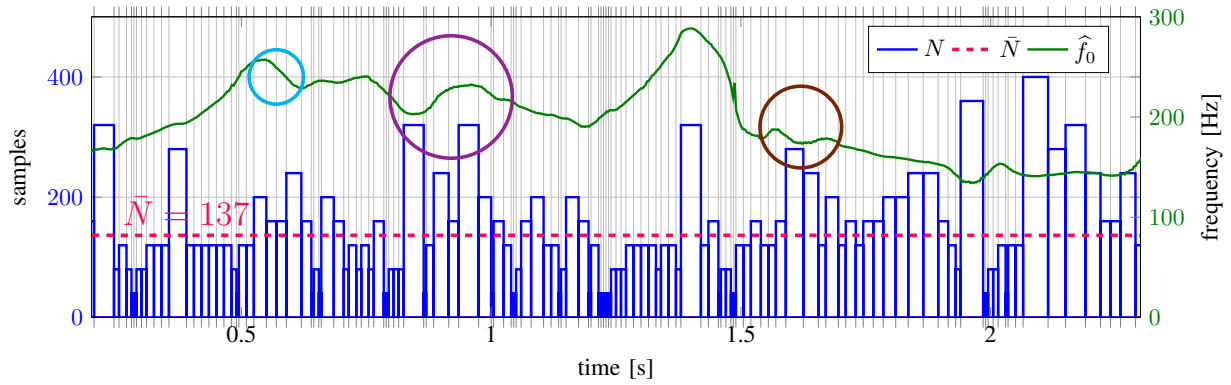
The segmentation is also tested for the signal in babble noise and prewhitened babble noise at an input SNR of 10 dB. The average segment lengths in the different cases are shown for the two models in Table V. In all cases, the signal is divided into longest segments when the chirp model is used. With respect to the different noise scenarios, the tendency is the same for the two models. The segments are shortest when the signal in babble noise is considered, followed closely by the prewhitened signal using FIR filtering. The longest segments are obtained with the LP filtered signal.

VIII. Conclusion

Traditionally, non-stationarity, fixed segment lengths and noise assumptions have limited the performance of fundamental frequency estimators. In this paper, we take these factors



(a) Harmonic chirp model



(b) Traditional harmonic model

Fig. 14: Segment length as a function of time for (a) the harmonic chirp model and (b) the traditional harmonic model. The average segment length, $\bar{N}$, is marked by the red line. The average segment length is 195 samples (24.4 ms) for the harmonic chirp model and 137 samples (17.1 ms) for the traditional harmonic model. The total number of segments is 105 for the chirp model and 150 for the harmonic model.

into account. We described the voiced part of a speech signal using a harmonic chirp model that allows the fundamental frequency to vary linearly within each segment. We proposed an iterative maximum likelihood estimator of the fundamental frequency and chirp rate based on this model. The estimator reaches the Cramer-Rao lower bound and shows expected correspondence between the estimate of the fundamental frequency and fundamental chirp rate of speech. Based on the maximum a posteriori (MAP) model selection criterion, the chirp model was shown to be preferred over the traditional harmonic model for long segments, suggesting that the chirp model is better at describing the non-stationary behaviour of voiced speech. Since the extent of the non-stationarity of speech changes over time, a fixed segment length is not optimal. Therefore, we also proposed varying the segment length based on the MAP criterion. Longer segments were obtained when the chirp model was used compared to the traditional harmonic model, again suggesting a better fit of the model to the speech. The maximum likelihood and MAP estimators are based on an assumption of white Gaussian noise. However, in real life the noise is rarely white. Therefore, we also suggested using two filters to prewhiten the noise, a simple FIR filter and one based on linear prediction (LP). They both have a minor influence on the speech signal, but the LP filter gives less error on the fundamental frequency estimate when the noise level is high. Further, the LP filter gives longer segment lengths in the optimal segmentation.

REFERENCES

[1] M. G. Christensen and J. R. Jensen, "Pitch estimation for non-stationary speech," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, Nov. 2014, pp. 1400–1404.

[2] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[3] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Elsevier Signal Process.*, vol. 80, no. 9, p. 19371944, 2000.

[4] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2012, pp. 409–412.

[5] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 76–87, 2004.

[6] M. G. Christensen, J. H. Jensen, A. Jakobsson, and S. H. Jensen, "Joint fundamental frequency and order estimation using optimal filtering." *EURASIP J. on Advances in Signal Processing*, vol. 2011, p. 13, 2011.

[7] P. Jain and R. B. Pachori, "Event-based method for instantaneous fundamental frequency estimation from voiced speech based on eigenvalue decomposition of the hankel matrix," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 22, no. 10, pp. 1467–1482, 2014.

[8] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[9] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear least squares methods for joint doa and pitch estimation," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 21, no. 5, pp. 923–933, 2013.

[10] A. M. Noll, "Short-time spectrum and "cepstrum" techniques for vocal-pitch detection," *J. Acoust. Soc. Am.*, vol. 36, no. 296, 1964.

[11] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[12] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 5, pp. 399–418, 1976.

[13] L. Deng and D. O'Shaughnessy, *Speech processing: a dynamic and optimization-oriented approach*. CRC Press, 2003.

[14] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 2, pp. 250–255, 1993.

[15] K. Molla, K. Hirose, N. Minematsu, and K. Hasan, "Voiced/unvoiced detection of speech signals using empirical mode decomposition model," in *Int. Conf. Information and Communication Technology*, March 2007, pp. 311–314.

[16] F. R. Drepper, "A two-level drive-response model of non-stationary speech signals," *Nonlinear Analyses and Algorithms for Speech Processing*, vol. 1, pp. 125–138, Apr. 2005.

[17] H. Kawahara, M. Morise, R. Nisimura, and T. Irino, "Higher order waveform symmetry measure and its application to periodicity detectors for speech and singing with fine temporal resolution," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2013, pp. 6797–6801.

[18] B. Yegnanarayana and K. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 4, pp. 614–624, 2009.

[19] D. Ealey, H. Kelleher, and D. Pearce, "Harmonic tunnelling: tracking non-stationary noises during speech," in *Proc. Eurospeech*, Sep. 2001, pp. 437–440.

[20] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 20, no. 7, pp. 1948–1963, Sep. 2012.

[21] T. Nilsson, S. I. Adalbjornsson, N. R. Butt, and A. Jakobsson, "Multi-pitch estimation of inharmonic signals," in *Proc. European Signal Processing Conf.*, 2013, pp. 1–5.

[22] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech and Language Process. (TASLP)*, vol. 22, no. 12, pp. 1931–1940, 2014.

[23] P. M. Djuric and S. M. Kay, "Parameter estimation of chirp signals," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 38, no. 12, pp. 2118–2126, 1990.

[24] Y. Pantazis, O. Rosec, and Y. Stylianou, "Chirp rate estimation of speech based on a time-varying quasi-harmonic model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2009, pp. 3985–3988.

[25] Y. Doweck, A. Amar, and I. Cohen, "Joint model order selection and parameter estimation of chirps with harmonic components," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1765–1778, Apr. 2015.

[26] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Enhancement of non-stationary speech using harmonic chirp filters," in *Proc. Interspeech*, Sep. 2015, accepted for publication.

[27] P. Prandoni, M. M. Goodwin, and M. Vetterli, "Optimal time segmentation for signal modeling and compression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1997, pp. 2029–2032.

[28] P. Prandoni and M. Vetterli, "R/D optimal linear prediction," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 646–655, 2000.

[29] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP J. on Advances in Signal Processing*, vol. 2007, no. 1, p. 24, Jun. 2007, article ID 092953.

[30] P. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[31] K. V. Sørensen and S. V. Andersen, "Speech enhancement with natural sounding residual noise based on connected time-frequency speech presence regions," *EURASIP J. on Advances in Signal Processing*, vol. 2005, no. 18, pp. 2954–2964, 2005.

[32] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[33] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

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

[35] A. Antoniou and W. S. Lu, *Practical Optimization - Algorithms and Engineering Applications*. Springer Science+Business Media, 2007.

[36] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, Inc., 1993.

[37] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006, vol. 1.

[38] P. M. Djuric, "A model selection rule for sinusoids in white gaussian noise," *IEEE Trans. Signal Process.*, vol. 44, no. 7, pp. 1744–1751, 1996.

[39] ——, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726–2735, 1998.

[40] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, 2004.

[41] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[42] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice Hall, Inc., 1998.

[43] E. Fisher, J. Tabrikian, and S. Dubnov, "Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 2, pp. 502–510, 2006.

[44] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. Eurospeech*, Sep. 1995, pp. 837–840.

[45] D. Pearce and H. G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. Int. Conf. Spoken Language Process.*, Oct 2000.

[46] P. C. Hansen and S. H. Jensen, "Prewhitening for rank-deficient noise in subspace methods for noise reduction," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3718–3726, 2005.

[47] N. S. Jayant and P. Noll, *Digital coding of wafeforms*. Prentice-Hall, 1984.

[48] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–1652, 2008.

Sidsel Marie Nørholm received her B.Sc. in electronics in 2007 and her M.Sc. in acoustics in 2012 both from the Technical University of Denmark. Currently, she is pursuing a Ph.D. in speech signal processing at the Audio Analysis Lab in the Department of Architecture, Design & Media Technology at Aalborg University in Denmark. She has been a visiting researcher at the University of Oldenburg, Germany.

Her research interests include signal processing methods, especially speech models, speech enhancement and parameter estimation.

Jesper Rindom Jensen (S'09–M'12) was born in Ringkøbing, Denmark in August 1984. He received the M.Sc. degree *cum laude* for completing the elite candidate education in 2009 from Aalborg University in Denmark. In 2012, he received the Ph.D. degree from Aalborg University. Currently, he is a Postdoctoral Researcher at the Department of Architecture, Design & Media Technology at Aalborg University in Denmark, where he is also a member of the Audio Analysis Lab. He has been a Visiting Researcher at the University of Quebec, INRS-EMT, in Montreal, Quebec, Canada, and at the Friedrich-Alexander Universität Erlangen-Nürnberg in Erlangen, Germany.

His research interests include signal processing theory and methods for, e.g., microphone array and joint audio-visual signal processing. Examples of more specific research interests within this scope are enhancement, separation, localization, tracking, parametric analysis, and modeling. He has published nearly more than 50 papers on these topics in top-tier, peer-reviewed conference proceedings and journals. Moreover, he has published 2 research monographs including the book "Speech Enhancement - A Signal Subspace Perspective" which is co-authored with Prof. Jacob Benesty, Prof. Mads Græsbøll Christensen, and Prof. Jingdong Chen.

He has received a highly competitive postdoc grant from the Danish Independent Research Council, as well as several travel grants from private foundations. Furthermore, he is an affiliate member of the IEEE Signal Processing Theory and Methods Technical Committee, and is Member of the IEEE.

Mads Græsbøll Christensen (S'00–M'05–SM'11)) received the M.Sc. and Ph.D. degrees in 2002 and 2005, respectively, from Aalborg University (AAU) in Denmark, where he is also currently employed at the Dept. of Architecture, Design & Media Technology as Professor in Audio Processing and is head and founder of the Audio Analysis Lab.

He was formerly with the Dept. of Electronic Systems at AAU and has been held visiting positions at Philips Research Labs, ENST, UCSB, and Columbia University. He has published 3 books and more than 150 papers in peer-reviewed conference proceedings and journals, and he has given tutorials at EUSIPCO and INTERSPEECH. His research interests include signal processing theory and methods with application to speech and audio, in particular parametric analysis, modeling, enhancement, separation, and coding.

Prof. Christensen has received several awards, including an ICASSP Student Paper Contest Award, the Spar Nord Foundations Research Prize, a Danish Independent Research Council Young Researchers Award, and the Statoil Prize, and he is also co-author of the paper *Sparse Linear Prediction and Its Application to Speech Processing* that received an IEEE Signal Processing Society Young Author Best Paper Award. Moreover, he is a beneficiary of major grants from the Danish Independent Research Council, the Villum Foundation, and Innovation Fund Denmark. He is an Associate Editor for IEEE/ACM Trans. on Audio, Speech, and Language Processing, a former Associate Editor of IEEE Signal Processing Letters, a member of the IEEE Audio and Acoustic Signal Processing Technical Committee, and a Senior Member of the IEEE.

Paper E

Enhancement and Noise Statistics Estimation for Non-Stationary Voiced Speech

S. M. Nørholm, J. R. Jensen, and **M. G. Christensen**

The paper has been published in the
IEEE/ACM Transacstions on Audio, Speech, and Language Processing,
vol. 24, no. 4, pp. 645-–658, April 2016.

Enhancement and Noise Statistics Estimation for Non-Stationary Voiced Speech

Sidsel Marie Nørholm, Jesper Rindom Jensen, *Member, IEEE,*
and Mads Græsbøll Christensen, *Senior Member, IEEE*

***Abstract*—In this paper, single channel speech enhancement in the time domain is considered. We address the problem of modelling non-stationary speech by describing the voiced speech parts by a harmonic linear chirp model instead of using the traditional harmonic model. This means that the speech signal is not assumed stationary, instead the fundamental frequency can vary linearly within each frame. The linearly constrained minimum variance (LCMV) filter and the amplitude and phase estimation (APES) filter are derived in this framework and compared to the harmonic versions of the same filters. It is shown through simulations on synthetic and speech signals, that the chirp versions of the filters perform better than their harmonic counterparts in terms of output signal-to-noise ratio (SNR) and signal reduction factor. For synthetic signals, the output SNR for the harmonic chirp APES based filter is increased 3 dB compared to the harmonic APES based filter at an input SNR of 10 dB, and at the same time the signal reduction factor is decreased. For speech signals, the increase is 1.5 dB along with a decrease in the signal reduction factor of 0.7. As an implicit part of the APES filter, a noise covariance matrix estimate is obtained. We suggest using this estimate in combination with other filters such as the Wiener filter. The performance of the Wiener filter and LCMV filter are compared using the APES noise covariance matrix estimate and a power spectral density (PSD) based noise covariance matrix estimate. It is shown that the APES covariance matrix works well in combination with the Wiener filter, and the PSD based covariance matrix works well in combination with the LCMV filter.**

***Index Terms*—Speech enhancement, chirp model, harmonic signal model, non-stationary speech.**

I. Introduction

SPEECH enhancement has many applications as in, e.g., mobile phones and hearing aids. Often, the speech enhancement is carried out in a transformed domain, a common one being the frequency domain. Here, the methods based on computational auditory scene analysis (CASA) [2], [3], spectral subtraction [4] and Wiener filtering [5] are well known methods. The CASA methods are based on feature extraction of the speech signal whereas spectral subtraction and Wiener filtering require an estimate of the power spectral density (PSD) of the noise. The PSD can be estimated in different ways [6]–[8], but common to these methods is that they primarily rely on periods without speech to update the noise statistics. In periods of speech, the PSD is mostly given by the previous estimate of the PSD. This update pattern makes the PSD estimates very vulnerable to non-stationary noise. Furthermore, in order to make enhancement in the frequency domain, the data needs to be transformed by use of the Fourier transform. This transform assumes that the signals are stationary within the analysis window which for speech signals is often between 20 ms and 30 ms. It is, however, well known that this assumption of stationary speech does not hold [9], [10], as, e.g., the fundamental frequency and formants vary continuously over time in periods of voiced speech, making the speech signal non-stationary. One example of this is the diphthong where one vowel is followed directly by another with a smooth transition. In [11], [12], it is suggested replacing the standard Fourier transform with a fan-chirp transform in the analysis of non-stationary harmonic signals. The voiced speech parts of a speech signal are often described by a harmonic model, and since voiced speech is the main constituent of speech, it makes good sense to use this transform on speech signals. The voiced speech can also easily be separated from the unvoiced speech by use of voiced/unvoiced detectors [13], [14]. The assumption behind the fan-chirp transform is that the harmonic frequencies of the signal vary linearly over time, and it is shown that spectra obtained using the fan-chirp transform have much more distinct peaks at the positions of the harmonic frequencies. Alternatively, the enhancement can be done directly in the time domain where, e.g., the Wiener filter has also been defined [15]. Most time domain filters also depend on noise statistics in the form of a covariance matrix. These are often obtained by averaging over a small frame of the observed signal, and, therefore, the signal in these frames is also assumed stationary. Also, a common way to filter speech in the time domain is by describing the voiced speech parts by a harmonic model [16]–[18]. The signal based on this model is composed of a set of sinusoids where the frequency of each sinusoid is given by an integer multiple of a fundamental frequency. The fundamental frequency in this model is constant within a frame, and so the voiced speech is assumed stationary. In [17], it is proposed estimating the noise by subtracting an estimate of the desired signal based on the harmonic model, and, from this, make a noise covariance matrix estimate. In doing so, the observed signal only needs to be stationary within the frame of 20 to 30 ms when the noise statistics are estimated and not from one speech free period to the next, as was mostly the case for the PSD. The non-stationarity of speech is considered in [19]–[21] in relation to modelling and parameter estimation. In these papers, a modified version of the harmonic model is used

This work was funded by the Villum Foundation and the Danish Council for Independent Research, grant ID: DFF 1337-00084. Part of this material will be published at INTERSPEECH 2015 [1].

S. M. Nørholm, J. R. Jensen and M. G. Christensen are with the Audio Analysis Lab, AD:MT, Aalborg University, DK-9000 Aalborg, Denmark, e-mail: {smn, jrj, mgc}@create.aau.dk

where a chirp parameter is introduced to allow the frequency of the harmonics to change linearly within each frame. In [19], the first model introduced to describe the speech signal is very flexible, but it is approximated with a Taylor expansion that leads to bigger and bigger deviations from the original model when the harmonic number increases, as mentioned in the paper. In [20], [21], a harmonic chirp model is used to describe the voiced speech, and the parameters of the model are estimated based on maximum likelihood estimation, but using different ways to avoid making a two dimensional search for the fundamental frequency and chirp rate.

In this work we want to explore if there is a benefit of taking the non-stationarity of speech into account when speech enhancement is considered. Therefore, we investigate the harmonic chirp model further in relation to speech enhancement. The linearly constrained minimum variance (LCMV) and the amplitude and phase estimation (APES) filters have previously been derived under the harmonic framework [18], [22], [23]. One objective of this work is to increase the performance of these filters by deriving them according to the harmonic chirp model. Both LCMV and APES filter have the goal of minimising the output noise power from the filter under the constraint that the desired signal is passed undistorted, or equivalently, when the constraint is fulfilled, to maximise the output signal-to-noise ratio (SNR). Therefore, we evaluate the performance of the filters by use of the output SNR and the signal reduction factor which measures the distortion of the desired signal introduced by the filters. Another objective is to investigate the noise covariance matrix that is obtained implicitly when the APES based filter is made in relation to other filters as, e.g., the Wiener filter. The noise covariance matrix estimate is made under the assumption of non-stationary speech when the harmonic chirp model is used. It is generated from the covariance matrix of the observed signal by subtracting the part that conforms to the harmonic chirp model. We propose using this estimate in combination with other filters as well and compare the performance of the Wiener filter using the APES noise covariance matrix to the chirp APES based filter. Alternatively, we suggest estimating the noise covariance matrix based on the earlier mentioned state of the art PSD estimates [7], [8] since more work has been put into noise PSD estimates than estimation of time domain noise statistics. The PSD is related through the Fourier transform to the autocorrelation and, thereby, to the covariance matrix as well.

In Section II, the harmonic chirp model is introduced. In Section III, the LCMV and APES based filters for harmonic chirp signals are derived. The Wiener filter and a family of trade-off filters are then introduced. In Section IV, the estimation of covariance matrices are discussed and suggestions on how to do it is given. In Section V, the performance of the LCMV and APES filters are considered through derivations of the used performance measures. In Section VI, experimental results on synthetic and real speech signals are shown and discussed, and the presented work is concluded in Section VII.

II. Framework

We are here considering the problem of recovering a desired signal, $s(n)$, from an observed signal, $x(n)$, with the desired signal buried in additive noise, i.e.,

$$x(n) = s(n) + v(n), \tag{1}$$

for discrete time indices $n = 0, ..., N-1$. The desired signal and noise are assumed to be zero mean signals and mutually uncorrelated. Further, we assume that the desired signal is quasi periodic which is a reasonable assumption for voiced speech. Often, voiced speech is described by a harmonic model [18], [24], [25], but here we are using a harmonic chirp model which makes the model capable of handling non-stationary speech.

The signal is built up by a set of harmonically related sinusoids as in the normal harmonic model where the sinusoid with the lowest frequency is the fundamental and the other sinusoids have frequencies given by an integer multiple of the fundamental. In the harmonic model, the speech signal is assumed stationary in short segments which is rarely the case. Instead the fundamental frequency is varying slowly over time which can be modelled by using a harmonic linear chirp model. In a linear chirp signal the instantaneous frequency of the l'th harmonic, $\omega_l(n)$, is not stationary but varies linearly with time,

$$\omega_l(n) = l(\omega_0 + kn), \tag{2}$$

where $\omega_0 = f_0/f_s 2\pi$, with f_s the sampling frequency, is the normalised fundamental frequency and k is the fundamental chirp rate. The instantaneous phase, $\theta_l(n)$, of the sinusoids are given by the integral of the instantaneous frequency as

$$\theta_l(n) = l\left(\omega_0 n + \frac{1}{2}kn^2\right) + \phi_l, \tag{3}$$

and, thereby, this leads to the harmonic chirp model for a voiced speech signal, $s(n)$:

$$s(n) = \sum_{l=1}^{L} A_l \cos\left(\theta_l(n)\right) \tag{4}$$

$$= \sum_{l=1}^{L} A_l \cos\left(l\left(\omega_0 n + \frac{k}{2}n^2\right) + \phi_l\right). \tag{5}$$

where L is the number of harmonics, $A_l > 0$ is the amplitude and ϕ_l is the initial phase of the l'th harmonic, respectively. A special case of the harmonic chirp model for $k = 0$ is then the traditional harmonic model:

$$s(n) = \sum_{l=1}^{L} A_l \cos\left(l\omega_0 n + \phi_l\right) \tag{6}$$

In the speech enhancement process later, it is instructive to make the relationship between the time dependent part of the instantaneous phase, $l(\omega_0 n + k/2n^2)$, and the initial phase, ϕ_l multiplicative instead of additive. This either leads to the real

signal model [16]:

$$s(n) = \sum_{l=1}^{L} a\cos\left(l\left(\omega_0 n + \frac{k}{2}n^2\right)\right) - b\sin\left(l\left(\omega_0 n + \frac{k}{2}n^2\right)\right), \quad (7)$$

where $a = A_l\cos(\phi_l)$ and $b = A_l\sin(\phi_l)$, or, by using Eulers formula, to the complex signal model:

$$\begin{aligned} s(n) &= \sum_{l=1}^{L} \alpha_l e^{jl\left(\omega_0 n + k/2n^2\right)} + \alpha_l^* e^{-jl\left(\omega_0 n + k/2n^2\right)} \\ &= \sum_{l=1}^{L} \alpha_l z^l(n) + \alpha_l^* z^{-l}(n), \end{aligned} \quad (8)$$

where

$$z(n) = e^{-j(\omega_0 n + k/2n^2)} \quad (9)$$

and $\alpha_l = \frac{A_l}{2}e^{j\phi}$. Since (7) and (8) are two ways of describing the same signal, it is possible to design optimal filters based on both, but the complex model in (8) gives a more intuitive and simple notation, and, therefore, we will use this model in the following instead of the real model in (7) [16].

Defining a subvector of samples

$$\mathbf{s}(n) = [s(n)\ s(n-1)\ \ldots\ s(n-M+1)]^T \quad (10)$$

where $M \leq N$ and $(\cdot)^T$ denotes the transpose, the signal model can be written as

$$\mathbf{s}(n) = \mathbf{Z}\mathbf{a}, \quad (11)$$

where $\mathbf{Z}$ is a matrix constructed from a set of L modified Fourier vectors matching the harmonics of the signal,

$$\mathbf{Z} = [\mathbf{z}(1)\ \mathbf{z}(-1)\ \mathbf{z}(2)\ \mathbf{z}(-2)\ \ldots\ \mathbf{z}(L)\ \mathbf{z}(-L)], \quad (12)$$

with

$$\mathbf{z}(l) = \begin{bmatrix} e^{-jl(\omega_0 n + k/2n^2)} \\ e^{-jl(\omega_0(n+1) + k/2(n+1)^2)} \\ \vdots \\ e^{-jl(\omega_0(n+M-1) + k/2(n+M-1)^2)} \end{bmatrix} = \begin{bmatrix} z(n)^l \\ z(n+1)^l \\ \vdots \\ z(n+M-1)^l \end{bmatrix}. \quad (13)$$

The vector $\mathbf{a}$ contains the complex amplitudes of the harmonics, $\mathbf{a} = [\alpha_1\ \alpha_1^*\ \alpha_2\ \alpha_2^*\ \ldots\ \alpha_L\ \alpha_L^*]^T$, where $\{\cdot\}^*$ denotes the complex conjugate.

The observed signal vector, $\mathbf{x}(n)$, is then given by

$$\mathbf{x}(n) = \mathbf{s}(n) + \mathbf{v}(n), \quad (14)$$

where $\mathbf{x}(n)$ and $\mathbf{v}(n)$ are defined in a similar way to $\mathbf{s}(n)$ in (10). Due to the assumption of zero mean uncorrelated signals, the variance of the observed signal is given by the sum of the variances of the desired signal and noise, $\sigma_x^2 = \sigma_s^2 + \sigma_v^2$, where the variance of a signal $g(n)$ is defined by $\sigma_g^2 = \mathbb{E}\{g^2(n)\}$ with $\mathbb{E}\{\cdot\}$ denoting statistical expectation. The level of the desired signal relative to the noise in the observed signal is described by the input signal-to-noise ratio (SNR):

$$\text{iSNR} = \frac{\sigma_s^2}{\sigma_v^2}. \quad (15)$$

The objective is then to recover the desired signal in the best possible way from the observed signal. This can be done by filtering $\mathbf{x}(n)$ with a filter $\mathbf{h} = [h(0)\ h(1)\ \ldots\ h(M-1)]^T$, where $M \leq N$ is the filter length and $\{\cdot\}^T$ denotes the transpose. However, because both the observed signal and the filter are real, multiplying the observed signal with the Hermitian transposed, $\{\cdot\}^H$, filter gives the same result as multiplying with the transposed filter. Due to the choice of a complex representation of the real signal, the Hermitian notation is used throughout the paper since this gives more intuitive interpretations of some intermediate variables such as covariance matrices. That is,

$$\hat{s}(n) = \mathbf{h}^H\mathbf{x}(n) = \mathbf{h}^H\mathbf{s}(n) + \mathbf{h}^H\mathbf{v}(n), \quad (16)$$

gives an estimate, $\hat{s}(n)$, of the desired signal, $s(n)$. The variance of the estimate is then $\sigma_{\hat{s}}^2 = \sigma_{x,\text{nr}}^2 = \sigma_{s,\text{nr}}^2 + \sigma_{v,\text{nr}}^2$, where $\sigma_{x,\text{nr}}^2$ is the variance of the observed signal after noise reduction, i.e.,

$$\sigma_{x,\text{nr}}^2 = \mathbb{E}\{(\mathbf{h}^H\mathbf{x}(n))^2\} = \mathbf{h}^H\mathbf{R}_x\mathbf{h}, \quad (17)$$

with $\mathbf{R}_x$ being the covariance matrix of the observed signal defined as:

$$\mathbf{R}_x = \mathbb{E}\{\mathbf{x}(n)\mathbf{x}^H(n)\}. \quad (18)$$

Similar definitions of the variance after noise reduction and the covariance matrix hold for the desired signal and the noise signal. Further, using the signal model in (11), the covariance matrix of the desired signal can be expressed as

$$\begin{aligned} \mathbf{R}_s &= \mathbb{E}\{\mathbf{s}(n)\mathbf{s}^H(n)\} && (19)\\ &= \mathbb{E}\left\{(\mathbf{Z}\mathbf{a})\,(\mathbf{Z}\mathbf{a})^H\right\} && (20)\\ &= \mathbf{Z}\mathbf{P}\mathbf{Z}^H, && (21) \end{aligned}$$

where

$$\mathbf{P} = \mathbb{E}\{\mathbf{a}\mathbf{a}^H\}. \quad (22)$$

Here, $\mathbf{P}$ is the covariance matrix of the amplitudes. If the phases are independent and uniformly distributed, it reduces to a diagonal matrix with the powers of the harmonics on the diagonal.

If $s(n)$ and $v(n)$ are uncorrelated, $\mathbf{R}_x$ is given by the sum of the covariance matrix of the desired signal, $\mathbf{R}_s$, and the covariance matrix of the noise, $\mathbf{R}_v$,

$$\mathbf{R}_x = \mathbf{R}_s + \mathbf{R}_v. \quad (23)$$

Like the input SNR, the output SNR is the ratio of the desired signal to noise but now after noise reduction

$$\begin{aligned} \text{oSNR}(\mathbf{h}) &= \frac{\sigma_{s,\text{nr}}^2}{\sigma_{v,\text{nr}}^2} && (24)\\ &= \frac{\mathbf{h}^H\mathbf{R}_s\mathbf{h}}{\mathbf{h}^H\mathbf{R}_v\mathbf{h}}. && (25) \end{aligned}$$

It is desirable to have as high an output SNR as possible, but if the filter distorts the desired signal along with removing the noise, it might be more beneficial to make a compromise between noise reduction and signal distortion. The signal distortion can be described by the signal reduction factor

which is the ratio between the variance of the desired signal before and after noise reduction:

$$\xi_{\text{sr}}(\mathbf{h}) = \frac{\sigma_s^2}{\sigma_{s,\text{nr}}^2} \tag{26}$$

$$= \frac{\sigma_s^2}{\mathbf{h}^H \mathbf{R}_s \mathbf{h}}. \tag{27}$$

A distortionless filter will give a signal reduction factor of one, even though a filter can introduce distortion in sub-bands and still have a signal reduction factor of one.

III. Filters

A. Traditional filters

A set of different filters can be defined by looking at the error, $e(n)$, between the desired signal, $s(n)$, and the estimate of the desired signal, $\hat{s}(n)$,

$$\begin{aligned} e(n) &= s(n) - \hat{s}(n) = s(n) - \mathbf{h}^H \mathbf{x}(n) \\ &= s(n) - \mathbf{h}^H \mathbf{s}(n) - \mathbf{h}^H \mathbf{v}(n). \end{aligned} \tag{28}$$

From this, the minimum mean squared error (MMSE) criterion can be defined

$$J(\mathbf{h}) = \mathbb{E}\{e(n)^2\} = \mathbb{E}\{(s(n) - \mathbf{h}^H \mathbf{x}(n))^2\} \tag{29}$$

$$= \mathbb{E}\{ \left(s(n) - \mathbf{h}^H \mathbf{s}(n) - \mathbf{h}^H \mathbf{v}(n)\right)^2 \} \tag{30}$$

Minimisation of $J(\mathbf{h})$ leads to the classical Wiener filter [15]:

$$\mathbf{h}_w = \mathbf{R}_x^{-1} \mathbf{R}_s \mathbf{i}_M, \tag{31}$$

where $\mathbf{i}_M$ is the first column of the $M \times M$ identity matrix. Using (23), the Wiener filter can be rewritten as

$$\mathbf{h}_w = \mathbf{R}_x^{-1} (\mathbf{R}_x - \mathbf{R}_v) \mathbf{i}_M, \tag{32}$$

which is often convenient when the covariance matrices are to be estimated.

More flexible filters can be obtained if the error signal, $e(n)$, is seen as composed of two parts, one expressing the signal distortion, $e_s(n)$, the other the amount of residual noise, $e_v(n)$,

$$e_s(n) = s(n) - \mathbf{h}^H \mathbf{s}(n), \tag{33}$$

$$e_v(n) = \mathbf{h}^H \mathbf{v}(n), \tag{34}$$

with the corresponding minimum mean squared errors (MSEs) being

$$J_s(\mathbf{h}) = \mathbb{E}\{e_s(n)^2\} = \mathbb{E}\{(s(n) - \mathbf{h}^H \mathbf{s}(n))^2\} \tag{35}$$

$$J_v(\mathbf{h}) = \mathbb{E}\{e_v(n)^2\} = \mathbb{E}\{(\mathbf{h}^H \mathbf{v}(n))^2\}. \tag{36}$$

These error measures make it possible to, e.g., minimise the noise power output of the filter while constraining the amount of signal distortion the filter introduces [26], i.e.,

$$\min_{\mathbf{h}} J_v(\mathbf{h}) \quad \text{s.t.} \quad J_s(\mathbf{h}) = \beta \sigma_s^2, \tag{37}$$

where β is a tuning parameter. Solving for the filter by use of the Lagrange multiplier λ gives:

$$\mathbf{h}_\lambda = \left(\mathbf{R}_s + \frac{1}{\lambda} \mathbf{R}_v \right)^{-1} \mathbf{R}_s \mathbf{i}_M, \tag{38}$$

where $\lambda > 0$ satisfies $J_s(\mathbf{h}) = \beta \sigma_s^2$. When $\lambda \to \infty$, $\mathbf{h} \to \mathbf{i}_M$ which gives $\beta \to 0$ and $\hat{s}(n) = x(n)$. When $\lambda = 1$ the filter reduces to the Wiener filter and $\lambda \to 0 \Rightarrow \beta \to 1$ which means that the difference in variance between the desired signal and the estimated signal is equal to the variance of the desired signal and so a large amount of signal distortion is introduced.

B. Parametric filters

The filter in (38) has no control over the distortion of the single harmonics in a voiced speech signal. This is, however, possible by minimisation of $J_v(\mathbf{h})$ under the constraint that the desired signal is passed undistorted, i.e.,

$$\min_{\mathbf{h}} J_v(\mathbf{h}) \quad \text{s.t.} \quad s(n) - \mathbf{h}^H \mathbf{s}(n) = 0. \tag{39}$$

Expressing the signal using the harmonic chirp model in (11), the restriction can be rewritten as

$$s(n) - \mathbf{h}^H \mathbf{s}(n) = 0 \Leftrightarrow \tag{40}$$

$$\mathbf{i}_M^T \mathbf{Z} \mathbf{a} - \mathbf{h}^H \mathbf{Z} \mathbf{a} = 0 \Leftrightarrow \tag{41}$$

$$\mathbf{i}_M^T \mathbf{Z} = \mathbf{h}^H \mathbf{Z} \Leftrightarrow \tag{42}$$

$$\mathbf{b}^H = \mathbf{h}^H \mathbf{Z}, \tag{43}$$

where $\mathbf{b}^H = \mathbf{i}_M^T \mathbf{Z}$ is an $1 \times L$ vector containing the constraints of each harmonic. Using the relation in (17), (39) can be rewritten as

$$\min_{\mathbf{h}} \mathbf{h}^H \mathbf{R}_v \mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^H \mathbf{Z} = \mathbf{b}^H \tag{44}$$

where the filter should be longer than the number of constraints, i.e., $M > 2L$ to ensure a nontrivial solution. If the signal is passed through the filter undistorted, the variance of the signal before and after filtering is the same, and the output SNR reduces to

$$\text{oSNR}(\mathbf{h}) = \frac{\sigma_s^2}{\mathbf{h}^H \mathbf{R}_v \mathbf{h}}. \tag{45}$$

Minimising $\mathbf{h}^H \mathbf{R}_v \mathbf{h}$ under the constraint of an undistorted signal will, therefore, lead to a filter that maximises the output SNR under the same constraint.

The solution to (44) is the linearly constrained minimum variance (LCMV) filter and is given by [22]:

$$\mathbf{h}_{\text{LCMV}} = \mathbf{R}_v^{-1} \mathbf{Z} (\mathbf{Z}^H \mathbf{R}_v^{-1} \mathbf{Z})^{-1} \mathbf{b} \tag{46}$$

The filter reduces to the LCMV filter for harmonic signals when $k = 0$. The covariance matrix of the noise signal is not known and has to be estimated. This is not trivial, but in an optimal situation where the signal model fits perfect, the noise covariance matrix can be replaced by the covariance matrix of the observed signal, $\mathbf{R}_x$, [17], which is easier to estimate, i.e.,

$$\mathbf{h}_{\text{LCMV}} = \mathbf{R}_x^{-1} \mathbf{Z} (\mathbf{Z}^H \mathbf{R}_x^{-1} \mathbf{Z})^{-1} \mathbf{b} \tag{47}$$

Another more empirical approach taking its starting point in the MSE is the amplitude and phase estimation (APES) filter [18]. Here, the harmonic chirp model is also assumed and the

expectation is approximated by an average over time, leading to the estimated MSE:

$$J_a(\mathbf{h}) = \frac{1}{N-M+1}\sum_{n=0}^{N-M}|s(n)-\mathbf{h}^H\mathbf{x}(n)|^2, \quad (48)$$

$$= \frac{1}{N-M+1}\sum_{n=0}^{N-M}|\mathbf{a}^H\mathbf{w}(n)-\mathbf{h}^H\mathbf{x}(n)|^2, \quad (49)$$

where

$$\mathbf{w}(n) = \begin{bmatrix} z(n)^1 & z(n)^{-1} & \cdots & z(n)^L & z(n)^{-L}\end{bmatrix}^T. \quad (50)$$

Writing out the terms in the quadratic expression and solving for the amplitudes [18] gives $\widehat{\mathbf{a}} = \mathbf{W}^{-1}\mathbf{G}\mathbf{h}$, and, thereby,

$$J_a(\mathbf{h}) = \mathbf{h}^H\mathbf{R}_x\mathbf{h} - \mathbf{h}^H\mathbf{G}^H\mathbf{W}^{-1}\mathbf{G}\mathbf{h} \quad (51)$$

$$= \mathbf{h}^H(\mathbf{R}_x - \mathbf{G}^H\mathbf{W}^{-1}\mathbf{G})\mathbf{h}, \quad (52)$$

$$= \mathbf{h}^H\mathbf{Q}\mathbf{h} \quad (53)$$

with

$$\mathbf{G} = \frac{1}{N-M+1}\sum_{n=0}^{N-M}\mathbf{w}(n)\mathbf{x}^H(n), \quad (54)$$

$$\mathbf{W} = \frac{1}{N-M+1}\sum_{n=0}^{N-M}\mathbf{w}(n)\mathbf{w}^H(n). \quad (55)$$

and

$$\mathbf{Q} = \mathbf{R}_x - \mathbf{G}^H\mathbf{W}^{-1}\mathbf{G}. \quad (56)$$

As with the LCMV filter, the MSE is minimised with a constraint that the desired signal should be passed undistorted, leading to a similar filter [18]:

$$\mathbf{h}_{\text{APES}} = \mathbf{Q}^{-1}\mathbf{Z}(\mathbf{Z}^H\mathbf{Q}^{-1}\mathbf{Z})^{-1}\mathbf{b} \quad (57)$$

IV. Covariance matrix estimates

The covariance matrices used in the derived filters are not known but have to be estimated. The covariance matrix of the observed signal can, e.g., be estimated by use of the sample covariance matrix estimate [22]:

$$\widehat{\mathbf{R}}_x = \frac{1}{N-M+1}\sum_{n=0}^{N-M}\mathbf{x}(n)\mathbf{x}^H(n). \quad (58)$$

In order to make the estimate nonsingular, it is required that $2M+1 \leq N$. For this to give a good estimate, the signal should be nearly stationary not only in the set of the filtered M samples, but for all N samples. Otherwise, the N samples are not a good representation of the signal within the M samples, and the sample covariance matrix will not be a good estimate of the observed signal covariance matrix. In such a case, the filters in (46) and (47) are not identical, and it is, therefore, necessary to find an estimate of the noise covariance matrix.

Exchanging $\mathbf{x}(n)$ in (54) with $\mathbf{Z}\mathbf{a}+\mathbf{v}(n)$, it can be shown that the term $\mathbf{G}^H\mathbf{W}^{-1}\mathbf{G}$ in (56) reduces to $\mathbf{Z}\mathbf{P}\mathbf{Z}^H$ for large sample sizes. This means that $\mathbf{G}^H\mathbf{W}^{-1}\mathbf{G}$ can be seen as an estimate of the covariance matrix of the desired signal, and, therefore, $\mathbf{Q}$ is an estimate of the noise covariance matrix. The APES filter is, therefore, an estimate of the optimal LCMV filter. These covariance matrix estimates are an implicit feature of the APES minimisation.

The APES based noise covariance matrix estimate is obtained using a signal driven approach. Alternatively, we suggest taking a noise driven approach and estimate the noise covariance matrix based on noise PSDs. This can be advantageous since several methods exist for estimating the noise power spectral density in the frequency domain, e.g., based on minimum statistics [7] or MMSE [8]. The power spectral density of a signal $g(n)$, $S_g(\omega)$, is related to the autocorrelation, $R_g(\tau)$, and, thereby, also to the covariance matrix of a signal through the Fourier transform [27]

$$R_g(\tau) = \int_{-\infty}^{\infty} S_g(\omega)e^{j\omega\tau}d\omega, \quad (59)$$

where τ denotes a time lag. The autocorrelation is also defined as

$$R_g(\tau) = \mathbb{E}\{g(n)g(n-\tau)\}. \quad (60)$$

In order to get a good approximation to the expectation by taking the mean over the samples and to make the covariance matrix full rank, the same restriction on M relative to N applies here, $2M+1 \leq N$.

The noise covariance matrix is then estimated as:

$$\mathbf{R}_v(p,q) = \begin{cases} R_v(q-p) & \text{for } q \geq p \\ R_v(N+q-p) & \text{for } q < p \end{cases} \quad (61)$$

for p and $q \in$ [1,M].

V. Performance of parametric filters

The theoretical performance of the LCMV filter in (46) can be found by inserting the expression for the filter in (25) and (27). Moreover, the expression for the covariance matrix of the desired signal introduced in (21) is used. The output power of the desired signal can be expressed as:

$$\mathbf{h}^H\mathbf{R}_s\mathbf{h} = \mathbf{b}^H(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z}\mathbf{P}\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z}(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})^{-1}\mathbf{b} = \mathbf{b}^H\mathbf{P}\mathbf{b} = \mathbf{1}^T\mathbf{P}\mathbf{1} = \sigma_s^2, \quad (62)$$

where $\mathbf{1}$ is a an $L\times 1$ vector of ones. The second last equality sign follows from the facts that $\mathbf{b}$ contains only unit amplitude exponential functions and that $\mathbf{P}$ is a diagonal matrix. The output power of the noise is:

$$\mathbf{h}^H\mathbf{R}_v\mathbf{h} = \mathbf{b}^H(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{R}_v\mathbf{R}_v^{-1}\mathbf{Z}(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})^{-1}\mathbf{b} = \mathbf{b}^H(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})^{-1}\mathbf{b}. \quad (63)$$

The output SNR and signal reduction factor then becomes:

$$\text{oSNR}(\mathbf{h}) = \frac{\sigma_s^2}{\mathbf{b}^H(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})^{-1}\mathbf{b}}, \quad (64)$$

and

$$\xi_{\text{sr}}(\mathbf{h}) = 1. \quad (65)$$

These expressions for output SNR and signal reduction are made under the assumption that the noise statistics and the

parameters of the signal are known, and that the model fits the desired signal perfectly. Looking at the expression for the output power of the desired signal from the filter in (62), it is seen that a distortionless response is dependent on the model of the signal. In order to let the signal pass undistorted through the filter, the model has to fit the signal, and a good estimation of the parameters is needed. The amount of distortion is independent of the noise covariance matrix. The output power of the noise from the filter is, on the other hand, not dependent on the parameters of the model, it is only dependent on a good noise covariance matrix estimate. Using the harmonic chirp model instead of the traditional harmonic model, should for all parametric filters decrease the amount of signal reduction since the model fits the signal better. For the APES filter, a better signal model will also lead to a better noise covariance matrix estimate, and, thereby, influencing both the power output of the desired signal and the noise.

VI. Experiments

The simulations are separated in three parts. In the first part, the filters based on the harmonic chirp model are tested on synthetic signals. This is done to verify that the derived filters work in an expected manner and to compare their performance to filters based on the traditional harmonic model under controlled conditions. In the second part, we turn to simulations on real speech signals to confirm that the harmonic chirp model describes voiced speech better than the traditional harmonic model, and that the harmonic chirp filters perform better than their harmonic counterparts. In the third part, the LCMV and APES filters are compared to the Wiener filter where the LCMV filter is combined with a PSD covariance matrix estimate, and the Wiener filter is combined with both an APES and a PSD covariance matrix estimate.

A. Synthetic signal

1) Setup: The LCMV and APES filters based on the harmonic chirp model were tested on a synthetic chirp signal made according to (5) with the same length as the segment length, N. The signal was generated with $L = 10$, $A_l = 1 \,\forall\, l$, random phase, fundamental frequency, and fundamental chirp rate, in the intervals $\phi_l \in [0, 2\pi]$, $f_0 \in [150, 250]$ Hz, $k \in [0, 200]$ Hz^2, respectively. The signal is sampled at 8 kHz and added to white Gaussian noise with a variance calculated to fit the desired input SNR.

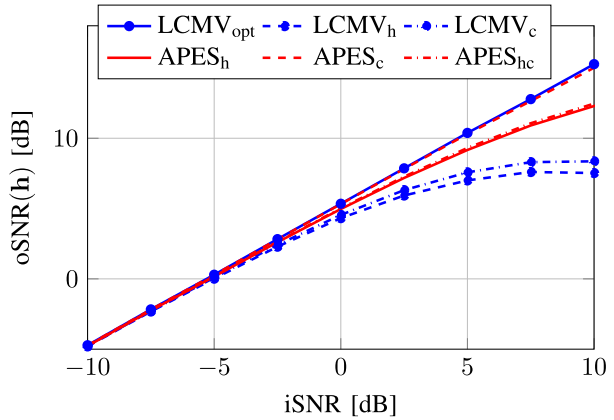
The filters are evaluated as a function of the input SNR, the segment length, N, and the filter length, M. When the parameters are not varied they are set to: iSNR $= 10$ dB, $N = 230$ and $M = 50$. Evaluating M with a fixed N makes it possible to have more elements in the sum in (58) when M is small compared to large, and, thereby, the statistical stability of $\mathbf{R}_x$ would be greater for shorter filters. To avoid this bias and make the conditions as similar as possible for all filter lengths, the same number of elements are used in the sum in (58) independent of the filter length. The fundamental frequency and fundamental chirp rate are assumed known when designing the filters for the synthetic signals. This assumption is made to evaluate the filters without the influence of the performance of a specific parameter estimation method. The results are averaged over 1000 Monte Carlo simulations (MCS). The filters are compared by means of the output SNR in (25) and the signal reduction factor in (27). Using these expressions, the output SNR and signal reduction factor are calculated sample wise based on the N samples used to generate the covariance matrix estimates $\mathbf{R}_s$ and $\mathbf{R}_v$, and afterwards they are averaged over the 1000 MCS.

2) Compared filters: The performance of the chirp based filters is compared to the same filter types based on the harmonic model. This will show whether it is beneficial to expand the traditional harmonic model based on the assumption of stationary speech to a harmonic chirp model where the fundamental frequency is assumed to change linearly within each segment. The LCMV and APES filters derived for the harmonic model can be obtained by setting $k = 0$ in the signal model. A set of six filters are compared in the simulations:

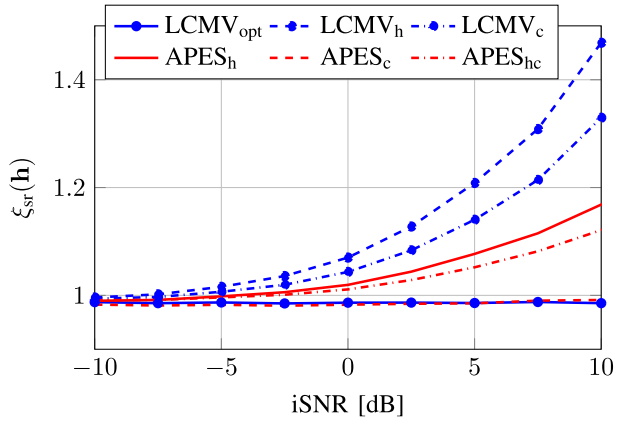
- **LCMV$_{\text{opt}}$**: chirp LCMV filter made according to (46) with $\mathbf{R}_v$ estimated from the clean noise signal. This filter will have the best possible performance a harmonic chirp LCMV filter can have, but can not be made in practice since there is no access to the clean noise signal.
- **LCMV$_{\text{h}}$**: harmonic LCMV filter made according to (47) with $k = 0$.
- **LCMV$_{\text{c}}$**: chirp LCMV filter made according to (47).
- **APES$_{\text{h}}$**: harmonic APES filter made according to (57) with $k = 0$.
- **APES$_{\text{c}}$**: chirp APES filter made according to (57).
- **APES$_{\text{hc}}$**: APES filter made as a combination of the chirp and normal harmonic model with $\mathbf{Z}$ based on the chirp model whereas the estimation of $\mathbf{Q}$ is based on the normal harmonic model. This filter is included to separate the contribution from the modified $\mathbf{Z}$ vector and the modified $\mathbf{Q}$ matrix.

3) Evaluation: The output SNR and signal reduction factor as a function of the input SNR are shown in Fig. 1. At an input SNR of -10 dB all filters perform equally well, but as the input SNR is increased the difference in performance between the filters is increased. As expected, the LCMV$_{\text{opt}}$ sets an upper bound for the performance with a similar gain in SNR at all considered levels of input SNR and no distortion of the desired signal. The harmonic chirp APES based filter, APES$_{\text{c}}$, has similar performance to the optimal LCMV filter. The difference between the two filters, APES$_{\text{h}}$ and APES$_{\text{hc}}$, is only minor. They deviate from the LCMV$_{\text{opt}}$ around 0 dB input SNR and at an input SNR of 10 dB the gain in SNR is around 3 dB less than for the optimal LCMV filter. They also introduce some distorion of the desired signal, with APES$_{\text{h}}$ distorting the desired signal slightly more than APES$_{\text{hc}}$. These two filters have the same noise covariance matrix estimate but different versions of the $\mathbf{Z}$ matrix, as is also the case for the two LCMV filters, LCMV$_{\text{h}}$ and LCMV$_{\text{c}}$, based on the covariance matrix of the observed signal. LCMV$_{\text{h}}$ and LCMV$_{\text{c}}$ have the worst performance of the compared filters, but show the same tendencies as APES$_{\text{h}}$ and APES$_{\text{hc}}$. The difference between the two filters is mainly a smaller signal distortion for the chirp based filter, but here also with a slight difference

in the output SNRs of the two filters. This shows, at least for relatively short filter lengths of $M = 50$, that the major change in performance comes from changing the covariance matrix, from the covariance matrix of the observed signal to the harmonic APES covariance matrix and further again to the harmonic chirp APES covariance matrix. Changing $\mathbf{Z}$ has a minor role but still has an influence, primarily with respect to the distortion of the desired signal.



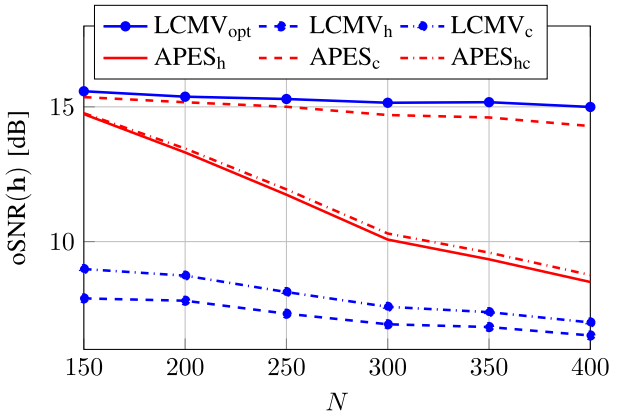
(a) Output SNR



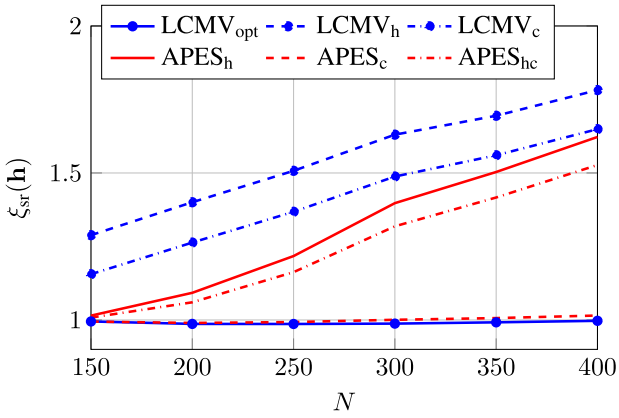
(b) Signal reduction factor

Fig. 1: Output SNR (a) and signal reduction factor (b) as a function of the input SNR for synthetic chirp signals.

The same relationships between the filters can be seen in Fig. 2 where the segment length, N, is varied. The LCMV$_{\text{opt}}$ has the best performance, LCMV$_{\text{c}}$ almost as good, LCMV$_{\text{h}}$ and LCMV$_{\text{c}}$ have the worst performances and APES$_{\text{h}}$ and APES$_{\text{hc}}$ have performances in between. The filters being most influenced by the change in segment length are APES$_{\text{h}}$ and APES$_{\text{hc}}$. They have a drop in output SNR of around 6 dB when the segment length is increased from 150 to 400 whereas the LCMV filters and the chirp APES based filter only give rise to a decrease in output SNR of 1 to 2 dB. Looking at the signal reduction factor, again the chirp APES based filter and the optimal LCMV filter have more or less no distortion of the desired signal whereas the other filters distort the signal more and more when N is increased.



(a) Output SNR

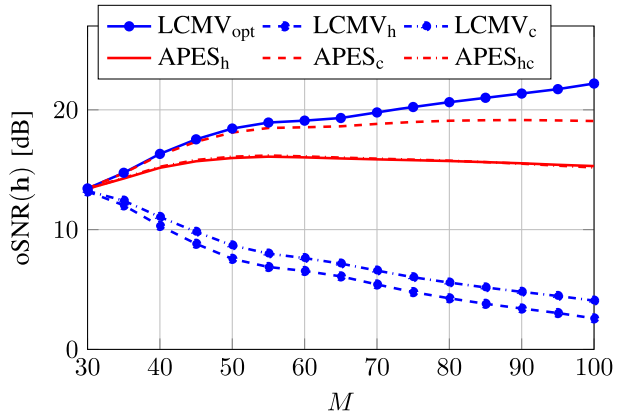


(b) Signal reduction factor

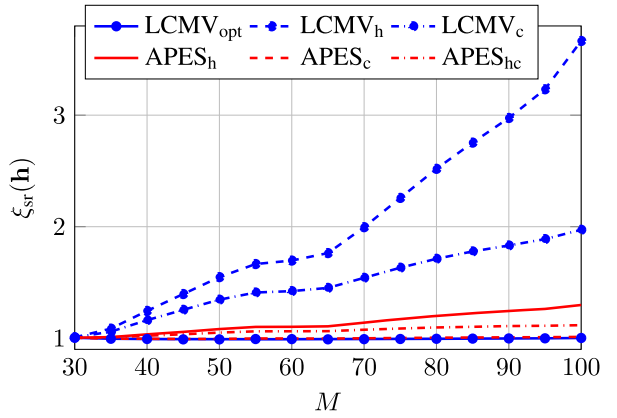
Fig. 2: Output SNR (a) and signal reduction factor (b) as a function of the number of samples N for synthetic chirp signals.

The filter length, M, is varied in Fig. 3. Also here, the difference between the filters increases with increasing filter length. Again, the optimal LCMV filter and the harmonic chirp APES based filter perform best whereas the other filters have a lower output SNR and more signal distortion. However, here the output SNR for APES$_{\text{c}}$ starts to deviate from LCMV$_{\text{opt}}$ for filter lengths above approximately 60.

As an example of the filtering, a signal with a length of 500 samples is generated. The fundamental frequency is set to $f_0 = 200$ Hz, the chirp rate to $k = 200\,\text{Hz}^2$, the initial phases are again random and the sampling rate is $f_s = 8$ kHz. The covariance matrices are based on $N = 230$ samples and the filter length is $M = 50$. The fundamental frequency and chirp rate are also here assumed known. The signal is added to white Gaussian noise to give an input SNR of 10 dB. The used filters are the APES$_{\text{h}}$ giving the estimated signal $\widehat{s}_{\text{h}}$ and APES$_{\text{c}}$ giving the signal $\widehat{s}_{\text{c}}$ since these two filters showed the best performance in the previous experiments. The estimates are compared to the clean signal and the noisy signal in Fig. 4. It is seen in the figure that the chirp filter gives a better estimate of the clean signal than the traditional harmonic filter, and the estimate is also closer to the clean signal than the noisy one



(a) Output SNR



(b) Signal reduction factor

Fig. 3: Output SNR (a) and signal reduction factor (b) as a function of the filter length M for synthetic chirp signals.

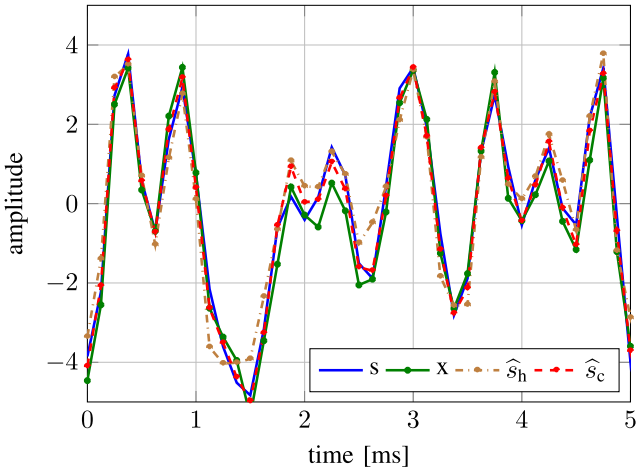


Fig. 4: Reconstructed signal using APES$_h$ and APES$_c$ filters compared to the clean and noisy signals. The noise is white Gaussian and the input SNR is 10 dB.

is.

B. Speech signals

1) Setup: The speech signals used are the 30 sentences included in the NOIZEUS database [28]. Three male and three female speakers produced the 30 Harvard sentences contained in the database. The signals are sampled at 8 kHz and corrupted with noise from the AURORA database [29]. In the first part of this evaluation of speech signals, where the chirp model is compared to the harmonic model, the parameters of the speech signals are estimated from the clean signals. This is done to be able to compare the results for speech signals with the simulations on synthetic data where the parameters were assumed known. In the second part, where the LCMV and Wiener filters are compared, results based on parameters estimated from the noisy signals are shown. The model order and a preliminary fundamental frequency are estimated for every 50 samples using a nonlinear least squares (NLS) estimator [22] with the lower and upper limit for the fundamental frequency given by 80 Hz and 400 Hz, respectively. This is followed by a smoothing [30] and joint estimation of the fundamental frequency and chirp parameter for each sample using the iterative NLS estimator described in [20]. Since the filters are independent of one another, and the fundamental frequency and chirp rate are estimated with reference to the sample being estimated, $\mathbf{Z}$ is also defined with reference to this sample, i.e., the time index in $\mathbf{Z}$ is going from 0 to $M+1$ in each filter. The filter length is increased to $M = 70$ because the real speech signals in many frames have more harmonics than the 10 used to create the synthetic signals. Therefore, a filter with more degrees of freedom is preferred. A good compromise between filter length and segment length for the LCMV and APES filters would according to [31] be $N = 4M$, but this would lead to quite long segments with the given filter length and, as a compromise, the segment length is again set to $N = 230$. The voiced periods are picked out using a generalised likelihood ratio test [32], [33]. Alternatively, the MAP criteria [22] or other voiced/unvoiced detectors can be used [13], [14]. In some cases where unvoiced speech is mistakenly assigned as voiced, the filters become numerically unstable, and these samples are, therefore, excluded from the evaluation. If the filter is not unstable, the unvoiced speech assigned as voiced is processed as if it was voiced speech. This is expected to give a slight decrease in the performance since it is not possible to obtain noise reduction without signal distortion when using the harmonic model in periods of unvoiced speech. In the first part, where the LCMV filters are compared, white Gaussian noise is used and the output SNR and signal reduction factor are calculated using (25) and (27) to facilitate the comparison with the results for the synthetic signal. As was the case for the synthetic signals, the performance measures are calculated sample wise and afterwards averaged over the entire speech signal and the NOIZEUS speech corpus. When the LCMV and APES filters are compared to the Wiener filter, babble noise is used, where the noisy signals are taken from the NOIZEUS speech corpus. The noise levels in the NOIZEUS speech corpus range from 0 dB to 15 dB. The babble noise is chosen because it is one of the most difficult noise types

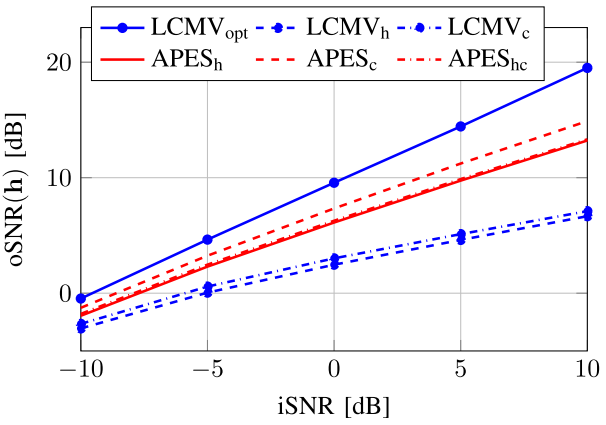
to remove. Results are shown both when the parameters are estimated from the clean signal and when the parameters are estimated from the noisy signals. Since the filters are made based on different ways to estimate the covariance matrices the filters are here compared by means of the output SNR in (24) and the signal reduction factor in (26). Before calculating the variance, the voiced speech parts have been concatenated. This way there will only be one value of the output SNR and signal reduction factor per speech signal which is then averaged over the speech corpus.

2) Compared filters: In the first part of the simulations with real speech, the same filters used for the synthetic signals are compared. In the second part, the LCMV and APES based filters are compared to the Wiener filter. This is done for two different choices of covariance matrices, the first one using the APES derivation, the other using (61) based on the MMSE criterion [8] for finding the PSD. Filters based on the PSD using MMSE and minimum statistics perform almost equally well, and, therefore, only one type of these filters is shown. Further, flexible Wiener filters with two different values of λ are included in the comparisons, leading to six filters:

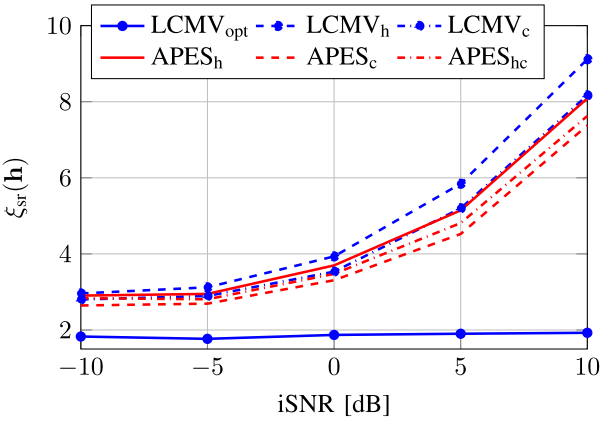
- $\mathbf{APES_c}$: chirp APES filter made according to (57).
- $\mathbf{LCMV_{MMSE}}$: chirp LCMV filter made according to (46) with $\mathbf{R}_v$ estimated from (61) using MMSE.
- $\mathbf{W_c}$: Wiener filter made according to (31) with $\mathbf{R}_s$ estimated using the APES principle as $\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G}$.
- $\mathbf{W_{MMSE}}$: Wiener filter made according to (32) with $\mathbf{R}_v$ estimated from (61) using MMSE.
- $\mathbf{W}_{\lambda=0.2}$: Trade-off Wiener filter made according to (38) with $\lambda = 0.2$ and $\mathbf{R}_s$ estimated using the APES principle as $\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G}$.
- $\mathbf{W}_{\lambda=5}$: Trade-off Wiener filter made according to (38) with $\lambda = 5$ and $\mathbf{R}_s$ estimated using the APES principle as $\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G}$.

Note that all filters except $\mathbf{W_{MMSE}}$ are in some way based on the harmonic chirp model. The $\mathbf{APES_c}$ through both the modified Fourier vector $\mathbf{Z}$ and the covariance matrix estimate $\mathbf{Q}$. The LCMV through $\mathbf{Z}$ and the three Wiener filters $\mathrm{W_c}$, $\mathrm{W}_{\lambda=0.2}$ and $\mathrm{W}_{\lambda=5}$ through the used covariance matrix.

3) Evaluation: In Fig. 5, the output SNR and signal reduction factor are shown as a function of the input SNR. The output SNR and signal reduction factor are calculated using (25) and (27) as was also the case for the synthetic signals. It is seen that the tendencies are the same as for the synthetic signal. $\mathrm{APES_c}$ does not follow the optimal LCMV filter as closely as it did for the synthetic signal, but this is not surprising since the synthetic signals were made according to the harmonic chirp model, and the parameters were assumed known. For the speech signals, the parameters are estimated, and the model does not fit perfectly since the fundamental frequency will not be completely linear in any considered piece within a speech signal. Even though the performance of the $\mathrm{APES_c}$ filter deviates more from the optimal LCMV filter than it did considering synthetic signals, it still has a better performance than the other considered filters. This means that the harmonic chirp model is better at describing the voiced parts of a speech signal and increased performance can be obtained by replacing the traditional harmonic filters with chirp filters.



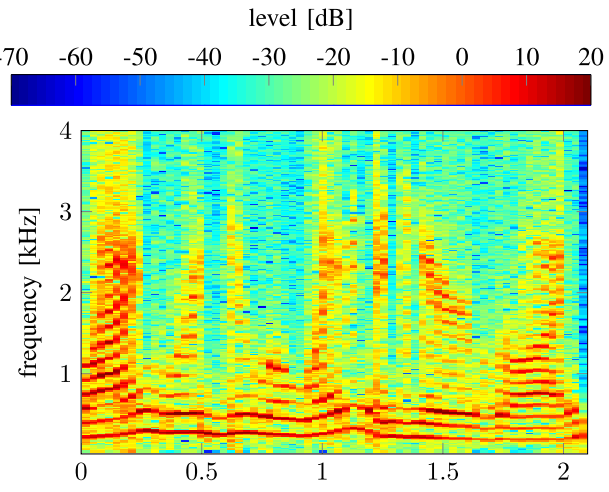
(a) Output SNR



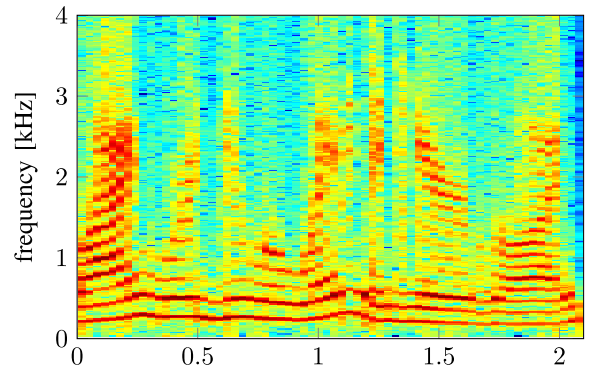
(b) Signal reduction factor

Fig. 5: Output SNR (a) and signal reduction factor (b) as a function of the input SNR, average over NOIZEUS speech corpus added white noise. Parameters estimated from clean speech signals.

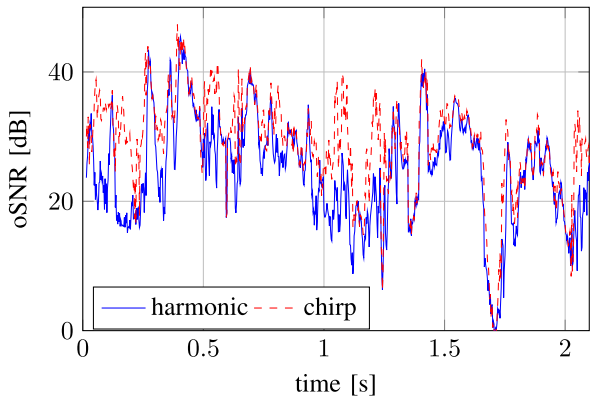
As an example, the speech signal 'Why were you away a year, Roy?' uttered by a female speaker is filtered. The signal has the advantage that it only contains voiced speech, and the entire signal can, therefore, be filtered by the proposed methods. The signal is sampled at 8 kHz, the segment length is 230, the filter length is 70, and the parameters are estimated in the same way as the previous speech signals. The noise is white Gaussian and the input SNR is 10 dB. The spectrograms of the filtered speech signal using $\mathrm{APES_h}$ and $\mathrm{APES_c}$ are shown in Fig. 6 together with the output SNR over time. It is seen that the output SNR of the chirp filter is larger or equal to the output SNR of the harmonic filter. The difference is most pronounced in the first 0.25 seconds and between 1 and 1.25 seconds where the fundamental frequency is changing the most. Here, it is also seen in the spectrograms that the harmonics look slightly cleaner when the chirp filter is used. The Perceptual Evaluation of Speech Quality (PESQ) score [34] for the speech filtered with the harmonic filter is 2.21



(a) Harmonic



(b) Chirp



(c) Output SNR over time

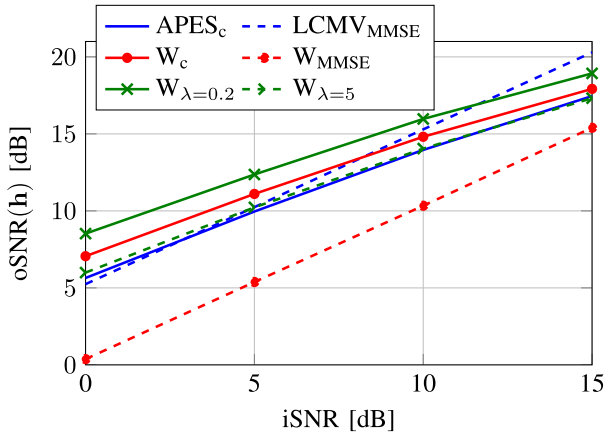
Fig. 6: Spectrograms of speech signal after filtering with (a) traditional harmonic filter and (b) harmonic chirp filter. In (c) the output SNR over time is shown. The input SNR is 10 dB and the noise is white Gaussian. The clean signal can be seen in Fig. 10.

whereas the chirp filter gives a PESQ score of 2.32 and the noisy signal gives a PESQ score of 1.57. The speech signals related to this comparison and the comparison in Fig. 10 can be found at http://www.create.aau.dk/smn.

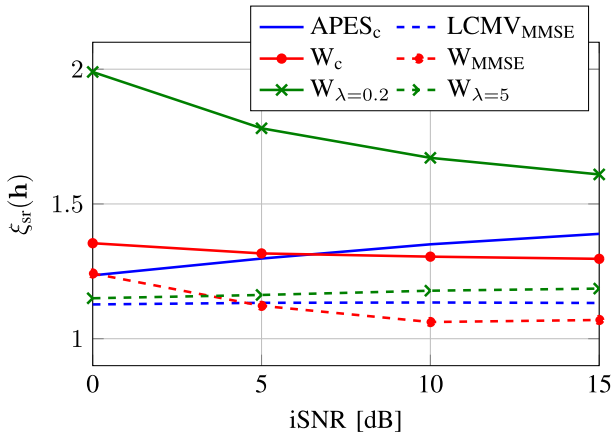
The increased performance of the harmonic chirp filters relative to the harmonic filters should of course be viewed in light of an increased computational complexity since the joint estimation of the fundamental frequency and chirp rate is based on a search in a two-dimensional space. However, [20] describes how to find the parameters iteratively which will decrease the complexity relative to a two-dimensional grid search, and the initial fundamental frequency estimate used in the algorithm is only estimated for every 50 samples in this work which seems to be sufficient for giving good estimates.

Now we turn to alternative combinations of filters and covariance matrices. Here, the output SNR and signal reduction factor are calculated according to (24) and (26). This ensures that no filter is favoured in the way the performance is calculated since the covariance matrices based on the sample covariance principle and the PSD are made in two fundamentally different ways. In Fig. 7a it is seen that five of the six filters work very similar. The Wiener filter in combination with the PSD noise covariance matrix perform significantly worse than the rest when it comes to output SNR. However, the PSD covariance matrix works quite well in combination with the LCMV filter. This filter is one of the better filters at higher input SNRs with respect to output SNR, and it has a low level of distortion at all input SNRs as is seen in Fig. 7b. This can probably be explained by looking at the filters in (31) and (46). The Wiener filter is dependent on two covariance matrices, and the relative levels of these two matrices are, therefore, important for the look of the filter. The LCMV based filters are only dependent on one covariance matrix, and in some way the denominator of the LCMV can be seen as a normalisation which makes the filter independent of the absolute size of the covariance matrix used. The trade-off Wiener filter with $\lambda = 0.2$ gives a higher output SNR than the Wiener filter but at the same time it also gives rise to a higher signal distortion. The flexible Wiener filter with $\lambda = 5.0$ works in the opposite way. It gives a lower output SNR, but also a lower degree of signal distortion. In Fig. 8, the parameters are estimated from the noisy signals whereas the voiced/unvoiced detection is based on the clean signal. The output SNR for the signal dependent filters is decreased a few dBs at low input SNRs whereas it is very similar at high input SNRs. This makes sense since the estimation of parameters is more difficult at low SNRs than at high SNRs. The Wiener filter dependent on the PSD has the same performance in the two situations. In Fig. 9, also the voiced/unvoiced detection is made based on the noisy signal. The overall performance of all filters is slightly decreased compared to making the detection based on the clean signal, but the tendency between the filters is very similar. This suggests that more unvoiced periods are assigned as voiced speech where the voiced signal model will not apply, and thus the performance will decrease slightly.

As an example, the speech signal 'Why were you away a year, Roy?' is again filtered, now in the presence of babble noise at an input SNR of 10 dB. The filters used for this comparison are the APES_c, $\text{LCMV}_{\text{MMSE}}$, W_c and W_{MMSE} and the spectrograms of the resulting signals are shown in Fig. 10 along with spectrograms of the clean and the noisy signal. From this figure, it seems like the Wiener filter in combination with the APES covariance matrix removes the

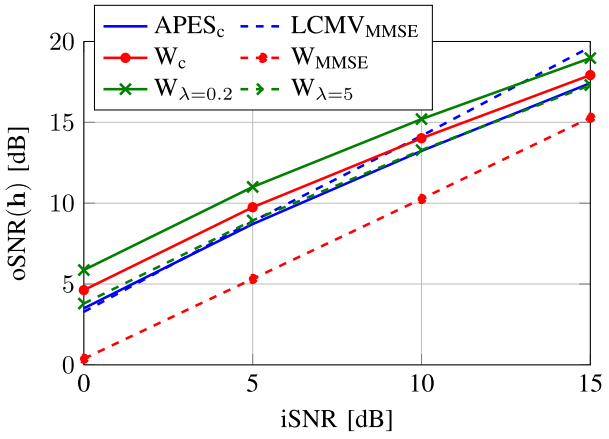


(a) Output SNR

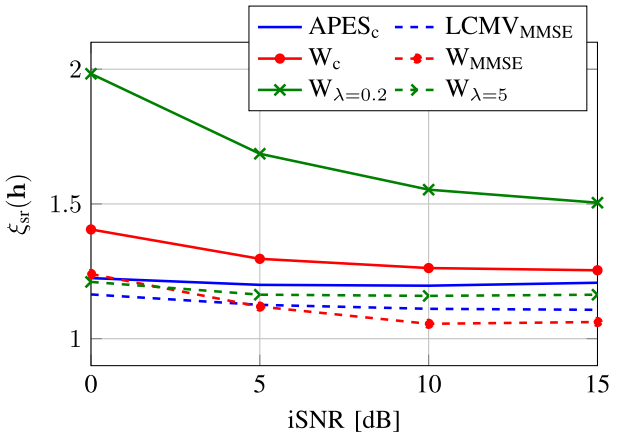


(b) Signal reduction factor

Fig. 7: Output SNR (a) and signal reduction factor (b) as a function of the input SNR, averaged over NOIZEUS corpus with babble noise. Parameters estimated from clean speech signals. Voiced/unvoiced detection based on clean signal.



(a) Output SNR



(b) Signal reduction factor

Fig. 8: Output SNR (a) and signal reduction factor (b) as a function of the input SNR, averaged over NOIZEUS corpus with babble noise. Parameters estimated from noisy speech signals. Voiced/unvoiced detection based on clean signal.

most noise between the harmonics whereas the APES filter and the LCMV filter remove the noise slightly less, both between the harmonics and outside the range of the speech signal. The Wiener filter in combination with the PSD noise covariance matrix seems to perform no noise reduction and the harmonics are even more difficult to distinguish than in the noisy signal. These observations are in line with the curves of output SNR when looking at an input SNR of 10 dB where the $\mathrm{W}_{\mathrm{MMSE}}$ performs worse than the noisy signal, the $\mathrm{APES}_{\mathrm{c}}$ and $\mathrm{LCMV}_{\mathrm{MMSE}}$ perform almost equally well and the W_{c} performs the best. The PESQ scores for the four filtered signals are, $\mathrm{APES}_{\mathrm{c}}$: 2.09, $\mathrm{LCMV}_{\mathrm{MMSE}}$: 2.25, W_{c}: 2.18 and $\mathrm{W}_{\mathrm{MMSE}}$: 1.54. It is interesting to see that the $\mathrm{LCMV}_{\mathrm{MMSE}}$ gives rise to the highest PESQ score since this was not clear from the spectrograms, but this filter gives a lower signal reduction factor than the $\mathrm{APES}_{\mathrm{c}}$ and W_{c} filters, and, therefore, it makes good sense. The noisy signal has a PESQ score of 2.06. Comparing to the signals in white Gaussian noise in Fig. 6, the PESQ score of the filtered signals decreased whereas the PESQ score of the noisy signal increased. This difference is mainly due to the different noise types while the fact that the parameters in Fig. 6 were estimated from the clean signal only contributes slightly. Since babble noise is noise made up from several speakers speaking at the same time, it is distributed in the same frequency range as the speech signal. This makes it more difficult to estimate the relevant parameters and also more difficult to filter out the noise afterwards. However, prewhitening of the noisy signal can help mediate this problem [35] with the noise statistics found using one of the methods in [36].

VII. Conclusion

In this paper, the non-stationarity of voiced speech is taken into account in speech enhancement. This is done by describing the speech by a harmonic chirp model instead of the traditional harmonic model. The chirp used is a linear chirp which allows the fundamental frequency to vary linearly within each segment, and, therefore, the speech signal is not

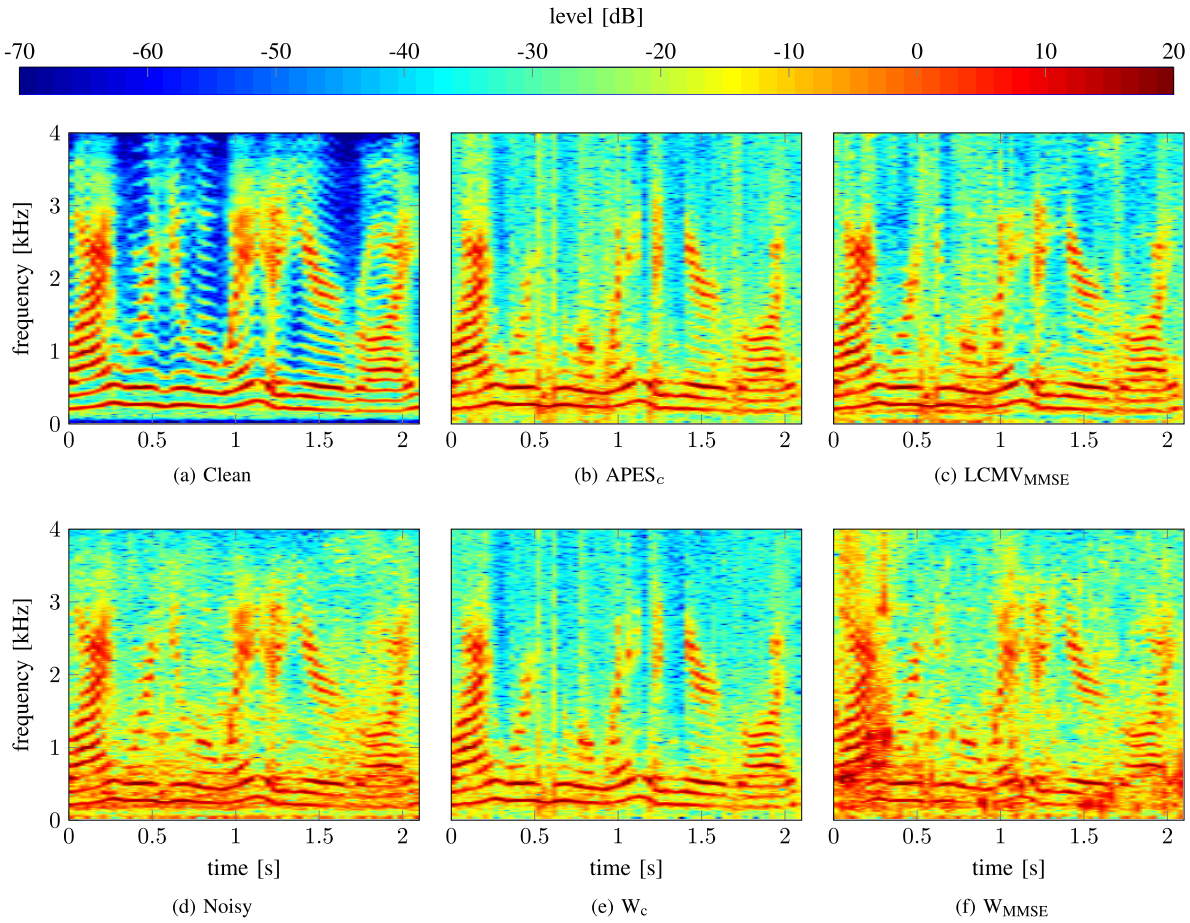


Fig. 10: Spectrograms of clean, noisy and filtered speech. Babble noise is mixed with speech at an input SNR of 10 dB.

assumed stationary within a segment. Versions of the linearly constraint minimum variance (LCMV) filter and amplitude and phase estimation (APES) filter are derived in the framework of harmonic chirp signals. As an implicit part of the APES filter, a noise covariance matrix estimate is derived. This makes the APES filter an estimate of the optimal LCMV filter which maximises the output SNR under the constraint that the desired signal is passed undistorted. APES gives a noise covariance matrix estimate which only assumes the noise signal to be stationary in frames of 20-30 ms as opposed to methods based on power spectral densities (PSDs) which primarily update the noise statistics in periods of unvoiced speech. It is shown through simulations on synthetic and speech signals that the chirp filters give rise to a higher output SNR and a lower signal distortion than their harmonic counterparts, and, therefore, the chirp model describes voiced speech better than the traditional harmonic model. We suggest also using the APES noise covariance matrix estimate in other filters as, e.g., the Wiener filter, and we compare it to a noise covariance matrix estimate based on the PSD. The APES noise covariance matrix estimate is shown to work well in combination with the Wiener and trade-off Wiener filters, whereas the PSD based noise covariance matrix estimate works well in combination with the LCMV filter. All chirp based Wiener and LCMV filters outperform the Wiener filter in combination with the PSD noise covariance matrix estimate.

REFERENCES

[1] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Enhancement of non-stationary speech using harmonic chirp filters," in *Proc. Interspeech*, Sep. 2015, accepted for publication.

[2] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.

[3] E. W. Healy, S. E. Yoho, J. Chen, Y. Wang, and D. Wang, "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," *J. Acoust. Soc. Am.*, vol. 138, no. 3, pp. 1660–1669, 2015.

[4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[5] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

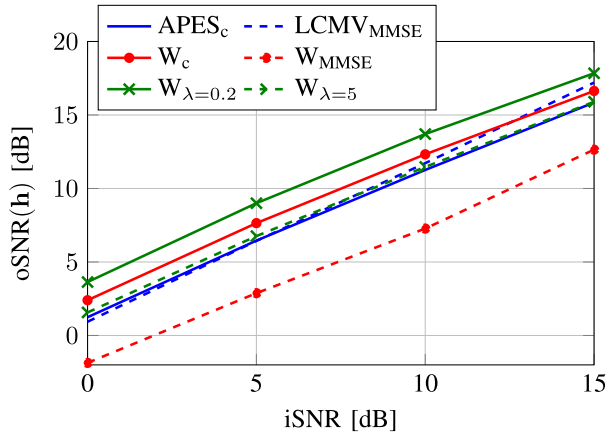
[6] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, Jan. 1999.

[7] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

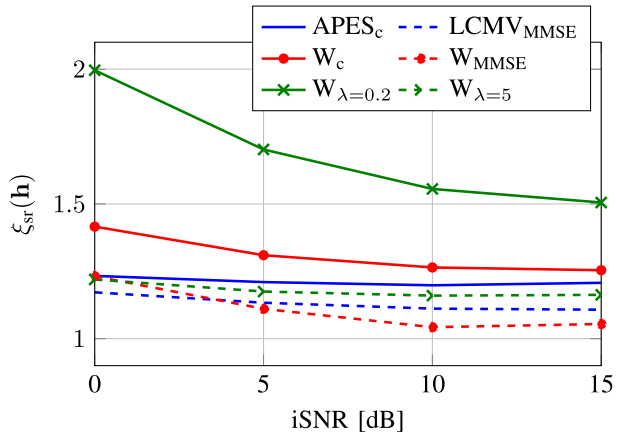
[8] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[9] F. R. Drepper, "A two-level drive-response model of non-stationary speech signals," *Nonlinear Analyses and Algorithms for Speech Processing*, vol. 1, pp. 125–138, Apr. 2005.

[10] L. Deng and D. O'Shaughnessy, *Speech processing: a dynamic and optimization-oriented approach*. CRC Press, 2003.



(a) Output SNR



(b) Signal reduction factor

Fig. 9: Output SNR (a) and signal reduction factor (b) as a function of the input SNR, averaged over NOIZEUS corpus with babble noise. Parameters estimated from noisy speech signals. Voiced/unvoiced detection based on noisy signal.

[11] M. Képesi and L. Weruaga, "Adaptive chirp-based time–frequency analysis of speech signals," *Speech Communication*, vol. 48, no. 5, pp. 474–492, 2006.

[12] L. Weruaga and M. Képesi, "The fan-chirp transform for non-stationary harmonic signals," *Signal Processing*, vol. 87, no. 6, pp. 1504–1522, 2007.

[13] K. I. Molla, K. Hirose, N. Minematsu, and K. Hasan, "Voiced/unvoiced detection of speech signals using empirical mode decomposition model," in *Int. Conf. Information and Communication Technology*, Mar. 2007, pp. 311–314.

[14] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 2, pp. 250–255, 1993.

[15] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1218–1234, 2006.

[16] A. Jakobsson, T. Ekman, and P. Stoica, "Capon and APES spectrum estimation for real-valued signals," *Eighth IEEE Digital Signal Processing Workshop*, 1998.

[17] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 7, pp. 1948–1963, Sep. 2012.

[18] M. G. Christensen and A. Jakobsson, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

[19] Y. Pantazis, O. Rosec, and Y. Stylianou, "Chirp rate estimation of speech based on a time-varying quasi-harmonic model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2009, pp. 3985–3988.

[20] M. G. Christensen and J. R. Jensen, "Pitch estimation for non-stationary speech," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, Nov. 2014, pp. 1400–1404.

[21] Y. Doweck, A. Amar, and I. Cohen, "Joint model order selection and parameter estimation of chirps with harmonic components," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1765–1778, Apr. 2015.

[22] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[23] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Education, Inc., 2005.

[24] D. Ealey, H. Kelleher, and D. Pearce, "Harmonic tunnelling: tracking non-stationary noises during speech," in *Proc. Eurospeech*, Sep. 2001, pp. 437–440.

[25] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1931–1940, 2014.

[26] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Springer-Verlag, 2009.

[27] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications*. Prentice Hall, Inc., 1996.

[28] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7–8, pp. 588 – 601, 2007.

[29] D. Pearce and H. G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. Int. Conf. Spoken Language Process.*, Oct 2000.

[30] H. Ney, "A dynamic programming algorithm for nonlinear smoothing," *Signal Processing*, vol. 5, no. 2, pp. 163–173, 1983.

[31] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: survey, new results, and an application," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 338–352, Feb. 2000.

[32] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice Hall, Inc., 1998.

[33] E. Fisher, J. Tabrikian, and S. Dubnov, "Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 2, pp. 502–510, 2006.

[34] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[35] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP J. on Advances in Signal Processing*, vol. 2007, no. 1, p. 24, Jun. 2007.

[36] P. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

Sidsel Marie Nørholm received her B.Sc. in electronics in 2007 and her M.Sc. in acoustics in 2012 both from the Technical University of Denmark. Currently, she is pursuing a Ph.D. in speech signal processing at the Audio Analysis Lab in the Department of Architecture, Design & Media Technology at Aalborg University in Denmark. She has been a visiting researcher at the University of Oldenburg, Germany.

Her research interests include signal processing methods, especially speech models, speech enhancement and parameter estimation.

Jesper Rindom Jensen (S'09–M'12) was born in Ringkøbing, Denmark in August 1984. He received the M.Sc. degree *cum laude* for completing the elite candidate education in 2009 from Aalborg University in Denmark. In 2012, he received the Ph.D. degree from Aalborg University. Currently, he is a Postdoctoral Researcher at the Department of Architecture, Design & Media Technology at Aalborg University in Denmark, where he is also a member of the Audio Analysis Lab. He has been a Visiting Researcher at the University of Quebec, INRS-EMT, in Montreal, Quebec, Canada, and at the Friedrich-Alexander Universität Erlangen-Nürnberg in Erlangen, Germany.

His research interests include signal processing theory and methods for, e.g., microphone array and joint audio-visual signal processing. Examples of more specific research interests within this scope are enhancement, separation, localization, tracking, parametric analysis, and modeling. He has published nearly more than 50 papers on these topics in top-tier, peer-reviewed conference proceedings and journals. Moreover, he has published 2 research monographs including the book "Speech Enhancement - A Signal Subspace Perspective" which is co-authored with Prof. Jacob Benesty, Prof. Mads Græsbøll Christensen, and Prof. Jingdong Chen.

He has received a highly competitive postdoc grant from the Danish Independent Research Council, as well as several travel grants from private foundations. Furthermore, he is an affiliate member of the IEEE Signal Processing Theory and Methods Technical Committee, and is Member of the IEEE.

Mads Græsbøll Christensen (S'00–M'05–SM'11)) received the M.Sc. and Ph.D. degrees in 2002 and 2005, respectively, from Aalborg University (AAU) in Denmark, where he is also currently employed at the Dept. of Architecture, Design & Media Technology as Professor in Audio Processing and is head and founder of the Audio Analysis Lab.

He was formerly with the Dept. of Electronic Systems at AAU and has been held visiting positions at Philips Research Labs, ENST, UCSB, and Columbia University. He has published 3 books and more than 150 papers in peer-reviewed conference proceedings and journals, and he has given tutorials at EUSIPCO and INTERSPEECH. His research interests include signal processing theory and methods with application to speech and audio, in particular parametric analysis, modeling, enhancement, separation, and coding.

Prof. Christensen has received several awards, including an ICASSP Student Paper Contest Award, the Spar Nord Foundations Research Prize, a Danish Independent Research Council Young Researchers Award, and the Statoil Prize, and he is also co-author of the paper *Sparse Linear Prediction and Its Application to Speech Processing* that received an IEEE Signal Processing Society Young Author Best Paper Award. Moreover, he is a beneficiary of major grants from the Danish Independent Research Council, the Villum Foundation, and Innovation Fund Denmark. He is an Associate Editor for IEEE/ACM Trans. on Audio, Speech, and Language Processing, a former Associate Editor of IEEE Signal Processing Letters, a member of the IEEE Audio and Acoustic Signal Processing Technical Committee, and a Senior Member of the IEEE.

Paper F

Sparse Linear Prediction and Its Applications to Speech Processing

D. Giacobello, **M. G. Christensen**, M. N. Murthi, S. H. Jensen, and M. Moonen,

The paper has been published in the
IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 5, pp. 1644-–1657, July 2012.

Sparse Linear Prediction and Its Applications to Speech Processing

Daniele Giacobello, *Member, IEEE,* Mads Græsbøll Christensen, *Senior Member, IEEE,*
Manohar N. Murthi, *Member, IEEE,* Søren Holdt Jensen, *Senior Member, IEEE,* and Marc Moonen, *Fellow, IEEE*

***Abstract*—The aim of this paper is to provide an overview of *Sparse Linear Prediction*, a set of speech processing tools created by introducing sparsity constraints into the linear prediction framework. These tools have shown to be effective in several issues related to modeling and coding of speech signals. For speech analysis, we provide predictors that are accurate in modeling the speech production process and overcome problems related to traditional linear prediction. In particular, the predictors obtained offer a more effective decoupling of the vocal tract transfer function and its underlying excitation, making it a very efficient method for the analysis of voiced speech. For speech coding, we provide predictors that shape the residual according to the characteristics of the sparse encoding techniques resulting in more straightforward coding strategies. Furthermore, encouraged by the promising application of compressed sensing in signal compression, we investigate its formulation and application to sparse linear predictive coding. The proposed estimators are all solutions to convex optimization problems, which can be solved efficiently and reliably using, e.g., interior-point methods. Extensive experimental results are provided to support the effectiveness of the proposed methods, showing the improvements over traditional linear prediction in both speech analysis and coding.**

***Index Terms*—Linear prediction, speech analysis, speech coding, sparse representation, 1-norm minimization, compressed sensing.**

I. Introduction

Linear prediction (LP) has been successfully applied in many modern speech processing systems in such diverse applications as coding, analysis, synthesis and recognition (see, e.g., [1]). The speech model used in many of these applications is the source-filter model where the speech signal is generated by passing an excitation through an all-pole filter, the predictor in the feedback loop. Typically, the prediction coefficients are identified such that the 2-norm of the residual, the difference between the observed signal and the predicted signal, is minimized. This works well when the excitation signal is Gaussian and independent and identically distributed (i.i.d.) [2], consistent with the equivalent maximum likelihood approach to determine the coefficients [3]. However, when the excitation signal does not satisfy these assumptions, problems arise [2]. This is the case for voiced speech where the excitation can be considered to be a spiky excitation of a quasi-periodic nature [1]. In this case, the spectral cost function associated with the minimization of the 2-norm of the residual can be shown to suffer from certain well-known problems such as overemphasis on peaks and cancellation of errors [2]. In general, the shortcomings of LP in spectral envelope modeling can be traced back to the 2-norm minimization approach: by minimizing the 2-norm, the LP filter cancels the input voiced speech harmonics causing the envelope to have a sharper contour than desired with poles close to the unit circle. A wealth of methods have been proposed to mitigate these effects. Some of the proposed techniques involve a general rethinking of the spectral modeling problem (see, e.g., [4], [5], [6], and [7]) while others are based on changing the statistical assumptions made on the prediction error in the minimization process (notably [8], [9], and [10]).

The above mentioned deficiencies of the 2-norm minimization in LP modeling have also repercussions in the speech coding scenario. In fact, while the 2-norm criterion is consistent with achieving minimal variance of the residual for efficient coding[1], sparse techniques are employed to encode the residual. Examples of this can be seen since early GSM standards with the introduction of multi-pulse excitation (MPE [12]) and regular-pulse excitation (RPE [13]) methods and, more recently, in sparse algebraic codes in code-excited linear prediction (ACELP [14]). In these cases, the sparsity of the RPE and ACELP excitation was motivated, respectively, by psychoacoustic and by the dimensionality reduction of the excitation vector space. Therefore, a better suited predictor for these two coding schemes, arguably, is not the one that

Daniele Giacobello is with the Office of the CTO, Broadcom Corporation, Irvine, CA 92617, USA (e-mail: giacobello@broadcom.com).

Mads Græsbøll Christensen is with the Department of Architecture, Design, and Media Technology, Aalborg University, 9220 Aalborg, Denmark (email: mgc@imi.aau.dk).

Manohar N. Murthi is with the Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33146, USA (e-mail: mmurthi@miami.edu).

Søren Holdt Jensen ia with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: shj@es.aau.dk).

Marc Moonen is with the Department of Electrical Engineering, Katholieke Universiteit Leuven, 3001 Leuven, Belgium (e-mail: marc.moonen@esat.kuleuven.be).

The work of Daniele Giacobello was supported by the Marie Curie EST-SIGNAL Fellowship (http://est-signal.i3s.unice.fr), contract no. MEST-CT-2005-021175 and was carried out at the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark.

The work of Manohar N. Murthi was supported by the National Science Foundation via awards CCF-0347229 and CNS-0519933.

[1]The fundamental theorem of predictive quantization [11] states that the mean squared reproduction error in predictive encoding is equal to the mean squared quantization error when the residual signal is presented to the quantizer. Therefore, by minimizing the 2-norm of the residual, these variables have a minimal variance whereby the most efficient coding is achieved.

minimizes the 2-norm, but the one that leaves the fewest non-zero pulses in the residual, i.e., the *sparsest residual*. Early contributions (notably [9], [15], and [16]) have followed this line of thought questioning the fundamental validity of the 2-norm criterion with regards to speech coding. Despite this research effort, to the authors' best knowledge, 2-norm minimization is the only criterion used in commercial speech applications.

Traditional usage of LP is confined to modeling only the the spectral envelope capturing the short-term redundancies of speech. Hence, in the case of voiced speech, the predictor does not fully decorrelate the speech signal because of the long-term redundancies of the underlying pitch excitation. This means that the residual will still have pitch pulses present. The usual approach is then to employ a cascaded structure where LP is initially applied to determine the short-term prediction coefficients to model the spectral envelope and, subsequently, a long-term predictor is determined to model the harmonic behavior of the spectrum [1]. Such a structure is arguably suboptimal since it ignores the interaction between the two different stages. Also in this case, while early contributions have outlined gains in performance in jointly estimating the two filters (the work in [17] is perhaps the most successful attempt), the common approach is to distinctly separate the two steps.

The recent developments in the field of sparse signal processing, backed up by significant improvements in convex optimization algorithms (e.g., interior point methods [18][19]), have recently encouraged the authors to explore the concept of sparsity in the LP minimization framework [20]. In particular, while reintroducing well-known methods to seek a short-term predictor that produces a residual that is sparse rather than minimum variance, we have also introduced the idea of employing high order sparse predictors to model the cascade of short-term and long-term predictors, engendering a joint estimation of the two [21]. This preliminary work has led the way for the exploitation of the sparse characteristics of the high order predictor and the residual to define more efficient coding techniques. Specifically, in [22], we have demonstrated that the new model achieves a more parsimonious description of a speech segment with interesting direct applications to low bit-rate speech coding. While in these early works, the 1-norm has been reasonably chosen as a convex approximation of the so-called 0-norm[2], in [23] we have applied the reweighted 1-norm algorithm in order to produce a more focused solution to the original problem that we are trying to solve. In this work, we move forward, introducing the novelty of a compressed sensing formulation [24] in sparse LP, that will not only offer important information on how to retrieve the sparse structure of the residual, but will also help reduce the size of the minimization problem, with a clear impact on the computational complexity.

The contribution of this paper is then twofold. Firstly, we put our earlier contributions in a common framework giving an introductory overview of Sparse Linear Prediction and we also introduce its compressed sensing formulation. Secondly, we provide a detailed experimental analysis of its usefulness in modeling and coding applications transcending the well-known limitations related to traditional LP.

The paper is organized as follows. In Section II, we provide a prologue that defines the mathematical formulations of the proposed sparse linear predictors. In Section III, we define the sparse linear predictors and, in Section IV, we provide their compressed sensing formulations. The results of the experimental evaluation of the analysis properties of the short-term predictors are outlined in Section V, while the experimental results of the coding properties and applications are outlined in Section VI. We provide a discussion on some of the drawbacks of sparse linear prediction in Section VII. Finally, Section VIII concludes our work.

II. Fundamentals of Linear Prediction

We consider the following speech production model, where a sample of speech $x(n)$ is written as a linear combination of K past samples:

$$x(n) = \sum_{k=1}^{K} a_k x(n-k) + r(n), \tag{1}$$

where $\{a_k\}$ are the prediction coefficients and $r(n)$ is the prediction error. In particular, we consider the optimization problem associated with finding the prediction coefficient vector $\mathbf{a} \in \mathbb{R}^K$ from a set of observed real samples $x(n)$ for $n = 1, \ldots, N$ so that the prediction error is minimized [18]. Considering the speech production model for a segment of N speech samples $x(n)$, for $n = 1, \ldots, N$, in matrix form:

$$\mathbf{x} = \mathbf{X}\mathbf{a} + \mathbf{r}, \tag{2}$$

the problem becomes:

$$\mathbf{a} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p + \gamma \|\mathbf{a}\|_k^k, \tag{3}$$

where:

$$\mathbf{x} = \begin{bmatrix} x(N_1) \\ \vdots \\ x(N_2) \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x(N_1-1) & \cdots & x(N_1-K) \\ \vdots & & \vdots \\ x(N_2-1) & \cdots & x(N_2-K) \end{bmatrix}. \tag{4}$$

The p-norm operator $\|\cdot\|_p$ is defined as $\|\mathbf{x}\|_p = (\sum_{n=1}^{N} |x(n)|^p)^{\frac{1}{p}}$. The starting and ending points N_1 and N_2 can be chosen in various ways by assuming $x(n) = 0$ for $n < 1$ and $n > N$. In this paper we will use the most common choice of $N_1 = 1$ and $N_2 = N+K$, which is equivalent, when $p = 2$ and $\gamma = 0$, to the *autocorrelation method* [25]. The introduction of the regularization term γ in (3) can be seen as being related to the prior knowledge of the coefficients vector $\mathbf{a}$, problem (3) then corresponds to the *maximum a posteriori* (MAP) approach for finding $\mathbf{a}$ under the assumptions that $\mathbf{a}$ has a Generalized Gaussian Distribution [26]. In finding a sparse signal representation, there is the somewhat subtle problem of how to measure sparsity. Sparsity is often measured as the cardinality, corresponding to the so-called 0-norm $\|\cdot\|_0$. Our optimization problem (3) would then become:

$$\mathbf{a} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_0 + \gamma \|\mathbf{a}\|_0, \tag{5}$$

[2]The 0-norm is not technically a norm since it violates the triangle inequality.

with the particular case in which we are only considering the sparsity in the residual ($\gamma = 0$):

$$\mathbf{a} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_0. \quad (6)$$

Unfortunately, these are combinatorial problems which generally cannot be solved in polynomial time. Instead of the cardinality measure, we will then use the more tractable 1-norm $\|\cdot\|_1$, which is known throughout the sparse recovery literature (see, e.g., [27]) to perform well as a relaxation of the 0-norm. We will also consider more recent variations of the 1-norm minimization criterion such as the reweighted 1-norm [28] to enhance the sparsity measure and moving the solution closer to the original 0-norm problem (5).

III. Sparse Linear Predictors

In this section, we will define the different sparse linear predictors and show their application in the context of speech processing. In particular, we will introduce the problem of determining a short-term predictor that engenders a sparse residual and the problem of finding a high order sparse predictor that also engenders a sparse residual. Since in Section II, we have introduced the 1-norm minimization as the sparsity measure, here we will also give a brief overview of the reweighted 1-norm algorithm to enhance this sparsity measure, moving closer to the original problem (0-norm minimization).

A. Finding a Sparse Residual

We consider the problem of finding a prediction coefficient vector $\mathbf{a}$ such that the resulting residual is sparse. Having identified the 1-norm as a suitable convex relaxation of the cardinality, the cost function for this problem is a particular case of (3). By setting $p = 1$ and $\gamma = 0$ we obtain the following optimization problem:

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1. \quad (7)$$

This formulation of the LP problem has been considered since the early works on speech analysis [9] [15] [16] and becomes particularly relevant for the analysis of voiced speech. In particular, compared to the traditional 2-norm minimization, the cost function associated with the 1-norm minimization deemphasize the impact of the spiky underlying excitation associated with voiced speech on the solution $\mathbf{a}$. Thus, there is an interesting connection between recovering a sparse residual vector and applying robust statistics methods to find the predictor [8]. An example of the more accurate recovery of the voiced excitation is shown in Figure 1. The effect of putting less emphasis on the outliers of the spiky excitation associated with voiced speech will reflect on the spectral envelope that will avoid the over-emphasis on peaks generated in the effort to cancel the pitch harmonics. An example of this property is shown in Figure 2.

While the 1-norm has been shown to outperform the 2-norm in finding a more proper LP model in speech analysis, in the case of unvoiced speech both approaches seem to provide appropriate models. However, by using the 1-norm minimization, we provide a residual that is sparser. In particular in [29] it is shown that, the residual vector provided by 1-norm minimization will have at least K components equal to zero.

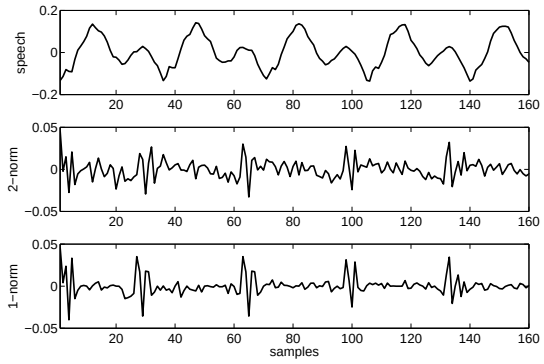


Fig. 1. An example of prediction residuals obtained by 2-norm and 1-norm error minimization. The speech segment analyzed is shown in the top box. The prediction order is $K = 10$ and the frame length is $N = 160$. It can be seen that the spiky pitch excitation is retrieved more accurately when 1-norm minimization is employed.

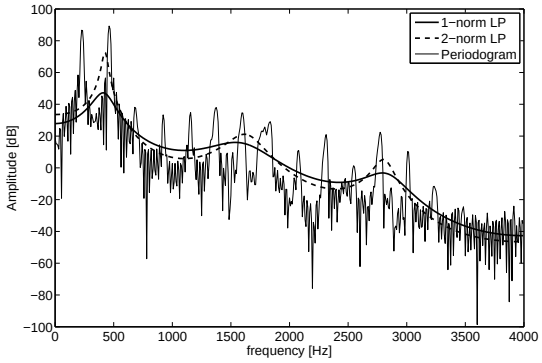


Fig. 2. An example of LP spectral model obtained by 1-norm and 2-norm error minimization for a segment of voiced speech. The prediction order is $K = 10$ and the frame length is $N = 160$. The lower emphasis on peaks in the envelope, when 1-norm minimization is employed, is a direct consequence of the ability to retrieve the spiky pitch excitation.

B. Finding a High Order Sparse Predictor

We now consider the problem of finding a high order sparse predictor that also engenders a sparse residual. This problem is particularly relevant when considering the usual modeling approach adopted in low bit-rate predictive coding for voiced speech segments. This corresponds to a cascade of a short-term linear predictor $F(z)$ and a long-term linear predictor $P(z)$ to remove respectively near-sample redundancies, due to the presence of formants, and distant-sample redundancies, due to the presence of a pitch excitation. The cascade of the predictors corresponds to the multiplication in the z-domain of the their transfer functions:

$$\begin{aligned} A(z) = F(z)P(z) &= 1 - \sum_{k=1}^{K} a_k z^{-k} \\ &= (1 - \sum_{k=1}^{N_f} f_k z^{-k})(1 - \sum_{k=1}^{N_p} g_k z^{-(T_p+k-1)}). \end{aligned} \quad (8)$$

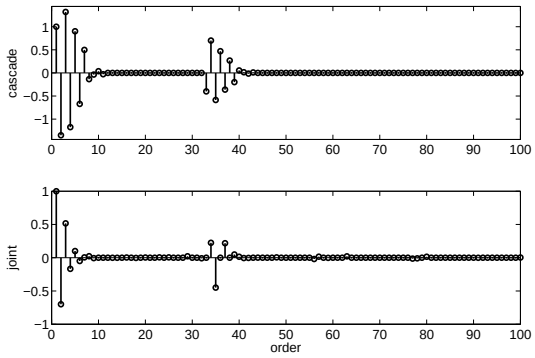


Fig. 3. An example of the high order predictor coefficient vector resulting from a cascade of long-term and short-term predictors (top box) and the solution of (9) for $\gamma = 0.1$ and order $K = 100$. The order is chosen sufficiently large to accommodate the filter cascade (8). It can be seen that the nonzero coefficient in the sparse prediction vector roughly coincide with the structure of the cascade of the two predictors.

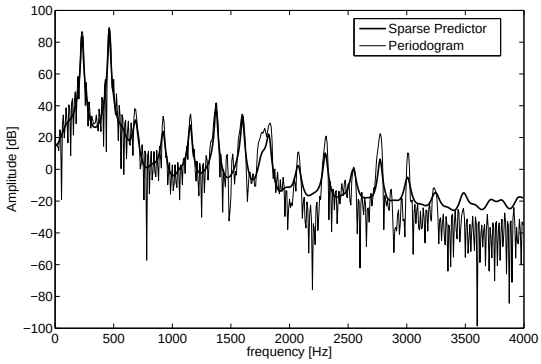


Fig. 4. Frequency response of the high order predictor of Figure 3. The order of the predictor is $K = 100$ and we consider only the nine nonzero coefficients of largest magnitude modeling the short-term and long-term predictors cascade.

The resulting prediction coefficient vector $\mathbf{a} = \{a_k\}$ of the high order polynomial $A(z)$ will therefore be highly sparse[3]. Taking this into account in our minimization process, and again considering the 1-norm as convex relaxation of the 0-norm, our original problem (5) becomes:

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1 + \gamma\|\mathbf{a}\|_1, \tag{9}$$

where the dimension of the prediction coefficient vector $\mathbf{a}$ (the order of the predictor) has to be sufficiently large to model the filter cascade ($K > N_f + T_p + N_p$) in (8). This approach, although maintaining resemblances to (7) looking for a sparse residual, is fundamentally different. While the predictor in (7) aims at modeling the spectral envelope, the purpose of the high order sparse predictor is to model the *whole* spectrum, i.e., the spectral envelope and the spectral harmonics. This can be easily achieved due to the strong ability of high order LP to resolve closely spaced sinusoids [30], [31]. Furthermore, considering the construction of the observation matrix $\mathbf{X}$, finding a high order sparse predictor is equivalent to identify which columns of $\mathbf{X}$, and in turn, which samples in $\mathbf{x}$ are important in the linear combination to predict a sample of speech (1). Thus, when a segment of voiced speech is analyzed with the predictive framework in (9), the nonzero coefficients roughly coincide with the structure in (8). An example of the predictor obtained as solution of (9) is shown in Figure 3. An example of the spectral modeling properties is shown in Figure 4.

There are mainly two problems associated with exploiting the modeling properties of the sparse high order predictor: determining an appropriate value of γ to solve (9) and using an approximate factorization to obtain again the initial formulation composed by the two predictors (8). Below we address these two issues.

1) Selection of γ: It is clear from (9) that γ controls *how sparse* the predictor should be and the trade-off between the sparsity of the predictor and the sparsity of the residual. In particular, by increasing γ, we increase the sparsity of the prediction coefficient vector, until all its entries are zero ($A(z) = 1$) for $\gamma \geq \|\mathbf{X}^T\mathbf{x}\|_\infty$ (where $\|\cdot\|_\infty$ denotes the dual norm to $\|\cdot\|_1$). More precisely, for $0 < \gamma < \|\mathbf{X}^T\mathbf{x}\|_\infty$, the solution vector $\mathbf{a}$ is a linear function of γ [32]. However, in general, the number of nonzero elements in $\mathbf{a}$ is not necessarily a monotonic function of γ.

There are obviously several ways of determining γ. In our previous work [21] [22], we have found the modified L-curve [33] as an efficient tool to find a balanced sparse representation between the two descriptions. The optimal value of γ (in the L-curve sense) is found as the point of maximum curvature of the curve $(\|\mathbf{x} - \mathbf{X}\mathbf{a}_\gamma\|_1, \|\mathbf{a}_\gamma\|_1)$. We have also observed that, in general, a constant value of γ, chosen for example as the average value of the set of γ's found with the L-curve based approach for a large set of speech frames, is an appropriate choice in the predictive problems considered. In the experimental analysis we will consider both approaches to defining γ.

2) Factorization of the high order polynomial: If γ is chosen appropriately, the considered formulation (9) results in a high order predictor $\hat{A}(z)$ with a clear structure that resembles the cascade of the short-term and long-term predictor (Figure 3). We can therefore bring $\hat{A}(z)$ to the original formulation in (8), by applying a simple and effective ad-hoc method to factorize the solution [22]. In particular, we use the first N_f coefficients of the high order predictor as the estimated coefficients of the short-term predictor:

$$\hat{F}(z) = 1 - \sum_{k=1}^{N_f} \hat{a}_k z^{-k}, \tag{10}$$

and then compute the quotient polynomial $\hat{Q}(z)$ of the division of $\hat{A}(z)$ by $\hat{F}(z)$ so that:

$$\hat{A}(z) = \hat{Q}(z)\hat{F}(z) + E(z) \approx \hat{Q}(z)\hat{F}(z), \tag{11}$$

where the deconvolution remainder $E(z)$ is considered to be negligible as most of the information of the coefficients has shown to be retained by $\hat{Q}(z)$ and $\hat{F}(z)$. From the polynomial $\hat{Q}(z)$ we can then extract the N_p taps predictor. In this paper,

[3]Traditionally, for speech sampled at 8 kHz, $N_f = 10$, $N_p = 1$, and T_p usually belongs in the range $[16, 120]$.

we will consider the most common pitch predictor where $N_p = 1$ ($P(z) = 1 - g_p z^{-T_p}$), then we merely identify the minimum value and its position in the coefficients vector of $\hat{Q}(z)$:

$$\begin{aligned} g_p &= \min\{q_k\}, \\ T_p &= \arg\min\{q_k\}. \end{aligned} \tag{12}$$

It is clear that, while heuristic, this factorization procedure is highly flexible. A different numbers of taps for both the short-term and long-term can be selected and also a voiced/unvoiced classification can be included, based on the presence or absence of long-term information, as described in [21], [22].

It should be noticed that the structure of the cascade can also be incorporated into the minimization scheme and can be potentially beneficial in reducing the size of the problem. This approach is then similar to the *One-Shot Combined Optimization* presented in [17] which is implicitly a sparse method looking for a similar high order factorisable predictor. The joint estimation in this case requires prior knowledge on the position of the pitch contributions (a pitch estimate) and the model order of both the short-term and long-term predictors. Differently from this method, in our approach, we obtain information on the model order of both short-term and long-term contribution and a pitch estimate, just by a simple post-processing the solution of (9).

C. Enhancing sparsity by reweighted 1-norm minimization

As shown throughout this section, the 1-norm is used as a convex relaxation of the 0-norm, because 0-norm minimization yields a combinatorial problem (NP-hard). We are therefore interested in adjusting the error weighting difference between the 1-norm and the 0-norm. A variety of recently introduced methods have dealt with reducing the error weighting difference between the 1-norm and the 0-norm by relying on the iterative reweighted 1-norm minimization (see, e.g., [34] and references therein). In particular, the iteratively reweighted 1-norm minimization may be used for estimating $\mathbf{a}$ and enhancing the sparsity of $\mathbf{r}$ (and $\mathbf{a}$), while keeping the problem solvable with convex tools [28] [23]. The predictor can then be seen as a solution of the following minimization problem:

$$\mathbf{a} = \arg\min_{\mathbf{a}} \lim_{p\to 0} \lim_{k\to 0} \{\|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p + \gamma \|\mathbf{a}\|_k^k\}, \tag{13}$$

where each iteration of the reweighting process brings us closer to the 0-norm.

The mismatch between the 0-norm and the 1-norm minimization can be seen more clearly in Figure 5, where larger coefficients are penalized more heavily by the 1-norm than small ones. From an optimization point of view, when $p \leq 1$, the cost functions will have lower emphasis on large values and sharper slopes near zero compared to the $p = 1$ case. In turn, from a statistical point of view, the density functions will have heavier tails and a sharper slope near zero. This means that the minimization will encourage small values to become smaller while enhancing the amplitude of larger values. The limit case for $p = 0$ will have an infinitely sharp slope in zero and equally weighted tails. This will introduce as many zeros as possible as these are infinitely weighted. In this sense, the 0-norm can be seen as more "impartial" by penalizing every nonzero coefficient equally. It is clear that if a very small value would be weighted as much as a large value, the minimization process will eliminate the smaller ones and enhance the larger ones.

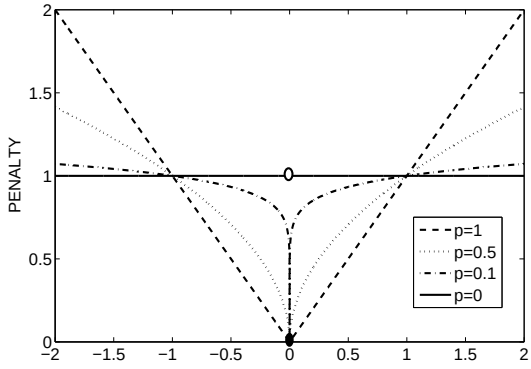


Fig. 5. Comparison between cost functions for $p \leq 1$. The 0-norm can be seen as more "democratic" than any other norm by weighting all the nonzero coefficients equally.

The algorithm to obtain a short-term predictor engendering a sparser residual, a reweighted formulation of (7), is shown in Algorithm 1. This approach, as we shall see, becomes beneficial in finding a predictor that produces a sparser residual, providing a tighter coupling between the prediction estimation and the search for the approximated sparse excitation. An example of the reweighted residual estimate is shown in Fig. 6.

When we impose sparsity both on the residual and on the high order predictor, as in (9), the algorithm is modified as shown in Algorithm 2. This formulation is relevant as it enhances the components that contain the information regarding the near-end and far-end redundancies in the high order predictor making the approximate factorization presented in III-B2 more accurate. In particular, the reweighting allows to reduce the spurious near-zero components in the high order predictor obtained (see Fig. 3) while enhancing the larger components that contain information of near-end and far-end redundancies.

It has been shown in [28] that $\|\hat{\mathbf{r}}^{i+1}\|_1 \leq \|\hat{\mathbf{r}}^{i}\|_1$, meaning that this is a descent algorithm. The halting criterion can therefore be chosen as either a maximum number of iterations or as a convergence criterion. In the experimental analysis we will give details on how many iterations are required in our setting. In both algorithms, the parameter $\epsilon > 0$ is used to provide stability when a component of $\hat{\mathbf{r}}$ goes to zero.

As a general remark, in [28] and [34], it is also shown that the reweighted 1-norm algorithm, at convergence, is equivalent to the minimization of the log-sum penalty function. This is relevant to what we are trying to achieve in (13): the log-sum cost function has a sharper slope near zero compared to the 1-norm, providing more effective sparsity inducing properties. Furthermore, since the log-sum is not convex, the iterative algorithm corresponds to minimizing a sequence of linearizations of the log-sum around the previous solution

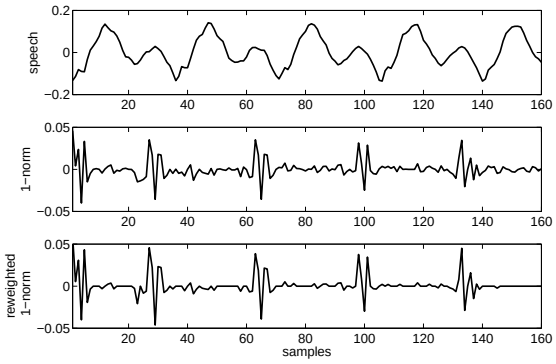


Fig. 6. An example of prediction residuals obtained through 1-norm and reweighted 1-norm error minimization using Algorithm 1. The speech segment analyzed is shown in the top box. The prediction order is $K = 10$ and the frame length is $N = 160$. Five iterations where made with $\epsilon = 0.01$.

Algorithm 1 Iteratively Reweighted 1-norm Minimization of the Residual

Inputs: speech segment $\mathbf{x}$
Outputs: predictor $\hat{\mathbf{a}}^i$, residual $\hat{\mathbf{r}}^i$
$i = 0$, initial weights $\mathbf{W}^{i=0} = \mathbf{I}$
while halting criterion false **do**
1. $\hat{\mathbf{a}}^i, \hat{\mathbf{r}}^i \leftarrow \arg\min_{\mathbf{a}} \|\mathbf{W}^i\mathbf{r}\|_1 \quad \text{s.t.} \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}$
2. $\mathbf{W}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{r}}^i\right| + \epsilon\right)^{-1}$
3. $i \leftarrow i + 1$
end while

estimate, providing at each step a sparser solution (until convergence).

IV. Compressed Sensing in Sparse Linear Prediction

The CS formulation is particularly relevant in our sparse recovery problems: by exploiting prior knowledge about the sparsity of the signal $\mathbf{x}$ we will show that a limited number of random measures are sufficient to recover our predictors and sparse residual with high accuracy. In particular, it has been shown [24] [35] that a random projection of a high-dimensional but sparse or compressible signal vector onto a lower-dimensional space contains enough information to be able to reconstruct, with high probability, the signal with small or zero error. The random measures in CS literature are usually obtained by projecting the considered measurement vectors onto a lower dimensional space, using random matrices.

In recent work [36], [37], CS formulations in the context of speech analysis and coding have been formulated in order to find a sparse approximation of the residual, given the predictor. It is then interesting to extend this work to the case where we want to find directly the predictor that engenders intrinsically a sparse residual. In particular, given the sparsity level of the sparse representation that we wish to retrieve in a given domain, we can determine an efficient *shrinkage* of the minimization problem in a lower dimensional space, with a clear impact on the computational complexity.

If we wish to perform CS, two main ingredients are needed: a domain where the analyzed signal is sparse and the sparsity

Algorithm 2 Iteratively Reweighted 1-norm Minimization of Residual and Predictor

Inputs: speech segment $\mathbf{x}$
Outputs: predictor $\hat{\mathbf{a}}^i$, residual $\hat{\mathbf{r}}^i$
$i = 0$, initial weights $\mathbf{W}^{i=0} = \mathbf{I}$ and $\mathbf{D}^{i=0} = \mathbf{I}$
while halting criterion false **do**
1. $\hat{\mathbf{a}}^i, \hat{\mathbf{r}}^i \leftarrow \arg\min_{\mathbf{a}} \|\mathbf{W}^i\mathbf{r}\|_1 + \gamma\|\mathbf{D}^i\mathbf{a}\|_1$ s.t. $\mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}$
2. $\mathbf{W}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{r}}^i\right| + \epsilon\right)^{-1}$
3. $\mathbf{D}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{a}}^i\right| + \epsilon\right)^{-1}$
4. $i \leftarrow i + 1$
end while

level of this signal T. In our case, the residual is the domain where the signal is sparse, while the linear transform that maps the original speech signal to the sparse residual is the sparse predictor. The sparsity in the residual domain is then imposed by our needs [35]. Let us now review the formulation presented in [37]:

$$\hat{\mathbf{r}} = \arg\min_{\mathbf{r}} \|\mathbf{r}\|_1 \quad \text{s.t.} \quad \mathbf{\Phi x} = \mathbf{\Phi H r} \tag{14}$$

where $\mathbf{x}$ is the $N \times 1$ analyzed segment of speech, $\mathbf{H}$ the $N \times (N+K)$ synthesis matrix, constructed from the truncated impulse response of the *known* predictor [38], $\mathbf{r}$ is the residual vector to be estimated (supposedly sparse) and $\mathbf{\Phi}$ is the sensing matrix of dimension $M \times N$. The dimensionality of the random linear projection M stems from the sparsity level T that one wishes to impose on the residual. In particular, based on empirical results, the number of projections is set equal to four times the sparsity, i.e. $M = 4T$. Furthermore, when the incoherence between the synthesis matrix and the random basis matrix $\mathbf{\Phi}$ holds ($\mu(\mathbf{\Phi}, \mathbf{H}) \approx 1$), even if $\mathbf{H}$ is not orthogonal the recovery of the sparse residual $\mathbf{r}$ is still possible and the linear program in (14) gives an accurate reconstruction of $\mathbf{x}$ with very high probability [24], [37]. As a general remark, the entries of the random matrix can be drawn from many different processes [39], in our case we will use a i.i.d. Gaussian process, as done in [36], [37].

To adapt CS principles to the estimation of the predictor as well, let us now consider the relation between the synthesis matrix $\mathbf{H}$ and the analysis matrix $\mathbf{A}$ where one is the pseudo-inverse of the other [40]:

$$\mathbf{A} = \mathbf{H}^{+}. \tag{15}$$

We can now replace the constraint $\mathbf{\Phi x} = \mathbf{\Phi H r}$ in (14) as

$$\mathbf{\Phi r} = \mathbf{\Phi A x}, \tag{16}$$

where $\mathbf{A}$ is the $(N+K) \times N$ analysis matrix that performs the whitening of the signal, constructed from the coefficients of the predictor $\mathbf{a}$ of order K [40], the dimension of the sensing matrix $\mathbf{\Phi}$ is now adjusted accordingly to $M \times (N+K)$. Notice that, due to the structure of $\mathbf{A}$ this can be rewritten equivalently to:

$$\mathbf{\Phi r} = \mathbf{\Phi A x} = \mathbf{\Phi}\left[\mathbf{x}|\mathbf{X}\right]\left[1|\mathbf{a}^T\right]^T, \tag{17}$$

where $[\mathbf{x}|\mathbf{X}]$ is the matrix obtained by stacking the vector $\mathbf{x}$ to the left of $\mathbf{X}$ in (4). The minimization problem can then be

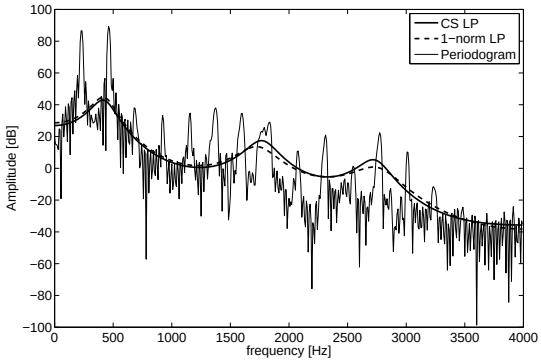


Fig. 7. An example of LP spectral model obtained through 1-norm minimization (7) and through CS based minimization (18) for a segment of voiced speech. The prediction order is $K = 10$ and the frame length is $N = 160$, for the CS formulation the dimension of the sensing matrix is $M = 80$, corresponding to the sparsity level $T = 20$.

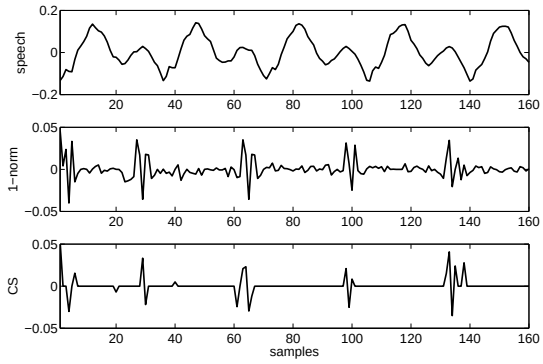


Fig. 8. An example of prediction residuals obtained through 1-norm minimization and CS recovery. The speech segment analyzed is shown in the top box. The prediction order is $K = 10$ and the frame length is $N = 160$. For the CS formulation, the imposed sparsity level is $T = 20$, corresponding to the size $M = 80$ for the sensing matrix.

rewritten as:

$$\min_{\mathbf{a},\mathbf{r}} \|\mathbf{r}\|_1 \quad \text{s.t.} \quad \mathbf{\Phi r} = \mathbf{\Phi}(\mathbf{x} - \mathbf{Xa}). \tag{18}$$

We can now see that (18) is *equivalent* to (7), the only difference being the projection onto the random basis in the constraint. Therefore, (7) can be seen as a particular case of the formulation in (18) where $\mathbf{\Phi} = \mathbf{I}$ and $\mathbf{I}$ is a identity matrix of size $(N + K) \times (N + K)$. In this case we are then not actually performing a projection in a random subspace. The minimization constraint on the left side of (18) would become:

$$\mathbf{\Phi r} = \mathbf{\Phi}(\mathbf{x} - \mathbf{Xa}) \quad \Rightarrow \quad \mathbf{r} = \mathbf{x} - \mathbf{Xa} \quad \text{for} \quad \mathbf{\Phi} = \mathbf{I}. \tag{19}$$

The results obtained will then be similar to our initial formulation (7), as long as the choice of $\mathbf{\Phi}$ is appropriate. In this case, the formulation in (18) will not only provide hints on the T pulses to be selected in the residual, but also a dimensionality reduction that will simplify the calculations. This computational complexity reduction, resulting from the dimensionality reduction given by the projection onto random basis has been also observed in [41] and arises from the Johnson-Lindestrauss lemma [42]. An example of an envelope estimation using the formulation in (18) is presented in Figure 7 while the recovered sparse residual is shown in Figure 8.

Similarly, if we are looking for a high order sparse predictor, the problem (9) can be cast into a CS framework leading to:

$$\arg\min_{\mathbf{a},\mathbf{r}} \|\mathbf{r}\|_1 + \gamma\|\mathbf{a}\|_1 \quad \text{s.t.} \quad \mathbf{\Phi r} = \mathbf{\Phi}(\mathbf{x} - \mathbf{Xa}). \tag{20}$$

The formulation (9) and (20), similarly to (7) and (18), become equivalent when $\mathbf{\Phi} = \mathbf{I}$ and the minimization constraint is then (19). Both formulations (18) and (20), can also be modified to involve iterative reweighting (Algorithm 3 shows the general case for $\gamma > 0$).

Algorithm 3 CS Formulation of the Iteratively Reweighted 1-norm Minimization of Residual and Predictor

Inputs: speech segment $\mathbf{x}$, desired residual sparsity level T
Outputs: predictor $\hat{\mathbf{a}}^i$, residual $\hat{\mathbf{r}}^i$
$i = 0$, initial weights $\mathbf{W}^{i=0} = \mathbf{I}$ and $\mathbf{D}^{i=0} = \mathbf{I}$,
random matrix $\mathbf{\Phi}$ of size $M \times (N + K)$, $M = 4T$
while halting criterion false **do**
1. $\hat{\mathbf{a}}^i, \hat{\mathbf{r}}^i \leftarrow \arg\min_{\mathbf{a}} \|\mathbf{W}^i\mathbf{r}\|_1 + \gamma\|\mathbf{D}^i\mathbf{a}\|_1$ s.t. $\mathbf{\Phi r} = \mathbf{\Phi}(\mathbf{x} - \mathbf{Xa})$
2. $\mathbf{W}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{r}}^i\right| + \epsilon\right)^{-1}$
3. $\mathbf{D}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{a}}^i\right| + \epsilon\right)^{-1}$
4. $i \leftarrow i + 1$
end while

V. Properties of Sparse Linear Prediction

As mentioned in the introduction, many problems appearing in traditional 2-norm LP modeling of voiced speech can be traced back to the inability of the predictor to decouple the vocal tract transfer function from the pitch excitation. This results in a lower spectral modeling accuracy and a strong dependence on the placement of the analysis window. In this section we provide some experiments to illustrate how the sparse linear predictors presented in the previous sections manage to overcome these problems. As a general remark, it is well-known that the $p-$norm LP estimate with $p \neq 2$ is not guaranteed to be stable [43]. Nevertheless, the results presented in this section concentrate on the spectral modeling properties of sparse LP, thus the stability of the predictor is simply imposed by pole reflection which stabilizes the filter without modifying the magnitude of the frequency response. We will provide a thorough discussion of the stability issues in the Section VII and in Section VI where the speech coding properties are analyzed and stability is critical.

The experimental analysis was done on 20,000 frames of length $N = 160$ (20 ms) of clean voiced speech coming from several different speakers with different characteristics (gender, age, pitch, regional accent) taken from the TIMIT database, downsampled at 8 kHz. The prediction methods we compare in this section are shown in Table I. The optimality of the methods **BE** and **RLP**, presented in [6], comes from the selection of the parameters which provided the lowest distortion compared with the reference envelope. For brevity and clarity of the presented results, we omitted the predictors obtained as solutions of the iterative reweighted algorithms

TABLE I
PREDICTION METHODS COMPARED IN THE MODELING PROPERTIES EVALUATION.

Method	Description
LP	Traditional 2-norm LP with 10Hz bandwidth expansion ($\gamma = 0.996$) and Hamming windowing.
SpLP10	1-norm LP presented in (III-A), solution of (7). Stability is imposed by pole reflection if unstable. No windowing is performed.
SpLP11	1-norm LP presented in (III-B). The order of (9) is $K = 110$ (covering accurately pitch delays in the interval $[N_f + 1, K - N_f - 1]$). γ is chosen as the point of maximum curvature in the L-curve. The short-term predictor coefficients are the first N_f coefficients of the high order polynomial. Stability is imposed by pole reflection if unstable. No windowing is performed.
BE	Optimally bandwidth expanded 2-norm LP as shown in [6]. Hamming window is used.
RLP	Optimally regularized 2-norm LP as shown in [6]. Hamming window is used.

presented in Section III-C and the CS formulation presented in Section IV. These methods, while presenting very similar modeling properties to **SpLP10** and **SpLP11**, produce predictors estimates with slightly higher variance, thus requiring few more bits to be encoded. Therefore, while it is hard to provide a fair comparison in terms of modeling, their properties become more interesting in the coding scenario that will thoroughly analyzed in Section VI; in particular, the differences in their bit allocation necessary for efficient coding and the information required in the residual will be analyzed.

A. Spectral Modeling

In this section, we provide results to the modeling properties of the short-term predictors. As a reference, we used the envelope obtained through a cubic spline interpolation between the harmonics peaks of the logarithmic periodogram. This method was presented in [6] and provided an approximation of the vocal tract transfer function, without the fine structure corresponding to the pitch excitation. We then calculated the log spectral distortion between our reference envelope $S_{int}(\omega)$ and the estimated predictive model $S(\omega, \mathbf{a})$ as:

$$\mathrm{SD}_m = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi} \left[10\log_{10} S_{int}(\omega) - 10\log_{10} S(\omega, \mathbf{a})\right]^2 d\omega}. \tag{21}$$

where the numerator gain is calculated as the variance of the residual.

The coefficients of the short-term predictors presented have also shown to be smoother and therefore they have a lower sensitivity to quantization. We also compared the log spectral distortion between our reference envelope $S_{int}(\omega)$ and the quantized predictive model $S(\omega, \hat{\mathbf{a}})$ for every predictor obtained with the presented methods. The quantizer used is the one presented in [44], with the number of bits fixed at 20 for the different prediction orders, providing in all the method presented a *transparent coding*[4]. The results are shown in Table II for different prediction orders. A critical analysis of the results showed the improved modeling properties of **SpLP11**. This was given by its ability to take into consideration the whole speech production model, thus decoupling more effectively the short-term contribution that provides the spectral envelope from the contribution given by the pitch excitation. **SpLP10** and **RLP** achieved similar performance, providing evidence supporting the generally good spectral modeling properties of the minimization problem in (7).

TABLE II
AVERAGE SPECTRAL DISTORTION FOR THE CONSIDERED METHODS IN THE UNQUANTIZED CASE (SD_m) AND QUANTIZED CASE (SD_q). A 95% CONFIDENCE INTERVAL IS GIVEN FOR EACH VALUE.

METHOD	K	SD_m	SD_q
LP	8	2.11±0.06	3.24±0.11
	10	1.97±0.03	2.95±0.09
	12	1.98±0.05	2.72±0.12
SpLP10	8	1.91±0.01	2.92±0.02
	10	1.78±0.01	2.53±0.02
	12	1.61±0.01	2.31±0.04
SpLP11	8	1.64±0.00	2.65±0.01
	10	1.69±0.00	2.37±0.01
	12	1.39±0.01	2.13±0.01
BE	8	2.04±0.03	3.11±0.08
	10	1.88±0.02	2.92±0.07
	12	1.83±0.10	2.71±0.04
RLP	8	1.89±0.02	2.93±0.04
	10	1.72±0.01	2.51±0.03
	12	1.53±0.02	2.22±0.04

B. Shift Invariance

In speech analysis, a desirable property for an estimator is to be invariant to the small shifts of the analysis window, since speech, and voiced speech in particular, is assumed to be short-term stationary. However, standard LP is well-known not to be shift invariant [8]. This is a direct consequence of the coupling between the vocal tract transfer function and the underlying pitch excitation that standard LP introduces in the estimate. To analyze the invariance of the LP methods to window shifts, we took the same 20,000 frames of clean voiced speech and we expanded them to the left and to the right with 20 samples, giving a total length $N = 200$. In each frame of length $N = 200$ we defined a $M = 160$ samples boxcar window and we shifted the window by $s = 1, 2, 5, 10, 20$ samples. The average log spectral difference of the 10^{th} order AR estimate between $S_0(\omega)$ and $S_s(\omega)$ was analyzed. The average differences obtained for the methods in Table I are shown in Table III. In Figure 9, we show an example of the shift invariance property. The results obtained indicate clearly the sparse predictors robustness to small shifts in the analyzed window. While the decay in performance for increasing shift

[4]According to [45], transparent coding of LP parameters is achieved when the two versions of coded speech, obtained using unquantized LP parameters and quantized LP parameters, are indistinguishable through listening. This is usually achieved with an average log distortion between quantized and unquantized spectra lower that 1 dB, with no outliers with log distortion greater than 4 dB and a number of outliers with 2-4 dB distortion lower than 2%. Furthermore, according to [46] the quality threshold for the model naturally follows from a distortion measure for the signal, the result being independent of rate, and giving the same well-known 1 dB without invoking notions of perception.

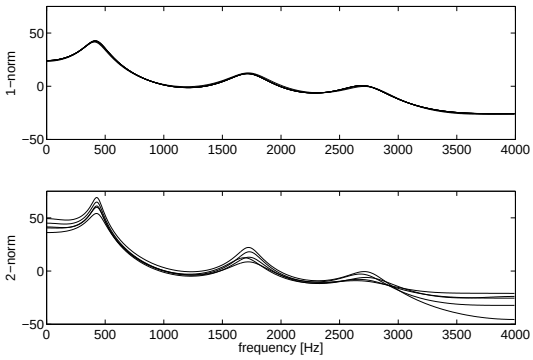


Fig. 9. An example of the shift invariance property of the sparse linear predictor (**SpLP11**) (top box), compared to traditional LP (**LP**). Ten envelopes are analyzed by shifting a the analysis window (160 samples) of $s = 1, 2, 5, 10, 20$ samples over a stationary voiced speech segment (length 200 samples).

TABLE III
AVERAGE SPECTRAL DISTORTION FOR THE CONSIDERED METHODS WITH SHIFT OF THE ANALYSIS WINDOW $s = 1, 2, 5, 10, 20$.

METHOD	SD_1	SD_2	SD_5	SD_{10}	SD_{20}
LP	0.113	0.128	0.223	0.452	1.262
SpLP10	0.003	0.003	0.011	0.017	0.032
SpLP11	0.001	0.002	0.005	0.006	0.009
BE	0.097	0.117	0.197	0.238	0.328
RLP	0.015	0.089	0.180	0.201	0.323

in the analysis window is comparable for all methods, the sparse predictors still retains better performance. Also in this case, the change in the frequency response in traditional LP is clearly given by the pitch bias in the estimate of the predictor, particularly dependent on the location of the spikes of the pitch excitation.

C. Pitch Independence

The ability of the sparse linear predictors to decouple the pitch excitation from the vocal tract transfer function is reflected also in the ability to have estimates of the envelope that are not affected by the pitch. In this experiment, we calculated the envelope using 10^{th} order regularized LP (**RLP**) and we modeled the underlying pitch excitation with an impulse train with different spacing. We then filtered this synthetic pitch excitation through the LP filter obtained and analyzed the synthetic speech applying the different LP methods in Table I. We divided the analysis into three subsets: high-pitched $T_p \in [16, 35]$ ($f_0 \in [228\text{Hz}, 500\text{Hz}]$), mid pitched $T_p \in [36, 71]$ ($f_0 \in [113\text{Hz}, 222\text{Hz}]$) and low pitched $T_p \in [72, 120]$ ($f_0 \in [67\text{Hz}, 111\text{Hz}]$). The shortcomings of LP can be particularly seen in high-pitched speech, as shown in the results of Table IV. Because high-pitched speakers have fewer harmonics within a given frequency range, modeling of the spectral envelope is more difficult and particularly problematic for traditional LP. The sparse linear predictors are basically unaffected by the underlying pitch excitation, which results in an improved spectral modeling. In particular for **SpLP11**, since the high order structure of the initial estimate includes the pitch harmonic structure, the extracted short-term predictor is particularly robustly independent from the underlying excitation.

TABLE IV
AVERAGE SPECTRAL DISTORTION FOR THE CONSIDERED METHODS WITH DIFFERENT UNDERLYING PITCH EXCITATION. A 95% CONFIDENCE INTERVAL IS GIVEN FOR EACH VALUE.

METHOD	low	mid	high
LP	0.81±0.12	1.04±0.23	1.32±0.56
SpLP10	0.02±0.00	0.09±0.00	0.11±0.01
SpLP11	0.00±0.00	0.00±0.00	0.01±0.00
BE	0.45±0.07	0.65±0.19	0.89±0.34
RLP	0.05±0.02	0.16±0.10	0.19±0.09

VI. CODING APPLICATIONS OF SPARSE LINEAR PREDICTION

By introducing sparsity in the residual, we can reasonably assume that only a small portion of the residual samples are sufficient to reconstruct the speech signal with high accuracy. We will corroborate our intuition by providing some experiments on the coding applications of sparse linear prediction. Specifically, in VI-A, we will first give experimental proof of the sparsity inducing effectiveness of the short-term predictors in the Analysis-by-Synthesis (AbS) scheme [38]. In this case, we used a very simple excitation model coding without long-term prediction where we exploit directly the information on the location of the nonzero samples. In VI-B, we will present a simple coding procedure that exploits the properties of the combined high order sparse LP and sparse residual. As we shall see in VI-C, this approach presents interesting properties such as noise robustness for which we give both objective and subjective evaluation.

As a general remark, since the stability of the short-term predictors is not assured, we consistently performed a stability check and, if the short-term predictor was found to be unstable, we performed a pole reflection. Note that this approach necessarily modifies the time domain behavior of the residual as well as the predictor coefficients. Nevertheless, since the rate of unstable filters is low and the instability is very mild (i.e., the magnitude of the poles is only very slightly higher than one), this can be considered as an adequate solution to this problem. We will return to the stability issue in Section VII.

A. Coding Properties of the Short-Term Sparse Linear Predictor

The first experiment regards the use of the short-term predictor in speech coding. In particular we compared the use of the multipulse encoding procedure in the case of bandwidth expanded linear prediction (**LP**) with a fixed bandwidth expansion of 60 Hz (done by lag-windowing the autocorrelation function [38]). We compared this approach with our introduced sparse linear predictors. The only difference is that, instead of performing the multipulse encoding, we performed the AbS procedure straight after selecting the T positions of the T largest samples that are located in the residual. In this experiment, we did not perform long-term prediction, focusing only on the coding properties of the sparsity inducing short-term predictors.

TABLE V
PREDICTION METHODS COMPARED IN THE CODING PROPERTIES EVALUATION.

Method	Description
LP	Traditional 2-norm LP with a fixed bandwidth expansion of 60 Hz (done by lag-windowing the autocorrelation function) and Hamming windowing.
SpLP10	1-norm LP solution of (7).
RWLP10	Reweighted 1-norm LP presented in Section III-C using Algorithm 1. Four reweighting iterations are performed (sufficient for convergence).
CSLP10	Compressed sensing formulation presented in Section IV, solution of (18). The size of the sensing matrix is given by the number of samples we want to retrieve in the residual.
RWCSLP10	Reweighed compressed sensing formulation of **CSLP10** using Algorithm 1. Four reweighting iterations are performed (sufficient for convergence).

TABLE VI
COMPARISON BETWEEN THE SPARSE PREDICTOR ESTIMATION METHODS. A 95% CONFIDENCE INTERVAL IS GIVEN FOR EACH VALUE.

METHOD	T	$\hat{\mathbf{a}}$	SSNR	MOS	t
LP	5	19	14.1±3.2	2.85±0.23	0.1±0.1
	10	19	19.1±2.9	3.01±0.16	0.9±0.3
SpLP10	5	18	15.3±2.1	2.87±0.12	1.3±0.2
	10	18	20.1±1.7	3.11±0.11	1.3±0.2
RWLP10	5	22	17.2±1.6	3.01±0.06	4.1±0.3
	10	22	21.4±1.5	3.19±0.03	4.1±0.3
CSLP10	5	19	16.9±1.9	2.97±0.04	0.4±0.0
	10	19	20.9±1.5	3.25±0.03	0.6±0.2
RWCSLP10	5	24	20.2±0.9	3.15±0.03	1.3±0.3
	10	24	24.4±0.4	3.43±0.01	1.9±0.2

We considered the formulation **SpLP10**, reweighted 1-norm **RWLP10**, and their CS formulations **CSLP10** and **RWCSLP10**. The methods compared are summarized in Table V. As mentioned in Section V, all these methods achieve similar modeling performance to **SpLP10**, although their estimate of the predictor requires a slightly larger number of bits. Here we will show this providing a comparison also in terms of bits needed for transparent quantization of the predictor. The methods **BE** and **RLP**, presented in the previous section (Table I) while offering better modeling properties than traditional LP, do not provide any significant improvement in the coding scenario, thus they will be omitted from the current experimental analysis.

We have performed the analysis on the same speech signals database considered in Section V. The frame size is $N = 40$, the 10^{th} order predictors were quantized transparently using the LSFs coding method in [44] while the T pulses are left unquantized. In the CS formulations the sensing matrix has $M = 4T$ rows; this means that just a slight reduction in the size of the problem was obtained when $T = 10$. Nevertheless we were able to obtain important information on the location of the pulses. In the reweighted schemes, the number of iterations is four, which was sufficient to reach convergence in all the analyzed frames.

In Table VI, we present the results in terms of Segmental SNR, Mean Opinion Score (obtained through PESQ evaluation) and empirical computational time t in elapsed CPU seconds for $T = 5$ and $T = 10$, and number of bits necessary to transparently encode the predictor ($\hat{\mathbf{a}}$) using LSFs [44]. The results demonstrate the effectiveness of the sparse linear predictors. These results also show that the predictors in the reweighted cases (**RWLP10** and **RWCSLP10**), need a larger number of bits for transparent quantization due to the larger variance of their estimates. This result is particularly interesting when considering the model in (2). In particular, the description of a segment of speech is distributed between its predictive model and the corresponding excitation. Thus, we can observe that the complexity of the predictor necessarily increases when the complexity of the residual decreases (less significant pulses). This also leaves open questions on the *optimal* bit distribution between the two descriptions. As a proof of concept, the results show how only 5 bits of difference between **LP** and **RWCSLP10** in the representation of the filter result in a significant improvement in performance: only 5 pulses in the residual are necessary in **RWCSLP10** to obtain similar performance to **LP** using 10 pulses.

A critical analysis of the results leads to another interesting conclusion. In fact, while 1-norm based minimization, with or without the *shrinkage* of the problem provided by the CS formulation in (18), is computationally more costly, than 2-norm minimization, it greatly simplifies the next stage where the excitation is selected in a closed-loop AbS scheme. In particular, the empirical computational time in Table VI refers to both the LP analysis stage and the search for the MPE excitation. Since the MPE search for the location is not performed in our sparse LP methods and we exploit directly the information regarding the T pulses of largest magnitude, the AbS procedure is merely a small least square problem where we find the T pulse amplitudes. We will come back to the discussion regarding complexity in VII-B. Furthermore, it should be noted that the CS formulation improves the selection of the T largest pulses. This is remarkable since while the predictor obtained with or without the random projection is similar, the reduction of the constraints helps us find a more specific solution for the level of sparsity T that we would like to retrieve in the residual. As mentioned above, the price to pay is a slightly higher bit allocation for the predictors obtained through CS formulation.

B. Speech coding based on Sparse Linear Prediction

As a proof of concept, we will now present a very simple coding scheme that incorporates all the previously introduced methods. We will use the method presented in Section III-B, exploiting the sparse characteristics of the high order predictor and the sparse residual. In order to reduce the number of constraints, we cast the problem in a CS formulation (20) that provides a shrinkage of the constraints according to the number of samples we wish to retrieve in the residual. Furthermore, in order to refine the initial sparse solution, we apply the reweighting algorithm. The core scheme is summarized in Algorithm 3. Differently from multistage coders, this method, with its joint estimation of a short-term and a long-term predictor and the presence of a sparse residual, provides a one-step approach to speech coding. In synthesis, given a segment of speech, a way to encode the speech signal can be as follows:

1) Define the desired level of sparsity of the residual T

TABLE VII
COMPARISON BETWEEN THE CODING PROPERTIES OF THE **AMR102** AND THE CODER BASED ON SPARSE LINEAR PREDICTION **SPLP**. A 95% CONFIDENCE INTERVAL IS GIVEN FOR EACH VALUE.

METHOD	rate	MOS	t
AMR102	10.2 kbps	4.02±0.11	0.1±0.0
SpLP	10.1 kbps	4.13±0.13	1.2±0.1

TABLE VIII
PERFORMANCES OF **AMR102** AND THE CODER BASED ON SPARSE LINEAR PREDICTION (**SPLP**) FOR DIFFERENT VALUES OF SNR (WHITE GAUSSIAN NOISE). A 95% CONFIDENCE INTERVAL IS GIVEN FOR EACH VALUE.

METHOD	clean	30dB	20dB	10dB
AMR102	4.02±0.11	3.88±0.21	3.25±0.19	2.76±0.23
SpLP	4.13±0.13	3.94±0.15	3.52±0.14	3.21±0.19

and define the sensing matrix dimensionality accordingly $M = 4T$.
2) Perform n steps of the CS reweighted minimization process (Algorithm 3).
3) Factorize the prediction coefficients into a short-term and long-term predictor using the procedure in III-B2.
4) Quantize short-term and long-term predictors.
5) Select the T positions where the values of largest magnitude are located.
6) Solve the analysis-by-synthesis equation keeping only the T nonzero positions.
7) Quantize the residual.

We have again analyzed about one hour of clean speech taken from the TIMIT database. In order to obtain comparable results, the frame length is now $N = 160$ (20 ms). The order of the high order predictor in (20) is $K = 110$ (meaning that we can cover accurately pitch delays in the interval $[N_f + 1, K - N_f - 1]$, including the usual range for the pitch frequency $[70\text{Hz}, 500\text{Hz}]$). the fixed regularization parameter is $\gamma = 0.12$ and the defined level of sparsity is $T = 20$. Four iterations of the reweighting minimization process are performed, sufficient to reach convergence in all the analyzed frames. The orders of the short-term and long-term predictors obtained from the factorization of the high order predictor are $N_f = 10$ and $N_p = 1$, respectively. 25 bits are used to transparently encode the LSF vector, 7 bits are used to quantize the pitch period T_p and 6 bits to quantize the pitch gain g_p. The stability of the overall cascade is imposed by pole reflection on the short-term predictor, and by limiting the pitch gain to be less that unity. As for the residual, the quantizer normalization factor is logarithmically encoded with 6 bits while a 8 levels uniform quantizer is used to quantize the normalized amplitudes; the signs are coded with 1 bit per each pulse. The upper bound given by the information content of the pulse location ($\log_2 \binom{160}{20}$ bits) is used as an estimate of the number of bits used for distortionless encoding of the location. No perceptual weighting is performed in our case. The total number of bits per frame used are 202, producing a 10.1 kbps rate. We will compare this method (**SpLP**) with the AMR coder in the 10.2 kbps mode (**AMR102**) [47]. The results in terms of MOS (obtained through PESQ evaluation) and empirical computation time are shown in Table VII and demonstrate similar performance but with a more straightforward approach to coding than AMR. The CS formulation also helps to generally keep the problem solvable in reasonable time.

C. Noise Robustness

This study is motivated by the ability of a sparse coder to identify more effectively the features of the residual signal that are important for its reconstruction, discarding those which probably are a result of the noise. The traditional encoding formulation, based on minimum variance analysis and residual encoding through pseudo-random sequences (i.e., algebraic codes), makes the identification of these important features basically impossible and requires, for low SNRs, noise reduction in the preprocessing. Interestingly enough, sparse LP based coding appears to be quite robust in the presence of noise. An example of the different performance in terms of MOS for different SNR under additive white Gaussian noise is given in table VIII.

D. Subjective assessment of speech quality

To further investigate the properties of our methods, we have conducted two MUSHRA listening tests [48] with 16 non-expert listeners. Ten speech clips were used in the listening test. In the first MUSHRA test we investigate what we have shown in VI-B, about the similarity in quality between the AMR coder and our method. In the second MUSHRA test the noise robustness of our method, discussed in VI-C, is proved. The test results are presented in Figure 10 where the score 100 corresponds to "Imperceptible" and the score 0 corresponds to "Very annoying" according to the 6-grade impairment scale. From the results, we can see that our method does not affect greatly the quality of the signal, given that our method is conceptually simpler and substantially less optimized compared to AMR. For example, we are not taking into account some of the main psychoacoustic criteria usually implemented in the AMR, such as the adaptive postfilter to enhance the perceptual quality of the reconstructed speech and the perceptual weighting filter employed in the analysis by synthesis search of the codebooks. Neverthless, in clean condition the average score was 89 for **AMR102**, and 82 for **SpLP**. The most significant results though, are the one related to the coding of noisy signals. In particular, we can see from Figure 10 that our method scores considerably better than the AMR showing how a sparse encoding technique can be more effective in noise robust speech coding. In fact, in noisy conditions, the average score was 62 for **AMR102**, and 75 for **SpLP**.

VII. DISCUSSION

A. Stability

In the presented applications of sparse linear predictors, the percentage of unstable filters was found to be low (around 2%) and the instability "mild[5]." This suggested the use of a simple stability check and pole reflection in our experimental analysis.

[5]The maximum absolute value for a root found in all our considered predictors is $\rho_{max} = 1.0259$.

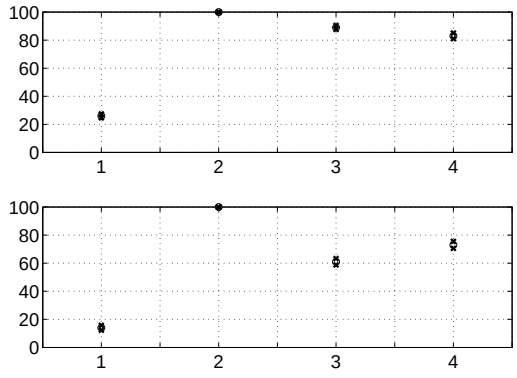


Fig. 10. MUSHRA test results. In the box above we show the results for clean speech and in the box below for speech corrupted by white noise (SNR=10dB). The four versions of the clips appear in the following order: Anchor, Hidden reference, **AMR102**, and **SpLP**. The anchor is the NATO standard 2400 bps LPC coding [49]. A 95% confidence interval is given for each value (upper and lower star).

Theorems exist to determine the maximum absolute value of the roots of a monic polynomial given the norm operator used in the minimization [43] but the bounds are generally too high to gain any real insight on how to create a intrinsic stable minimization problem, as done in [50].

The stability problem in (7) was already tackled in [9] by introducing the Burg method for prediction parameters estimation based on the least absolute forward-backward error. In this approach, however, the sparsity is not preserved. This is mostly due to the decoupling of the main K-dimensional minimization problem in K one-dimensional minimization subproblems. Therefore this method is suboptimal and produces results, as we have observed, somewhere in between those of the 2-norm and 1-norm approach. Also, the approach is only valid in (7) and not in all the other minimization schemes presented.

B. Computational Cost

As for the computational cost, finding the solution of the overdetermined system of equations in (7) using a modern interior point algorithm [19] can be shown to be equivalent to solving around 20-30 least square problems. Nevertheless, implementing this procedure in an AbS coder, as done in Section VI-A, is shown to greatly simplify the search for the sparse approximation of the residual in a closed-loop configuration, without compromising the overall quality. Furthermore, in the case of (9), the advantage is that a one step approach is taken to calculate both the short-term and the long-term predictors while the encoding of the residual is facilitated by its sparse characteristics.

The introduction of a compressed sensing formulation for the prediction problem has helped reduce dramatically the computational costs. An example of this can be seen in the coding scheme presented in VI-B. Retrieving $T = 20$ samples reduces the number of constraints of the minimization problem from 270 ($N + K$) to 80 ($M = 4T$). Since for each constraint we have a dual variable, by reducing the number of the constraints we also reduce the number of the dual variables [18]. In turn, the whole coding scheme, as shown empirically, is only about one order of magnitude more expensive than a 2-norm LP based coder, although with added improvements such as noise robustness and a fairly high conceptual simplicity.

C. Uniqueness

The minimization problems considered do not necessarily have a unique solution. In these rare cases with multiple solutions, due to the convexity of the cost function, we can immediately state that all the possible multiple solutions will still be optimal [18]. Viewing the non-uniqueness of the solution as a weakness is also arguable: in the set of possible optimal solutions we can probably find one solution that offers better properties for our modeling or coding purposes. A theorem to verify uniqueness is discussed in [52].

D. Frequency Domain Interpretation

The standard linear prediction method exhibits spectral matching properties in the frequency domain due to Parseval's theorem [2]:

$$\sum_{n=-\infty}^{\infty} |e(n)|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(e^{j\omega})|^2 d\omega. \tag{22}$$

It is also interesting to note that minimizing the squared error in the time domain and in the frequency domain leads to the same set of equations, namely the Yule-Walker equations [25]. To the best of our knowledge, the only relation existing between the time and frequency domain error using the 1-norm is the trivial Hausdorff-Young inequality [53]:

$$\sum_{n=-\infty}^{\infty} |e(n)| < \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(e^{j\omega})| d\omega, \tag{23}$$

which implies that time domain minimization does not corresponds to frequency domain minimization. It is therefore difficult to say if the 1-norm based approach is always advantageous compared to the 2-norm based approach for spectral modeling, since the statistical character of the frequency errors is not clear. However, the numerical results in Tables II, III and IV clearly show better spectral modeling properties of the sparse formulation.

VIII. Conclusions

In this paper, we have given an overview of several linear predictors for speech analysis and coding obtained by introducing sparsity into the linear prediction framework. In speech analysis, the sparse linear predictors have been shown to provide a more efficient decoupling between the pitch harmonics and the spectral envelope. This translates into predictors that are not corrupted by the fine structure of the pitch excitation and offer interesting properties such as shift invariance and pitch invariance. In the context of speech coding, the sparsity of residual and of the high order predictor provides a more synergistic new approach to encode a speech segment. The sparse residual obtained allows a more compact representation, while the sparse high order predictor engenders

joint estimation of short-term and long-term predictors. A compressed sensing formulation is used to reduce the size of the minimization problem, and hence to keep the computational costs reasonable. The sparse linear prediction based robust encoding technique provided a competitive approach to speech coding with a synergistic multistage approach and a slower decaying quality for decreasing SNR.

ACKNOWLEDGMENT

The authors would like to thank Dr. Tobias Lindstrøm Jensen (Aalborg University), Dr. Shaminda Subasingha (University of Miami) and L. Anders Ekman (Royal Institute of Technology, Stockholm) for providing part of the code used in the evaluation procedures as well as useful suggestions.

REFERENCES

[1] J. H. L. Hansen, J. G. Proakis, and J. R. Deller, Jr., *Discrete-Time Processing of Speech Signals*, Prentice-Hall, 1987.

[2] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[3] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method,"*Rep. 6th Int. Congr. Acoustics* pp. C17–C20, Paper C-5-5, 1968.

[4] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," *IEEE Trans. Signal Processing*, vol. 39, pp. 411–423, 1991.

[5] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech and Audio Processing*, vol. 8, pp. 221–239, 2000.

[6] L. A. Ekman, W. B. Kleijn and M. N. Murthi, "Regularized Linear Prediction of Speech," *IEEE Trans. Audio, Speech, Language Processing*, vol. 16, no. 1, pp. 65–73, 2008.

[7] H. Hermansky, H. Fujisaki, Y. Sato, "Spectral envelope sampling and interpolation in linear predictive analysis of speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 9, pp. 53–56, 1984.

[8] C.-H. Lee, "On Robust Linear Prediction of Speech," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 36, No. 5, pp. 642–650, 1988.

[9] E. Denoël and J.-P. Solvay, "Linear prediction of speech with a least absolute error criterion," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 6, pp. 1397–1403, 1985.

[10] J. Schroeder, R. Yarlagadda, "Linear predictive spectral estimation via the L_1 norm," *Signal Processing*, Vol. 17, No. 1, pp. 19–29, 1989.

[11] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1993.

[12] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 7, pp. 614–617, 1982.

[13] P. Kroon, E. D. F. Deprettere, and R. J. Sluyter, "Regular-pulse excitation - a novel approach to effective multipulse coding of speech", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp. 1054–1063, 1986.

[14] W. C. Chu, *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*, Wiley, 2003

[15] J. Lansford and R. Yarlagadda, "Adaptive L_p approach to speech coding," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp 335–338, 1988.

[16] M. N. Murthi and B. D. Rao, "Towards a synergistic multistage speech coder," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp 369–372, 1998.

[17] P. Kabal and R. P. Ramachandran, "Joint Optimization of Linear Predictors in Speech Coders," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37(5), pp. 642–650,May 1989.

[18] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[19] S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, 1997.

[20] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Sparse Linear Predictors for Speech Processing," *Proc. Interspeech*, pp. 1353–1356, 2008.

[21] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Joint estimation of short-term and long-term predictors in speech coders," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 4109–4112, 2009.

[22] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Speech Coding Based on Sparse Linear Prediction," *Proc. European Signal Proc. Conf.*, pp. 2524–2528, 2009.

[23] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Enhancing sparsity in linear prediction of speech by iteratively reweighted 1-norm minimization," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2010.

[24] D. L. Donoho, "Compressed sensing," *IEEE Trans. on Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[25] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Pearson Prentice Hall, 2005.

[26] S. Nadarajah, "A generalized normal distribution," *Journal of Applied Statistics*, vol. 32, no. 7, pp. 685–694, 2005.

[27] D. L. Donoho and M. Elad, "Optimally sparse representation from overcomplete dictionaries via ℓ^1-norm minimization," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, no. 5, pp. 2197–2202, 2002.

[28] E. J. Candés, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[29] J. A. Cadzow, "Minimum ℓ_1, ℓ_1, and ℓ_∞ norm approximate solutions to an overdetermined system of linear equations," *Digital Signal Processing*, vol. 12, no. 4, pp. 524–560, 2002.

[30] P. Stoica and T. Söderström, "High Order Yule-Walker equations for estimating sinusoidal frequencies: the complete set of solutions," *Signal Processing*, vol. 20, pp. 257–263, 1990.

[31] D. Giacobello, T. van Waterschoot, M. G. Christensen, S. H. Jensen, M. Moonen, "High-Order Sparse Linear Predictors for Audio Processing," in *Proc. European Signal Proc. Conf.*, pp. 234–238, 2010.

[32] J. J. Fuchs, "On sparse representations in arbitrary redundant bases," in *IEEE Trans. on Inf. Theory*, vol. 50, no. 6, 2004.

[33] P. C. Hansen and D. P. O'Leary, "The use of the L-curve in the regularization of discrete ill-posed problems," *SIAM Journal on Scientific Computing*, vol. 14, no. 6, pp. 1487–1503, 1993.

[34] D. Wipf, S. Nagarajan, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 317–329, 2010.

[35] E. J. Candés, and M. B. Wakin, "An introduction to compressive sampling," *IEEE Sig. Proc. Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[36] T. V. Sreenivas and W. B. Kleijn, "Compressive sensing for sparsely excited speech signals," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 4125–4128, 2009.

[37] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Retrieving sparse patterns using a compressed sensing framework: applications to speech coding based on sparse linear prediction," *IEEE Sig. Proc. Letters*, vol. 17, no. 1, pp. 103–106, 2010.

[38] P. Kroon and W. B. Kleijn, "Linear-prediction based analysis-by-synthesis coding", in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal Eds., Elsevier Science B.V., ch. 3, pp. 79–119, 1995.

[39] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A Simple Proof of the Restricted Isometry Property for Random Matrices,", *Constructive Approximation*, vol. 28, no, 3, pp. 253–263, 2008.

[40] L. Scharf, *Statistical Signal Processing*, Addison-Wesley, 1991.

[41] M. G. Christensen, J. Østergaard, and S. H. Jensen, "On compressed sensing and its applications to speech and audio signals," in *Rec. Asilomar Conf. Sig., Sys., and Comp.*, 2009.

[42] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mapping into Hilbert space," *Conf. in modern analysis and probability*, vol. 26, pp. 189–206, 1984.

[43] L. Knockaert, "Stability of linear predictors and numerical range of shift operators in normed spaces," *IEEE Trans. on Inf. Theory*, vol. 38, no. 5, pp. 1483–1486, 1992.

[44] A. D. Subramaniam, B. D. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies", *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 2, 2003.

[45] K. K. Paliwal, B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 1, pp 3–14, 1993.

[46] W. B. Kleijn and A. Ozerov, "Rate Distribution Between Model and Signal," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 243–246, 2007.

[47] "Adaptive Multi-Rate (AMR) speech codec; Transcoding functions," 3GPP TS 26.190, 2004.

[48] Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems, ITU-R BS.1534-1 2003.

[49] NATO (unclassified), "Parameters and coding characteristics that must be common to assure interoperability of 2400 bps linear predictive encoded digital speech," Annex X to AC/302 (NBDS) R/2.

[50] C. Magi, J. Pohjalainen, T. Bäckström and P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 401–411, 2009.
[51] W. F. G. Mecklenbrauker, "Remarks on the minimum phase property of optimal prediction error filters and some related questions," *IEEE Sig. Proc. Letters*, vol. 5, no. 4, pp. 87–88, 1998.
[52] P. Bloomfield and W. Steiger, "Least absolute deviations curve-fitting", *SIAM J. on Scientific and Statistical Computing*, vol. 1, no. 2, pp. 290–301, 1980.
[53] M. Reed and B. Simon, *Methods of Modern Mathematical Physics II: Fourier Analysis, Self-adjointness*, Academic Press, 1975.

Daniele Giacobello (S'06, M'10) was born in Milan, Italy in 1981. He completed each of his Telecommunications Engineering degrees, Laurea (B.Sc.), 2003 and Laurea Specialistica (M.Sc., with distinction), 2006 at Politecnico di Milano, Italy. In 2007, as a recipient of the EST-SIGNAL Marie Curie Fellowship, he joined the Department of Electronic Systems at Aalborg Universitet, Denmark, where he received his Ph.D. degree in Electrical and Electronic Engineering in 2010. During his graduate studies, he has been a visiting scholar at Delft University of Technology, University of Miami, and Katholieke Universiteit Leuven. As a professional, he was an integral part of research and development teams at Nokia Siemens Networks and Asahi-Kasei Corporation. He is now a Staff Scientist in the office of the CTO at Broadcom Corporation.

Dr. Giacobello research interests include digital signal processing theory and methods with applications to speech and audio signals, in particular sparse representation, statistical modeling, coding, and recognition. He is a reviewer of the Elsevier Signal Processing Journal, the IEEE Signal Processing Letters, the IEEE Journal of Selected Topics in Signal Processing, the IEEE Transactions on Speech, Audio, and Language Processing, the EURASIP Journal on Advances in Signal Processing, and the European Signal Processing Conference.

Mads Græsbøll Christensen (S'00, M'05, SM'11) was born in Copenhagen, Denmark, in March 1977. He received the M.Sc. and Ph.D. degrees from Aalborg University, Denmark, in 2002 and 2005, respectively.

He was formerly with the Department of Electronic Systems, Aalborg University, and is currently an Associate Professor in the Department of Architecture, Design and Media Technology. He has been a Visiting Researcher at Philips Research Labs, Ecole Nationale Supérieure des Télécommunications (ENST), University of California, Santa Barbara (UCSB), and Columbia University. He has published about 100 papers in peer-reviewed conference proceedings and journals and is coauthor (with A. Jakobsson) of the book *Multi-Pitch Estimation* (Morgan & Claypool Publishers, 2009). His research interests include digital signal processing theory and methods with application to speech and audio, in particular parametric analysis, modeling, and coding.

Dr. Christensen has received several awards, namely an IEEE International Conference on Acoustics, Speech and Signal Processing Student Paper Contest Award, the Spar Nord Foundation's Research Prize for his Ph.D. dissertation, and a Danish Independent Research Councils Young Researcher's Award. He is an Associate Editor for the IEEE Signal Processing Letters.

Manohar N. Murthi received his B.S. degree in electrical engineering and computer science from the University of California, Berkeley, in 1990, and his M.S. and Ph.D. degrees in electrical engineering (communication theory and systems) from the University of California, San Diego, CA, in 1992 and 1999, respectively.

He has previously worked at Qualcomm in San Diego, CA, KTH (Royal Institute of Technology), Stockholm, Sweden, and Global IP Sound in San Francisco, CA. In September 2002 he joined the Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL, where he is an Associate Professor. His research interests are in the general areas of signal and data modeling, compression, fusion and learning, and networking. He is a recipient of a National Science Foundation CAREER Award.

Søren Holdt Jensen (S'87, M'88, SM'00) received the M.Sc. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 1988, and the Ph.D. degree in signal processing from the Technical University of Denmark, Lyngby, Denmark, in 1995. Before joining the Department of Electronic Systems, Aalborg University, he was with the Telecommunications Laboratory of Telecom Denmark, Ltd., Copenhagen, Denmark; the Electronics Institute of the Technical University of Denmark; the Scientific Computing Group, Danish Computing Center for Research and Education, Lyngby; the Electrical Engineering Department, Katholieke Universiteit Leuven, Leuven, Belgium; and the Center for PersonKommunikation (CPK), Aalborg University. He is a Full Professor and is currently heading a research team working in the area of numerical algorithms and signal processing for speech and audio processing, image and video processing, multimedia technologies, and digital communications.

Dr. Jensen was an Associate Editor for the IEEE Transactions on Signal Processing and is currently Member of the Editorial Board of Elsevier Signal Processing and the EURASIP Journal on Advances in Signal Processing. He is a recipient of an European Community Marie Curie Fellowship, former Chairman of the IEEE Denmark Section, and Founder and Chairman of the IEEE DenmarkSections Signal Processing Chapter.

Marc Moonen (M'94, SM'06, F'07) is a Full Professor at the Electrical Engineering Department of Katholieke Universiteit Leuven, where he is heading a research team working in the area of numerical algorithms and signal processing for digital communications, wireless communications, DSL and audio signal processing.

He received the 1994 K.U.Leuven Research Council Award, the 1997 Alcatel Bell (Belgium) Award (with Piet Vandaele), the 2004 Alcatel Bell (Belgium) Award (with Raphael Cendrillon), and was a 1997 "Laureate of the Belgium Royal Academy of Science". He received a journal best paper award from the IEEE Transactions on Signal Processing (with Geert Leus) and from Elsevier Signal Processing (with Simon Doclo).

He was chairman of the IEEE Benelux Signal Processing Chapter (1998-2002), and is currently President of EURASIP (European Association for Signal Processing).

He has served as Editor-in-Chief for the "EURASIP Journal on Applied Signal Processing" (2003-2005), and has been a member of the editorial board of "IEEE Transactions on Circuits and Systems II" (2002-2003) and "IEEE Signal Processing Magazine" (2003-2005) and "Integration, the VLSI Journal". He is currently a member of the editorial board of "EURASIP Journal on Applied Signal Processing", "EURASIP Journal on Wireless Communications and Networking", and "Signal Processing".

Paper G

Joint Spatio-Temporal Filtering Methods for DOA and Fundamental Frequency Estimation

J. R. Jensen, **M. G. Christensen**, J. Benesty and S. H. Jensen

The paper has been published in the
IEEE/ACM Transactions on Audio, Speech, and Language Processing,
vol. 23, no. 1, pp. 174-–185, January 2015.

Joint Spatio-Temporal Filtering Methods for DOA and Fundamental Frequency Estimation

Jesper Rindom Jensen*, *Member, IEEE,* Mads Græsbøll Christensen, *Senior Member, IEEE,* Jacob Benesty, and Søren Holdt Jensen, *Senior Member, IEEE*

***Abstract*—In this paper, spatio-temporal filtering methods are proposed for estimating the direction-of-arrival (DOA) and fundamental frequency of periodic signals, like those produced by the speech production system and many musical instruments using microphone arrays. This topic has quite recently received some attention in the community and is quite promising for several applications. The proposed methods are based on optimal, adaptive filters that leave the desired signal, having a certain DOA and fundamental frequency, undistorted and suppress everything else. The filtering methods simultaneously operate in space and time, whereby it is possible resolve cases that are otherwise problematic for pitch estimators or DOA estimators based on beamforming. Several special cases and improvements are considered, including a method for estimating the covariance matrix based on the recently proposed iterative adaptive approach (IAA). Experiments demonstrate the improved performance of the proposed methods under adverse conditions compared to the state of the art using both synthetic signals and real signals, as well as illustrate the properties of the methods and the filters.**

***Index Terms*—Fundamental frequency estimation, DOA estimation, joint estimation, 2-D filtering, LCMV beamformer, periodogram-based beamformer**

I. Introduction

A fundamental property of speech and audio signals is the so-called pitch. For many signals, namely periodic signals, the pitch is equivalent to the fundamental frequency, i.e., the frequency of which integer multiples form the frequencies of the individual harmonics, even though there exists some pathological examples where it is not quite that simple. In some applications, the pitch itself is of interest or is being studied for other purposes, some examples being prosody analysis and transcription of music. The pitch also often forms the basis of the processing of such signals. Some well-known examples include speech coding, wherein long-term predictors are used to exploit the correlation caused by the quasi-periodicity that causes the pitch, and noise reduction, wherein the pitch can be used to either directly enhance the signal of interest [1] or to estimate the properties of the noise [2]. Filters that extract or attenuate the harmonics of periodic sounds are often referred to as comb filters, due to their characteristic frequency response. Such comb filters have played a prominent role in the history of signal processing, dating back to 1970's [3], and new forms of comb filters keep emerging. The classical comb filter is based on signal-independent FIR or IIR filters with poles or zeros, respectively, close to the unit circle at the harmonic frequencies. Later, it was shown that more efficient filters can be obtained via a set of notch or peak filters [4], and a few other examples of such approaches can be found in [5] and the references therein. More recently, it was shown that by generalizing the principle of the Capon spectral estimator, it is possible to design optimal, adaptive FIR comb filters [1], [6]. These filters have a number of properties that make them desirable in several applications. The filters are distortionless, i.e., they let the signal of interest, i.e., periodic signals, pass undistorted. They are adaptive and, hence, automatically adapt to the conditions under which the signal of interest has been recorded. This means that they can cancel strong interferences, including also other periodic signals, without prior knowledge of their properties. The filters also, curiously, reduce to evaluating Fourier transforms at certain frequencies or projecting onto the space spanned by Fourier bases under certain conditions.

In microphone arrays, the direction-of-arrival (DOA) is often used as a means of locating, tracking and separating signals, something that is often done using spatial filters, i.e., beamformers [7], [8]. Since speech and audio signals are generally broadband, unlike, for example, communication and radar signals, many of the clever narrowband beamforming techniques cannot be applied directly to such signals. Instead, speech and audio signals are often decomposed into a set of subbands, each of which are then processed as narrowband signals. However, periodic signals can be modeled efficiently using the harmonic model [9], in which the signal of interest is modeled as a set of narrowband signals, namely sinusoids corresponding to the individual harmonics. This means that such signals can in fact be treated as multiple narrowband signals that share some common parameters: the fundamental frequency and the DOA. In fact, by finding, jointly, both the fundamental frequency (i.e., the pitch) and the DOA, it is possible to mitigate some of the severe problems that pitch estimators encounter for multiple sources, and it is possible to overcome some of the problems that DOA estimators have with distinguishing between different sources when these

*J. R. Jensen is the corresponding author.

This research was funded by the Villum Foundation and the Danish Council for Independent Research, grant ID: DFF – 1337-00084.

J. R. Jensen and M. G. Christensen are with the Audio Analysis Lab, AD:MT, Aalborg University, DK-9220 Aalborg, Denmark, e-mail: {jrj, mgc}@create.aau.dk.

J. Benesty is with the INRS-EMT, University of Quebec, 800 de la Gauchetiere Ouest, Suite 6900, Montreal, QC H5A 1K6, Canada, e-mail: benesty@emt.inrs.ca and with the Audio Analysis Lab, AD:MT, Aalborg University, Denmark.

S. H. Jensen is with the Dept. of Electronic Systems, Aalborg University, DK-9220 Aalborg, Denmark, e-mail: shj@es.aau.dk.

impinge on the array from angles that are close. It should also be noted that the DOA along with the pitch also are believed to be some of the governing factors that the human auditory system uses for separating sources. This line of reasoning has, quite recently, led to some joint DOA and fundamental frequency estimators, including maximum likelihood based [10], [11], subspace-based [12]–[14], correlation-based [15], [16], and filtering-based [17]–[19] methods. Notably, the problem of joint DOA and fundamental frequency estimation was formalized and thoroughly analyzed in [10], and a maximum likelihood estimator that achieves the highest possible accuracy (under certain conditions) was proposed.

In this paper, we propose spatio-temporal filtering methods for joint DOA and fundamental frequency estimation for periodic signals, like, for example, speech signals or signals produced by musical instruments. The filters are based on the principle of the Capon and Frost beamformers [20], [21] and spectral estimators combined and generalized to account for the nature of periodic signals, and they are controlled by two parameters: the DOA and the fundamental frequency. The proposed filters are optimal and adaptive, and should, hence, be capable of dealing with adverse conditions, like when strong background noise or interference is present but guarantee that the signal of interest is left undistorted. The filters can be thought of as jointly performing beamforming and enhancement, i.e., they are spatio-temporal. In this paper, however, we consider only the application of these filters to parameter estimation, i.e., estimation of the DOA and the fundamental frequency. We consider various variations and simplifications of the filters, including optimal filters for white noise and for infinitely long filters. We also consider the application of the principle of the iterative adaptive approach (IAA) [19], [22], [23] for finding the covariance matrix, which is required to compute the optimal filters. This can be used to obtain longer filters for a given number of samples, something that often results in an improved estimation of parameters, especially under adverse conditions. While the IAA based estimators are computationally more complex than using the traditional sample covariance matrix estimate, it has been shown [24], [25] that the computational complexity of the IAA can be reduced dramatically.

The rest of the paper is organized as follows: In Section II, we introduce the problem formulation along with some useful notation, and we proceed to motivate the usage of joint DOA and fundamental frequency estimation in more detail. We then, in Section III, introduce the filter designs and consider, as mentioned, various special cases and the IAA method for estimation of the covariance matrix. In Section IV, the experimental results are presented, after which the conclusion follows in Section V.

II. Problem Statement

Consider a scenario where K microphones are recording a mixture of a desired, noise, and interfering sources. At time instance n, we can then model the signal observed using the k'th microphone as

$$y_k(n) = x_k(n) + v_k(n), \tag{1}$$

where $x_k(n)$ is the recording of the desired source, and $v_k(n)$ is the sum of the recorded noise and interference. In this paper, we assume that the desired signal is periodic, which is a reasonable assumption for, e.g., voiced speech and many musical instruments. The noise can, for example, be background noise such as sensor noise, whereas the interference covers other periodic signal not being of interest. Utilizing the periodicity assumption and by exploiting that the desired signal observations across the microphones are just delayed and attenuated version of each other, the signal model can be further specified as

$$\begin{aligned} y_k(n) &= \beta_k s(n - f_\mathrm{s}\tau_k) + v_k(n) \\ &= \beta_k \sum_{l=1}^{L} \alpha_l e^{jl\omega_0(n-f_\mathrm{s}\tau_k)} + v_k(n), \end{aligned} \tag{2}$$

with L being the number of harmonics, $\alpha_l = A_l e^{j\phi_l}$ being the complex amplitude of the l'th harmonic with A_l and ϕ_l denoting the positive real amplitude and phase, respectively, ω_0 is the fundamental frequency, f_s is the sampling frequency, τ_k is the delay of the desired signal from microphone 0 to microphone k, and β_k is the attenuation of the desired signal at sensor k. Note that, by using this model, we have implicitly assumed no reverberation. When the array of microphones is organized in a known way, we can also model the time delay τ_k. For example, if the microphones are organized in a uniform linear array structure, we have that

$$\tau_k = k\frac{d\sin\theta}{c}, \tag{3}$$

where d is the inter microphone spacing, θ is the direction-of-arrival (DOA), and c is the wave propagation speed. That is,

$$y_k(n) = \beta_k \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n} e^{-jl\omega_\mathrm{s} k} + v_k(n) \tag{4}$$

with

$$\omega_\mathrm{s} = \omega_0 f_\mathrm{s} \tau_1 \tag{5}$$

being the so-called spatial frequency. In the remainder of the paper, we assume, for simplicity, that $\beta_p = \beta_q = 1$ for $p \neq q$, which is a reasonable assumption for arrays with closely spaced microphones. When this assumption does not hold, the βs can be estimated using, e.g., the techniques presented in [26].

In practice, N time-consecutive samples from each microphone are used for the estimation of the pitch and DOA. These data can be organized in a matrix like

$$\mathbf{Y}(n) = \begin{bmatrix} y_0(n) & \cdots & y_0(n-N+1) \\ \vdots & \ddots & \vdots \\ y_{K-1}(n) & \cdots & y_{K-1}(n-N+1) \end{bmatrix}. \tag{6}$$

If we consider a subblock of $M \times P$ samples from the above matrix, which is useful for the filter designs to follow later,

we can write the signal model on vector form as

$$\mathbf{Y}_k(n) = \begin{bmatrix} y_k(n) & \cdots & y_k(n-M+1) \\ \vdots & \ddots & \vdots \\ y_{k+P-1}(n) & \cdots & y_{k+P-1}(n-M+1) \end{bmatrix} \quad (7)$$
$$= \sum_{l=1}^{L} \alpha_l(n)\mathbf{z}_{\mathrm{s}}(l\omega_{\mathrm{s}})\mathbf{z}_{\mathrm{t}}^T(l\omega_0) + \mathbf{V}_k(n),$$

where $\alpha_l(n) = e^{jl\omega_0 n}$, and

$$\mathbf{z}_{\mathrm{s}}(l\omega_{\mathrm{s}}) = \begin{bmatrix} 1 & e^{-jl\omega_{\mathrm{s}}} & \cdots & e^{-j(P-1)l\omega_{\mathrm{s}}} \end{bmatrix}^T, \quad (8)$$

$$\mathbf{z}_{\mathrm{t}}(l\omega_0) = \begin{bmatrix} 1 & e^{-jl\omega_0} & \cdots & e^{-j(M-1)l\omega_0} \end{bmatrix}^T, \quad (9)$$

$$\mathbf{V}_k(n) = \begin{bmatrix} v_k(n) & \cdots & v_k(n-M+1) \\ \vdots & \ddots & \vdots \\ v_{k+P-1}(n) & \cdots & v_{k+P-1}(n-M+1) \end{bmatrix}. \quad (10)$$

In the optimal filter designs considered in Sec. III, it is useful to stack the columns of the subblocks of the observed signal matrix (denoted $\mathrm{vec}\{\cdot\}$), which yields

$$\mathbf{y}_k(n) = \mathrm{vec}\{\mathbf{Y}_k(n)\} = \sum_{l=1}^{L} \alpha_l(n)\mathbf{z}_l + \mathbf{v}_k(n), \quad (11)$$

with $\mathbf{v}_k(n) = \mathrm{vec}\{\mathbf{V}_k(n)\}$, and

$$\mathbf{z}_l = \mathrm{vec}\{\mathbf{z}_{\mathrm{s}}(l\omega_{\mathrm{s}})\mathbf{z}_{\mathrm{t}}^T(l\omega_0)\} = \mathbf{z}_{\mathrm{s}}(l\omega_{\mathrm{s}}) \otimes \mathbf{z}_{\mathrm{t}}(l\omega_0), \quad (12)$$

where $\otimes$ denotes the Kronecker product of two vectors or matrices.

A. Motivation for Joint Estimation

Instead of estimating the DOA and pitch jointly, we could estimate those parameters separately with a much lower computational complexity. However, there are a number of significant benefits by conducting the estimation jointly. First of all, in scenarios where multiple periodic sources are present simultaneously, joint estimators may be able to resolve those sources even if either the pitch frequencies or the DOAs of one or more of those sources are similar. This would be impossible if the parameters are estimated separately, since the search is here in only one dimension. Another benefit is a potentially higher estimation accuracy. In [10], it was shown that the asymptotic Cramér-Rao bounds (CRBs) for the DOA and pitch are given by

$$\mathrm{CRB}(\omega_0) \approx \frac{6}{N^3 K}\mathrm{PSNR}^{-1}, \quad (13)$$

$$\mathrm{CRB}(\theta) \approx \left[\left(\frac{c}{\omega_0 f_{\mathrm{s}} d\cos\theta}\right)^2 \frac{6}{NK^3} + \left(\frac{\tan\theta}{\omega_0}\right)\frac{6}{K^3 N}\right]\mathrm{PSNR}^{-1} \quad (14)$$

for the scenario described by (11) when $v(n)$ is white noise and $\beta_p \approx \beta_q$ for $p \neq q$, with PSNR denoting the pseudo signal-to-noise ratio. The PSNR is defined as

$$\mathrm{PSNR} = \frac{\sum_{l=1}^{L} l^2 A_l^2}{\sigma_v^2}, \quad (15)$$

and σ_v^2 is the variance of the noise. Close investigation of these expressions reveals the fact that the CRB of the pitch decreases cubically and linearly for increasing N's and K's, respectively. In other words, the pitch estimate can be more accurate when multiple microphone recordings are used. Moreover, we can see that the DOA can be estimated more accurately when taking the harmonic structure of the periodic signal into account as opposed to if the DOA was estimated from, e.g., just the fundamental tone.

Another way of estimating the DOA and pitch is to use a cascaded approach where the DOA is first estimated from the multiple microphone recordings. Then, the signal impinging from this direction is extracted using a beamformer, whereupon the pitch is estimated from the beamformer output. This traditional and cascaded way of estimating the parameters will most likely increase the CRB of the parameter estimated in the second step of this procedure. The cause of this increase, is the linear transformation of the spatio-temporal data introduced by the signal extraction after estimation of the first parameter [10].

III. Spatio-Temporal Filtering Methods

In this section, we present filtering methods for joint estimation of the DOA and pitch from noisy, spatio-temporal, observed data that can be modeled by (11). We assume that we have sampled a signal N times in time and using K sensors in space, which gives us the $K \times N$ data matrix $\mathbf{Y}(n)$ in (6). Then, based on these data, we can design optimal filterbanks or filters for estimating the aforementioned parameters. In all of the presented filtering methods, the idea is to design a filterbank or filter that has minimum output power, while it passes the desired signal undistorted. The joint parameter estimates can then be obtained by maximizing the output power of the so-obtained optimal filters.

A. Optimal Filterbanks

In the filterbank approach, the idea is to design a bank of L FIR filters, where the lth filter should pass the lth harmonic of the desired, periodic signal undistorted. Applying such a bank of FIR filters on a block of the observed signal, we get

$$z_k(n) = \sum_{l=1}^{L}\sum_{m=0}^{M-1}\sum_{p=0}^{P-1} h_{mp}^l y_{k+p}(n-m) \quad (16)$$
$$= \sum_{l=1}^{L} \mathbf{1}^T\left[\mathbf{H}_l \circ \mathbf{Y}_k(n)\right]\mathbf{1}$$
$$= \sum_{l=1}^{L} \mathbf{h}_l^H \mathbf{y}_k(n),$$

where $\circ$ denotes the Hadamard product, M and P are the temporal and spatial filter lengths, respectively, h_{mp}^l is the (m,p)th coefficient of the lth filter in the filterbank, $\mathbf{h}_l = \mathrm{vec}\{\mathbf{H}_l\}$, $\mathbf{y}_k(n) = \mathrm{vec}\{\mathbf{Y}_k(n)\}$, $\mathbf{1}$ is a column vector of ones,

and

$$\mathbf{H}_l = \begin{bmatrix} h_{00}^l & \cdots & h_{(M-1)0}^l \\ \vdots & \ddots & \vdots \\ h_{0(P-1)}^l & \cdots & h_{(M-1)(P-1)}^l \end{bmatrix}. \quad (17)$$

Then, we can design a filterbank where the sum of the output powers from the individual filters is minimized, while the lth filter passes the lth harmonic undistorted and cancels out the other harmonics. The sum of the output powers of the filters is given by

$$\sum_{l=1}^{L} \mathrm{E}\left[\left|\mathbf{h}_l^H \mathbf{y}_k(n)\right|^2\right] = \sum_{l=1}^{L} \mathbf{h}_l^H \mathbf{R}_{\mathbf{y}} \mathbf{h}_l, \quad (18)$$

where $\mathbf{R}_{\mathbf{y}} = \mathrm{E}[\mathbf{y}_k(n)\mathbf{y}_k^H(n)]$ is the covariance matrix of $\mathbf{y}_k(n)$. From (11) and (18), it is clear that the aforementioned design goal can be achieved by solving the following optimization problem:

$$\min_{\mathbf{H}} \mathrm{Tr}[\mathbf{H}^H \mathbf{R}_{\mathbf{y}} \mathbf{H}] \quad \text{s.t.} \quad \mathbf{H}^H \mathbf{Z} = \mathbf{I}, \quad (19)$$

with

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \cdots & \mathbf{z}_L \end{bmatrix}. \quad (20)$$

The well known solution to this second order optimization problem can obtained using Lagrange multipliers, and it is given by

$$\mathbf{H}_{\text{opt}} = \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}\left(\mathbf{Z}^H \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}\right)^{-1}. \quad (21)$$

Note that for $\mathbf{Z}^H\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}$ to be invertible, we must require that $M \geq L$, and this is also the case for the other optimal filter designs proposed in the remainder of the section. Interestingly, it can be shown that a filter identical to the one in (21) can be designed by minimizing the sum of the powers of the noise at the output of all the filters [27], which gives [18]

$$\mathbf{H}_{\text{opt}} = \mathbf{R}_{\mathbf{v}}^{-1}\mathbf{Z}\left(\mathbf{Z}^H \mathbf{R}_{\mathbf{v}}^{-1}\mathbf{Z}\right)^{-1}. \quad (22)$$

This fact can be exploited to achieve some computationally more efficient filter designs. If we, for a moment, assume that the noise is white such that $\mathbf{R}_{\mathbf{v}} = \sigma_v^2\mathbf{I}$, where σ_v^2 is the variance of the noise, we get that

$$\mathbf{H}_{\text{wn}} = \mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}. \quad (23)$$

This filter will, of course, only be optimal with respect to the aforementioned design criteria when the noise is indeed white, but it may still be useful even in other noise settings due to its simplicity. Finally, we can achieve an approximative filter design by exploiting that [9]

$$\lim_{MP\to\infty} \frac{1}{MP}\mathbf{Z}^H\mathbf{Z} = \mathbf{I}. \quad (24)$$

In this case, the optimal filter for the white noise scenario becomes

$$\mathbf{H}_{\text{awn}} = \frac{1}{MP}\mathbf{Z}. \quad (25)$$

This approximative filter design can be interpreted as a filterbank of spatio-temporal, periodogram-based filters [18], and it can be applied efficiently in practice using FFTs.

Using either of the aforementioned filter designs, the fundamental frequency and the DOA can then be estimated jointly. This is achieved by maximizing the sum of the output powers of the filters in these filterbanks over sets of candidate fundamental frequencies and DOAs, i.e.,

$$\{\widehat{\omega}_0, \widehat{\theta}\} = \arg \max_{\{\omega_0,\theta\}\in\Omega\times\Theta} \mathrm{Tr}\left[\mathbf{H}^H\mathbf{R}_{\mathbf{y}}\mathbf{H}\right], \quad (26)$$

where Ω and Θ are the sets of candidate fundamental frequencies and DOAs, respectively. We note that, in practice, $\mathbf{R}_{\mathbf{y}}$ is most likely not known and therefore has to be estimated. Moreover, it is worth mentioning that, while the above estimator is only for estimating the pitch and DOA of a single source, the estimator can be used in the iterative RELAX algorithm in [28] to estimate the parameters of multiple sources.

B. Optimal Single Filters

An alternative filtering approach to joint fundamental frequency and DOA estimation is the single filter approach. In this approach, the idea is to apply a single FIR filter on a block of the observed signal, yielding the output:

$$\begin{aligned} z_k(n) &= \sum_{m=0}^{M-1}\sum_{p=0}^{P-1} h_{mp} y_{k+p}(n-m) \quad (27)\\ &= \mathbf{1}^T[\mathbf{H} \circ \mathbf{Y}_k(n)]\mathbf{1} \\ &= \mathbf{h}^H\mathbf{y}_k(n), \end{aligned}$$

where $\mathbf{H}$ is defined similarly to $\mathbf{H}_l$, i.e.,

$$\mathbf{H} = \begin{bmatrix} h_{00} & \cdots & h_{(M-1)0} \\ \vdots & \ddots & \vdots \\ h_{0(P-1)} & \cdots & h_{(M-1)(P-1)} \end{bmatrix}, \quad (28)$$

and $\mathbf{h} = \mathrm{vec}\{\mathbf{H}\}$. We then want to design a single filter that passes all of the harmonics undistorted while the output power of the filter is minimized. The output power of the single filter is given by

$$\mathrm{E}\left[\left|\mathbf{h}^H\mathbf{y}_k(n)\right|^2\right]. \quad (29)$$

That is, a solution to the above filter design problem is, from (11), clearly achieved by solving:

$$\min_{\mathbf{h}} \mathbf{h}^H\mathbf{R}_{\mathbf{y}}\mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^H\mathbf{z}_l = 1, \quad \text{for} \quad l = 1,\ldots,L. \quad (30)$$

Like in the filterbank approach, the solution to this optimization problem can be obtained using Lagrange multipliers and is given by

$$\mathbf{h}_{\text{opt}} = \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1}. \quad (31)$$

The same filter is obtained if we minimize the power of the noise after filtering under the same constraints, in which case the optimal filter is given by [27]

$$\begin{aligned} \mathbf{h}_{\text{opt}} &= \mathbf{R}_{\mathbf{v}}^{-1}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{R}_{\mathbf{v}}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1} \quad (32)\\ &= \mathbf{H}_{\text{opt}}\mathbf{1}. \end{aligned}$$

Table I
COST FUNCTIONS INVOLVED IN THE ESTIMATORS OBTAINED USING THE DIFFERENT FILTERING APPROACHES AND DESIGN TOPOLOGIES.

	Filterbank	**Single Filter**
Optimal	$\text{Tr}\left[\left(\mathbf{Z}^H\mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}\right)^{-1}\right]$	$\mathbf{1}^T\left(\mathbf{Z}^H\mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1}$
White Noise	$\text{Tr}\left[\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{R}_\mathbf{y}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\right]$	$\mathbf{1}^T\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{R}_\mathbf{y}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{1}$
Approx.	$\frac{1}{M^2P^2}\text{Tr}\left[\mathbf{Z}^H\mathbf{R}_\mathbf{y}\mathbf{Z}\right]$	$\frac{1}{M^2P^2}\mathbf{1}^T\mathbf{Z}^H\mathbf{R}_\mathbf{y}\mathbf{Z}\mathbf{1}$

If we again assume that the noise is white, with $\mathbf{R}_\mathbf{v} = \sigma_v^2\mathbf{I}$, we get that

$$\mathbf{h}_{\text{wn}} = \mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{1} = \mathbf{H}_{\text{wn}}\mathbf{1}. \tag{33}$$

Then, by applying the approximation in (24), we can obtain an approximative single filter design as

$$\mathbf{h}_{\text{awn}} = \frac{1}{MP}\mathbf{Z}\mathbf{1} = \mathbf{H}_{\text{awn}}\mathbf{1}. \tag{34}$$

This filter can be seen as a sum of spatio-temporal, periodogram-based filters.

In the single filter approach, the fundamental frequency and DOA are then jointly estimated by simply maximizing the output power of either of the filter designs proposed above over the sets of candidate fundamental frequencies Ω and DOAs Θ. Mathematically speaking, the joint estimates obtained using the single filter approach can be put as

$$\{\widehat{\omega}_0, \widehat{\theta}\} = \arg\max_{\{\omega_0,\theta\}\in\Omega\times\Theta} \mathbf{h}^H\mathbf{R}_\mathbf{y}\mathbf{h}. \tag{35}$$

In Table I, an overview of the estimators obtained using the two different filtering approaches and the different filter design topologies is found. These have been found by first inserting the expressions in (21), (23), and (25) in (26), and then inserting (31), (33), and (34) in (35). We note that, except for a scaling factor of $\frac{1}{M^2P^2}$, the cost function for the approximative filterbank approach resembles that of the approximative NLS estimator in [10]. Furthermore, as mentioned in Section III-A, the above estimator can be used in an iterative algorithm to estimate the parameters of multiple sources [28].

When the assumptions on white Gaussian noise and large sample sizes do not hold, the white noise and approximative filtering methods will consequently not be optimal, and most often yield less accurate estimates compared to the optimal filtering methods as we also discovered in the experiments in Section IV. The approximative filtering methods, however, are computationally simpler compared to the optimal filtering methods by not requiring any inversions. That is, the filter design can be chosen to achieve a certain tradeoff between computational complexity and estimation accuracy.

C. Estimation of the Covariance Matrix

In the estimators presented in this section, knowledge about the covariance matrix $\mathbf{R}_\mathbf{y}$ is needed. This covariance matrix is obviously not known in most practical scenarios, so we need to replace it by an estimate. One possible estimate is the outer product estimate, which is commonly used, e.g., in single-channel, fundamental frequency estimation. In the multichannel, spatio-temporal case, the outer product estimate of $\mathbf{R}_\mathbf{y}$ is given by

$$\widehat{\mathbf{R}}_\mathbf{y} = \sum_{k=0}^{K-P}\sum_{m=0}^{N-M}\frac{\mathbf{y}_k(n-m)\mathbf{y}_k^H(n-m)}{(K-P+1)(N-M+1)}. \tag{36}$$

The optimal estimators for the general noise case (see Table I) require the covariance matrix of the observed signal to be inverted. To ensure that $\widehat{\mathbf{R}}_\mathbf{y}$ is invertible, we must require it to be full-rank, i.e.,

$$(K-P+1)(N-M+1) \geq MP \tag{37}$$

needs to be fulfilled. Typically, $K \ll N$ and P is desired to be as large as possible to attain a reasonable spatial resolution. If we, for example, choose $P = K$ we have that

$$M \leq \frac{N+1}{K+1}. \tag{38}$$

As a result of that, M may need to be very small or a large amount of temporal samples N is needed if K is relatively large.

Alternatively, to circumvent this issue, an iterative adaptive approach (IAA) [22], [23] on the estimation of the covariance matrix can be taken. First, let the amplitude of a spatio-temporal frequency component of interest be denoted by $\alpha_{\gamma',\psi'}$, where γ' is a frequency index, and ψ' is a direction index corresponding to the DOA. Then, using the covariance matrix model, the noise covariance matrix can be approximated as

$$\begin{aligned}\mathbf{Q}_{\gamma',\psi'} &\approx \sum_{\gamma=1}^{\Gamma}\sum_{\psi=1}^{\Psi}|\alpha_{\gamma,\psi}|^2\mathbf{z}_{\gamma,\psi}\mathbf{z}_{\gamma,\psi}^H - |\alpha_{\gamma',\psi'}|^2\mathbf{z}_{\gamma',\psi'}\mathbf{z}_{\gamma',\psi'}^H \\ &= \widetilde{\mathbf{R}}_\mathbf{y} - |\alpha_{\gamma',\psi'}|^2\mathbf{z}_{\gamma',\psi'}\mathbf{z}_{\gamma',\psi'}^H,\end{aligned} \tag{39}$$

where γ and ψ denote frequency and direction indices, respectively, Γ is the number of frequency grid points utilized

Table II
IAA FOR SPATIO-TEMPORAL SPECTRUM AND COVARIANCE ESTIMATION

initialization

$$\hat{\alpha}_{\gamma,\psi} = \frac{\mathbf{z}_{\gamma,\psi}^H\mathbf{y}(n)}{\mathbf{z}_{\gamma,\psi}^H\mathbf{z}_{\gamma,\psi}}, \quad \gamma = 1,\ldots,\Gamma, \quad \psi = 1,\ldots,\Psi$$

repeat

$$\widetilde{\mathbf{R}}_\mathbf{y} = \sum_{\gamma=1}^{\Gamma}\sum_{\psi=1}^{\Psi}|\alpha_{\gamma,\psi}|^2\mathbf{z}_{\gamma,\psi}\mathbf{z}_{\gamma,\psi}^H$$

$$\hat{\alpha}_{\gamma,\psi} = \frac{\mathbf{z}_{\gamma,\psi}^H\widetilde{\mathbf{R}}_\mathbf{y}^{-1}\mathbf{y}(n)}{\mathbf{z}_{\gamma,\psi}^H\widetilde{\mathbf{R}}_\mathbf{y}^{-1}\mathbf{z}_{\gamma,\psi}}, \quad \gamma = 1,\ldots,\Gamma, \quad \psi = 1,\ldots,\Psi$$

until (convergence)

in the IAA, Ψ is the number direction grid points utilized in the IAA,

$$\widetilde{\mathbf{R}}_{\mathbf{y}} = \sum_{\gamma=1}^{\Gamma} \sum_{\psi=1}^{\Psi} |\alpha_{\gamma,\psi}|^2 \mathbf{z}_{\gamma,\psi} \mathbf{z}_{\gamma,\psi}^H, \tag{40}$$

is an estimate of the observed signal covariance matrix, and

$$\mathbf{z}_{\gamma,\psi} = \mathbf{z}_{\mathrm{s}}(\psi,\gamma) \otimes \mathbf{z}_{\mathrm{t}}(\gamma), \tag{41}$$

$$\mathbf{z}_{\mathrm{t}}(\gamma) = \begin{bmatrix} 1 & e^{-j\omega_\gamma} & \cdots & e^{-j(N-1)\omega_\gamma} \end{bmatrix}^T, \tag{42}$$

$$\mathbf{z}_{\mathrm{s}}(\psi,\gamma) = \begin{bmatrix} 1 & e^{-j\omega_{\mathrm{s},\psi,\gamma}} & \cdots & e^{-j(K-1)\omega_{\mathrm{s},\psi,\gamma}} \end{bmatrix}^T, \tag{43}$$

with ω_γ denoting the frequency corresponding to the γ'th grid point, and $\omega_{\mathrm{s},\psi,\gamma}$ denoting the spatial frequency corresponding to the grid points ψ and γ, i.e.,

$$\omega_{\mathrm{s},\psi,\gamma} = \omega_\gamma f_{\mathrm{s}} \frac{d \sin\theta_\psi}{c}. \tag{44}$$

In (44), θ_ψ is the DOA corresponding to the ψ'th grid point.

The IAA is then used to obtain an estimate of the amplitude $\alpha_{\gamma',\psi'}$ by minimizing a weighted least-squares (WLS) cost function J_{WLS} given by

$$J_{\mathrm{WLS}} = \left[\mathbf{y}(n) - \alpha_{\gamma',\psi'}\mathbf{z}_{\gamma',\psi'}\right]^H \mathbf{Q}_{\gamma',\psi'} \left[\mathbf{y}(n) - \alpha_{\gamma',\psi'}\mathbf{z}_{\gamma',\psi'}\right], \tag{45}$$

with $\mathbf{y}(n) = \mathrm{vec}\{\mathbf{Y}(n)\}$. Minimizing the cost function with respect to the unknown amplitude $\alpha_{\gamma',\psi'}$ yields the following closed-form estimate

$$\hat{\alpha}_{\gamma',\psi'} = \frac{\mathbf{z}_{\gamma',\psi'}^H \mathbf{Q}_{\gamma',\psi'}^{-1} \mathbf{y}(n)}{\mathbf{z}_{\gamma',\psi'}^H \mathbf{Q}_{\gamma',\psi'}^{-1} \mathbf{z}_{\gamma',\psi'}}. \tag{46}$$

Using the matrix inversion lemma on (39), it can be shown that the amplitude estimate is equivalently found from

$$\hat{\alpha}_{\gamma',\psi'} = \frac{\mathbf{z}_{\gamma',\psi'}^H \widetilde{\mathbf{R}}_{\mathbf{y}}^{-1} \mathbf{y}(n)}{\mathbf{z}_{\gamma',\psi'}^H \widetilde{\mathbf{R}}_{\mathbf{y}}^{-1} \mathbf{z}_{\gamma',\psi'}}. \tag{47}$$

This expression is preferred over (46) as the covariance matrix estimate $\widetilde{\mathbf{R}}_{\mathbf{y}}$ needs to be formed only once, while $\mathbf{Q}_{\gamma',\psi'}$ needs to be updated per frequency and direction grid point. We note that the amplitude estimate depends on the estimate of the covariance matrix and vice versa, so these are estimated by iterating between (40) and (47), hence the method is termed the IAA. While the IAA has historically been used for amplitude spectrum estimation, we here utilize it for estimation of the covariance matrix of the observed signal herein. As opposed to the sample covariance matrix estimate, this estimate is formed from a single observation, $\mathbf{y}(n)$, while also being full-rank. This enables us to choose $M = N$ and $P = K$, but of course it is computationally more complex to obtain this estimate than the sample covariance matrix estimate. The algorithm is summarized in Table II. As it can be seen, the algorithm is initialized with $\widetilde{\mathbf{R}}_{\mathbf{y}} = \mathbf{I}$. Typically, 10-15 iterations is sufficient to achieve convergence in practice.

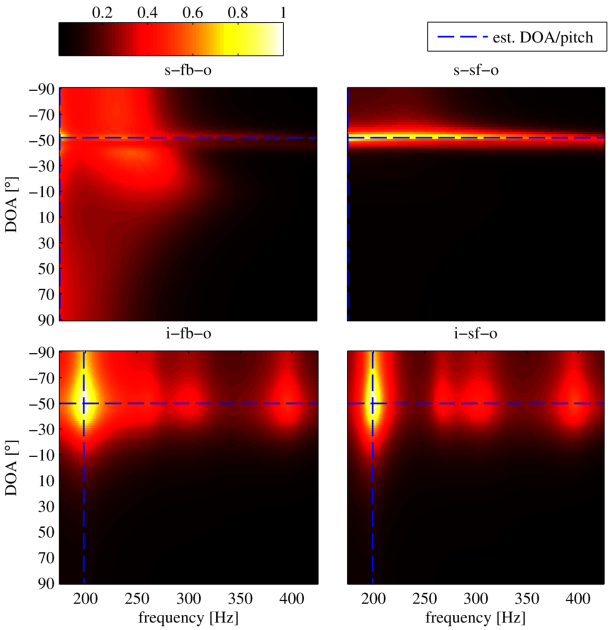


Figure 1. Plots of the cost functions for the optimal (left) filterbank and (right) single filter methods implemented using the (top) sample and (bottom) IAA-based covariance matrix estimates when applied on a synthetic, multichannel, periodic signal for $N = 20$.

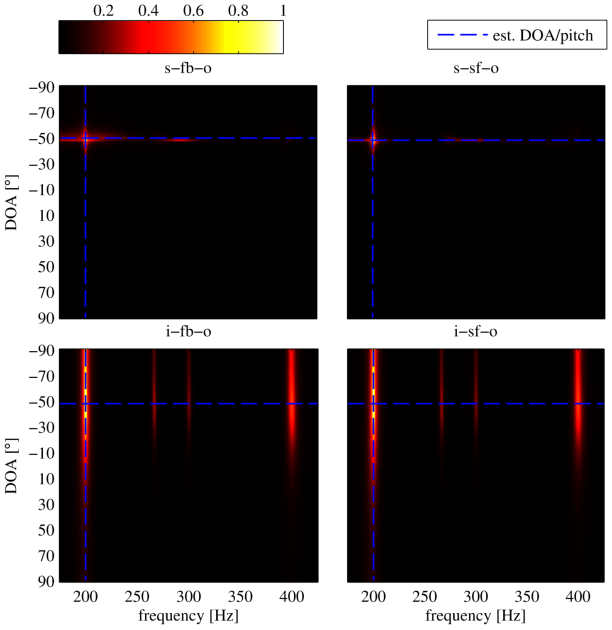


Figure 2. Plots of the cost functions for the optimal (left) filterbank and (right) single filter methods implemented using the (top) sample and (bottom) IAA-based covariance matrix estimates when applied on a synthetic, multichannel, periodic signal for $N = 60$.

IV. Experimental Results

We now proceed with an experimental evaluation of the proposed filter designs. The evaluation is split into three parts: 1) a qualitative comparison of the proposed filters, 2) a thorough statistical evaluation of the proposed filters through Monte-Carlo simulations including comparison with state of the art, and 3) qualitative evaluation of the filters on a real-life signal. First, we compare the cost functions of the optimal filterbank

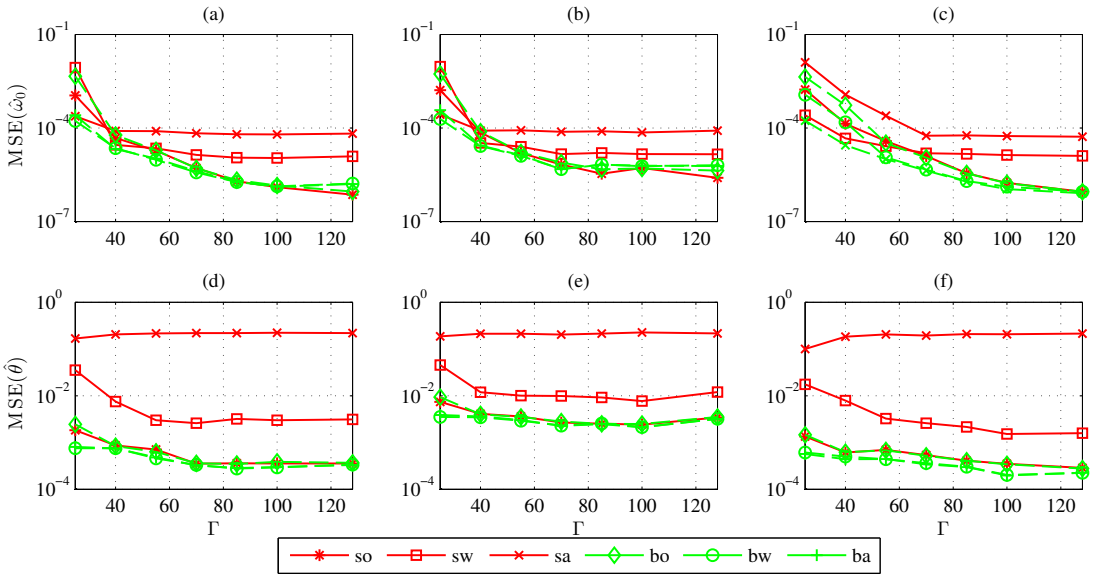


Figure 3. MSEs of pitch and DOA estimates for different Γ's in scenarios with (a,d) $N = 20$ and an SNR of 30 dB, (b,e) $N = 20$ and an SNR of 20 dB, and (c,f) $N = 25$ and an SNR of 30 dB for the proposed methods.

and single filter when using both the sample and IAA-based covariance matrix estimates. For this experiment, we used a synthetic, periodic signal with $L = 4$ harmonics with unit amplitudes, $\theta = -50°$, $f_0 = 200$ Hz, $f_s = 4$ kHz, and white noise was added to each sensor signal at an SNR of 30 dB. The other parameters of interest in the simulation were chosen as follows: $K = 3$, $P = 3$, $N = 20$, $M = \lfloor (N+1)/(K+1) \rfloor$, $\Gamma = 512$, $\Psi = 128$, 10 iterations was used to obtain the IAA estimate, $c = 343$ m/s, and $d = 0.04$ m. Using this setup, we then evaluated the cost functions of the optimal filters in Table I for different candidate pitch frequencies and DOAs when using the sample and IAA-based covariance matrix estimates, and the results are depicted in Fig. 1. From the figures, we can see that none of the methods show a distinct peak at the true DOA and pitch when the sample covariance matrix estimate is used. This is opposed to when using the IAA-based covariance matrix estimate, in which case both optimal filtering methods each yield a distinct, maximum peak near the true parameters. This indicates that for low numbers of samples, the IAA-based covariance matrix estimate should be used. Moreover, it supports the practicability of optimal filtering with the IAA despite its computational complexity, since small sample sizes are generally preferred when the signal of interest is nonstationary, violating the stationarity assumption in (4). This is often the case in practice, e.g., when processing speech signals. The same experiment was conducted for $N = 80$ resulting in the cost functions in Fig. 2. For this sample length, both optimal filtering methods seem to provide a good estimate of the DOA and pitch for both covariance matrix estimates. However, the sample covariance matrix estimate seems to give the best resolution in this case due to narrower peaks around the true parameters, with the optimal single filter having the narrowest peak. This indicates that, for longer sample sizes, the sample covariance matrix estimate may be preferred.

In the another series of experiments, we evaluated the statistical performance in terms of mean squared errors (MSEs) of the proposed estimators implemented using the IAA-based covariance matrix estimate (since relatively small sample sizes are considered) through Monte Carlo simulations. In all these experiments, 100 Monte Carlo simulations were conducted for each parameter setting, and, in each simulation, the noise and the phases of the harmonics were randomized. The MSEs of the pitch and DOA estimates (MSE_{ω_0} and MSE_{θ}), respectively) obtained from these simulations were calculated as

$$\text{MSE}(\widehat{\omega}_0) = \frac{1}{Q}\sum_{q=1}^{Q}(\omega_{0,q} - \widehat{\omega}_{0,q})^2, \tag{48}$$

$$\text{MSE}(\widehat{\theta}) = \frac{1}{Q}\sum_{q=1}^{Q}(\theta_q - \widehat{\theta}_q)^2, \tag{49}$$

where Q is the number of Monte Carlo simulations, q is the simulation number, $\omega_{0,q}$ and θ_q are the true pitch and DOA in simulation q, and $\widehat{(\cdot)}$ denotes an estimate of a parameter.

Moreover, a synthetic, multichannel periodic signal was used in every simulation with $L = 4$ harmonics with unit amplitudes, and, in each simulation, the pitch and DOA were sampled from $\mathcal{U}(250 \text{ Hz}, 300 \text{ Hz})$ and $\mathcal{U}(15°, 35°)$, respectively, where $\mathcal{U}(a, b)$ denotes the continuous uniform distribution in the interval $[a; b]$. The methods evaluated in these experiments are the optimal, white noise, and approximate filterbank ('bo', 'bw', and 'ba') and single filter ('so', 'sw', and 'sa') methods, the multichannel pitch estimator ('am') in [29], the steered response power method with phase transform ('sp') [30], and the exact and asymptotic nonlinear least squares (NLS) methods ('n' and 'an') in [10]. First, the performance of the proposed methods was evaluated for different Γ's in scenarios with 1) $N = 20$ and an SNR of 30 dB, 2) $N = 20$ and an SNR of 20 dB, and 3) $N = 25$ and an SNR of 30 dB. The other simulation parameters were: $f_s = 4$ kHz,

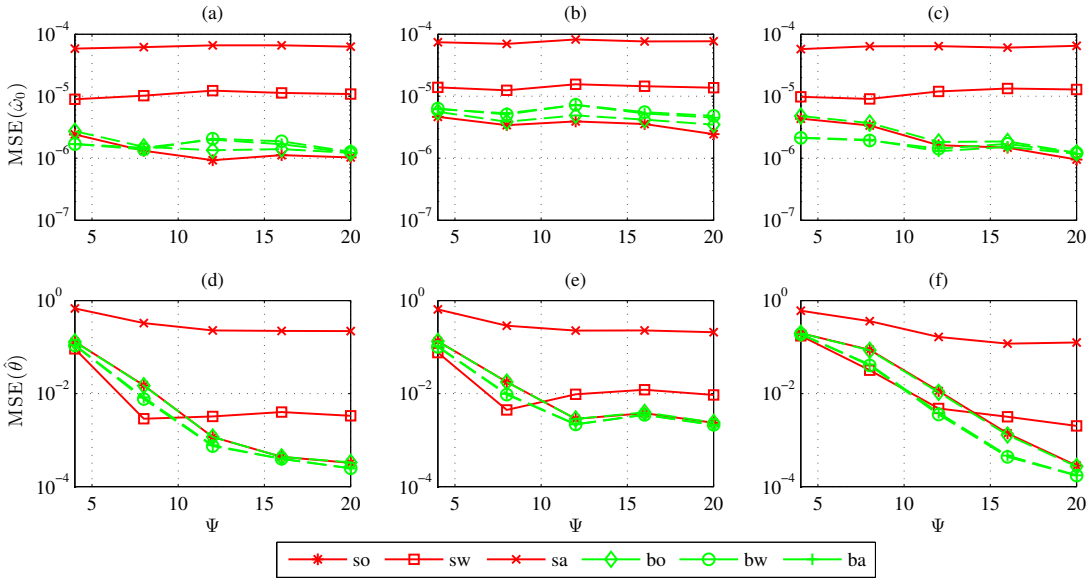


Figure 4. MSEs of pitch and DOA estimates for different Ψ's in scenarios with (a,d) $K = 2$ and an SNR of 30 dB, (b,e) $K = 2$ and an SNR of 20 dB, and (c,f) $K = 3$ and an SNR of 30 dB for the proposed methods.

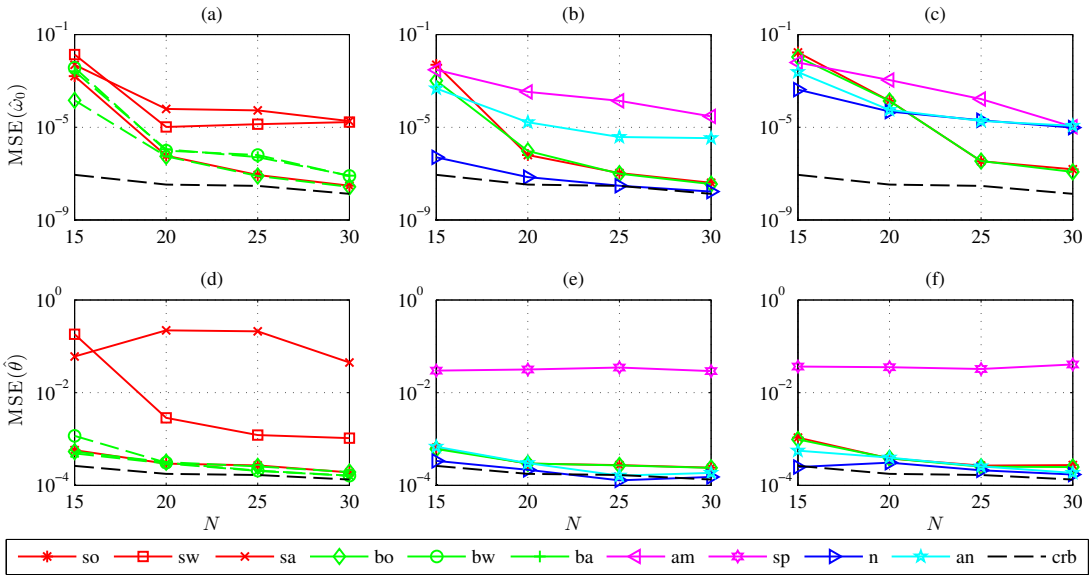


Figure 5. MSEs of pitch and DOA estimates for different N's in scenarios with (a,b,d,e) white noise, and (c,f) white noise and an interfering source for the proposed and state-of-the-art methods.

$c = 343$ m/s, $K = 2$, $d = 0.04$ m, and $\Psi = 64$. The results from this series of simulations are depicted in Fig. 3. From this figure, we make two important observations: first, the performances of the 'sw' and 'sa' methods are generally worse than those of the other proposed methods. Moreover, the results indicate that the higher the SNR and number of samples N, the more frequency grid points Γ is needed in the IAA-based covariance matrix estimation to achieve the highest possible performance. A similar series of simulations were conducted where the performance of the proposed methods were evaluated for different Ψ's. In these experiments, three scenarios were considered: 1) $K = 2$ and an SNR of 30 dB, 2) $K = 2$ and an SNR of 20 dB, and 3) $K = 3$ and an SNR of 30 dB. The other simulation parameters were the same as in the previous series of simulations except that $N = 20$ and $\Gamma = 100$, and the results are provided in Fig. 4. As in the previous series of simulations, we observe that the higher the SNR and number of sensors, the more spatial frequency grid points is needed in the IAA-based covariance matrix estimation to achieve the maximum possible performance.

Then, we conducted other series of simulations where the performance of the proposed methods were also compared with the state-of-the-art methods mentioned before. In the first of these evaluations, the performance was measured for different N's in two scenarios: 1) a scenario where the periodic signal was added with white noise at an SNR of 30 dB, and

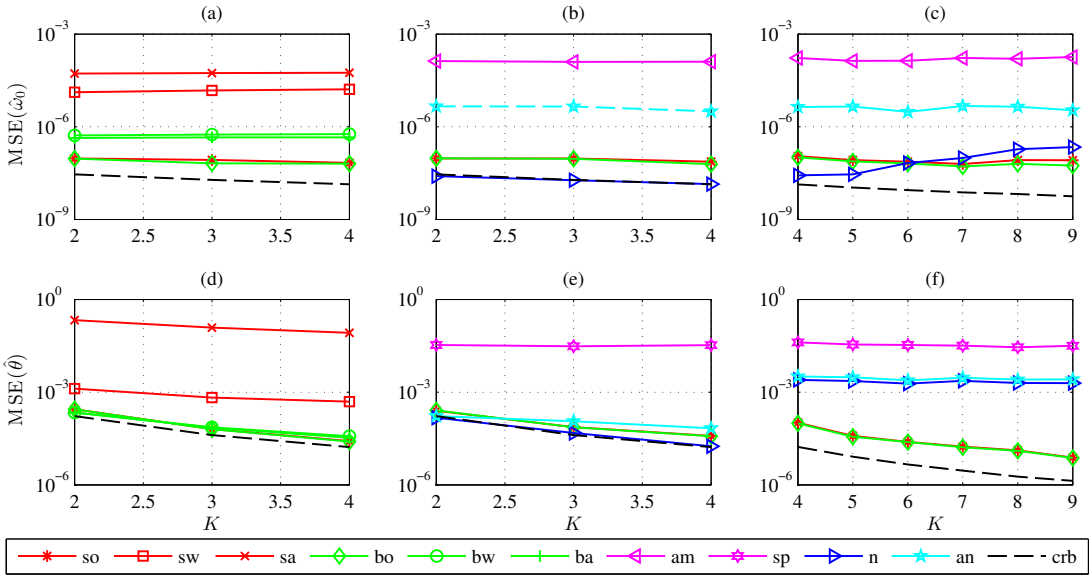


Figure 6. MSEs of pitch and DOA estimates for different K's in scenarios with (a,b,d,e) white noise, and (c,f) white noise and an interfering source for the proposed and state-of-the-art methods.

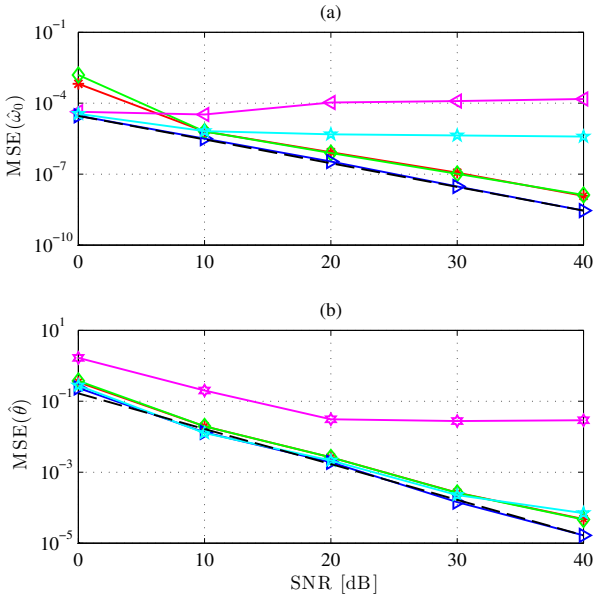


Figure 7. MSEs of pitch and DOA estimates for different SNRs in a scenario with white noise for the proposed and state-of-the-art methods. The labels for the plot are similar to those in Fig. 5.

2) a scenario with both white noise and an interfering source added where the SNR was 30 dB, and the interfering source was a single sinusoid with unit amplitude and random phase. The interfering sinusoid had the same DOA as the desired signal, but a frequency equal to $f_\mathrm{i} = f_0 + 60$ Hz. Otherwise, the simulation parameters were chosen as in the previous Monte-Carlo simulations except that $\Gamma = 512$, $\Psi = 64$, and $K = 2$. The results are found in Fig. 5. First of all, we observe that the proposed 'so', 'bo', 'bw', and 'ba' methods all yield similar performance, and that they outperform the 'sw' and 'sa' methods for the whole range of N's. In the comparison with

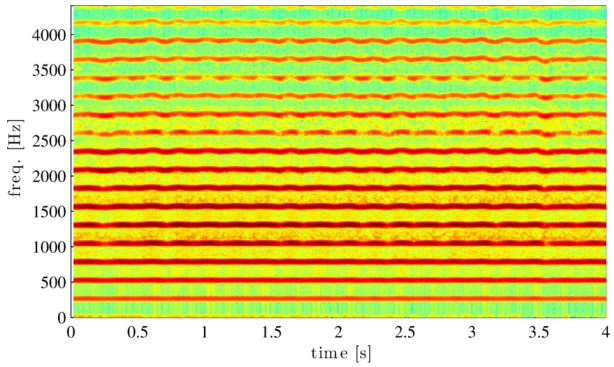


Figure 8. Plot of the spectrogram of a single-channel trumpet signal with vibrato.

state of the art, we see that the 'n' method generally has the best performance in the white noise only scenario. However, for higher N's (≥ 25), there is not much difference between the proposed optimal filtering methods and the 'n' method and, for $N \geq 20$, the proposed methods clearly outperform the 'an' and 'am' methods for pitch estimation and the 'sp' method for DOA estimation. Finally, in the scenario with an interfering source, the proposed optimal filters clearly outperform all other methods for pitch estimation in the range $25 \leq N \leq 30$, while they are only slightly worse than the 'n' and 'an' methods for DOA estimation in general. Similarly, we also evaluated the performance for different K's. Again, a scenario with white noise and a scenario with white noise and an additional interfering sinusoid with unit amplitude were considered. In this evaluation, however, the interfering source had the same frequency as the pitch of the harmonic signal, while its DOA was $\theta_\mathrm{i} = \theta - 80°$. The IAA grid size parameters were $\Gamma = 256$ and $\Psi = 128$, the number of temporal samples was $N = 25$, and otherwise the simulation parameters were the same as in the previous Monte-Carlo simulations. The results from this

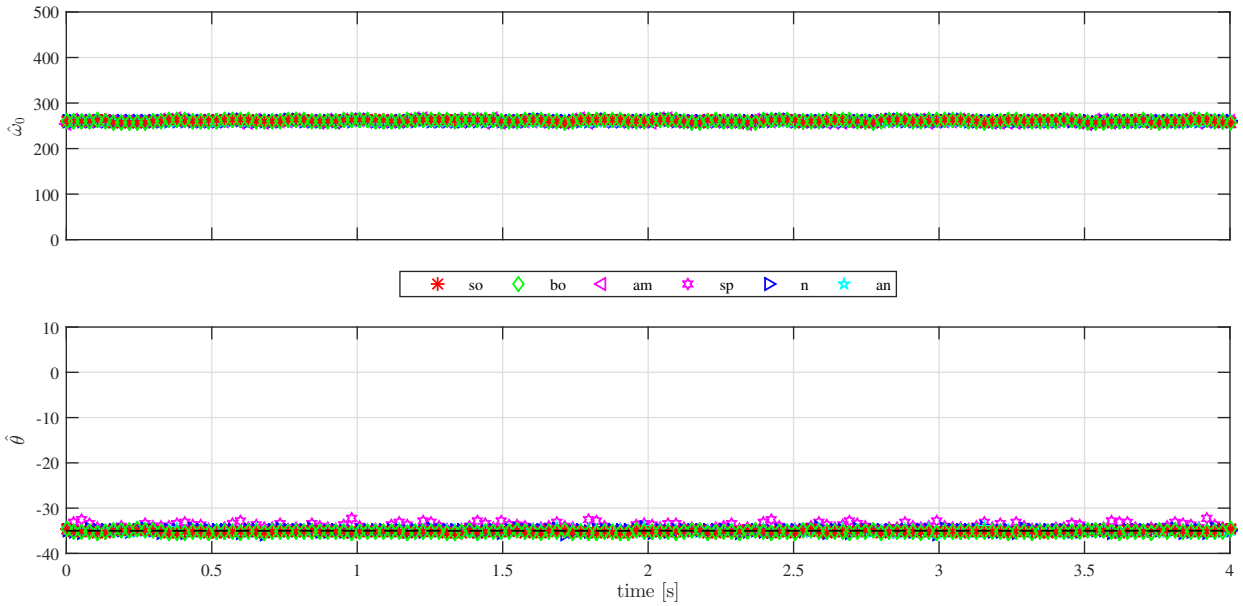


Figure 9. Plots of (top) pitch and (bottom) DOA estimates obtained from the spatially resynthesized, trumpet signal in a scenario with no reverberation.

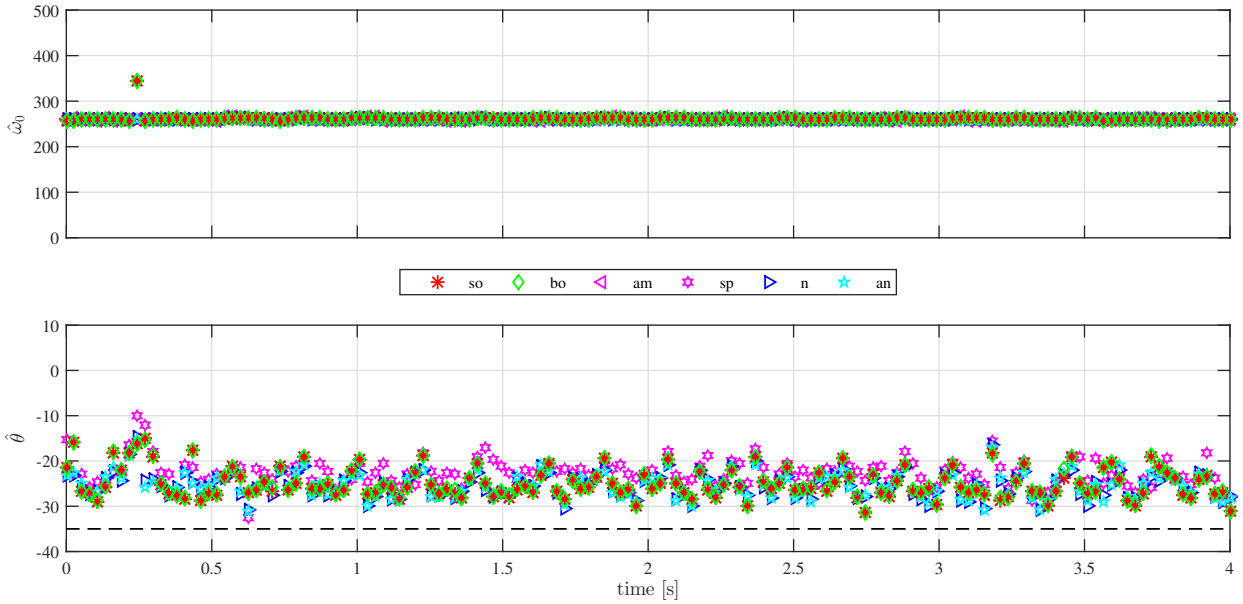


Figure 10. Plots of (top) pitch and (bottom) DOA estimates obtained from the spatially resynthesized, trumpet signal in a scenario with reverberation and a reverberation time of $T_{60} = 0.5$ s.

experiment are depicted in Fig. 6. For pitch estimation, the 'bw' and 'ba' generally yield the best performance of the proposed methods, followed closely by the 'bo', 'so', 'sa', and 'sw' methods in this order. For DOA estimation, the 'so', 'bo', 'bw', and 'ba' methods yield similar performance and outperform the 'sw' and 'sa' methods. In comparison with state of the art in the white noise scenario, we see that the 'so' and 'bo' methods have similar performance to the 'an' method for pitch estimation and that they outperform the 'am' method. The 'n' method generally yields the most accurate pitch estimates, though. The same observations are also valid for DOA estimation for the 'so', 'bo', 'an' and 'n' methods, whereas the 'sp' method shows a much worse performance. In the scenario where an additional interfering sinusoid is added, the proposed 'so' and 'bo' outperform the 'an' and 'am' methods for all K's, wheres the 'n' method shows better performance for low K's due to bias and worse performance for higher K's. For DOA estimation the proposed optimal filtering methods clearly outperforms all other methods in the comparison.

In the last series of Monte-Carlo simulations, the performances were measured for different SNRs in a scenario with white noise only. The setup for these simulations was: $N = 25$, $K = 2$, $\Gamma = 512$, $\Psi = 50$, and the remaining parameters were setup as in the previous Monte-Carlo simulations. We see, from the results in Fig. 7, that the 'n' method has the best performance as expected for both DOA and pitch estimation in all scenarios, however, the difference in terms of DOA estimation performance between the 'n', 'an', 'so', and 'bo'

methods is negligible for low SNRs ($\leq$ 30 dB). In terms of pitch estimation, the proposed methods outperform the 'am' and 'an' methods for SNR $\geq$ 20 dB, and, for DOA estimation, the 'sp' method is outperformed in all scenarios.

A final evaluation of the proposed filtering methods was conducted on a real-life signal. The signal used in this experiment was a 4 seconds long, single-channel trumpet signal with vibrato. The spectrogram of the signal is shown in Fig. 8, and it can be seen that it has a pitch fluctuating around $\approx$260. Based on the spectrogram, we chose a fixed model order for the experiment of $L = 5$. To obtain a multichannel signal, the signal was resynthesized spatially, using an online available room impulse response (RIR) generator [31]. The RIR generator was set up as follows: $c = 343$ m/s, $f_{\rm s} = 8,820$ Hz, the microphones of a ULA with 5 sensors was located at $\langle 2 + d[k - (K-1)/2] \rangle$ m $\times$ 0.5 m $\times$ 1.5 m, for $k = 1, \ldots, 5$, $d = 0.04$ m, the source was located at $\theta = -35^\circ$ at a distance of 3 m from the center of the array, the room dimensions were 4 m $\times$ 4 m $\times$ 3 m, the length of the RIRs was 2048, the microphones had cardioid responses with orientation $(90^\circ, 0^\circ)$ [(azimuth, elevation)], and the reflection order was 0. Then, we generated the multichannel, real data using this setup, and applied the proposed optimal filtering methods and the state of the art methods on time-consecutive frame of length $N = 40$ of the signal. The methods were implemented with $\Gamma = 128$, $\Psi = 64$, and, in the 'sp' method, we used an FFT length of 256 and integrated over frequencies in the interval $[150 \text{ Hz}, f_{\rm s}/2]$. From this experiment, we obtained the results depicted in Fig. 9. The results show that all methods yield pitch estimates close to the true pitch by comparing the estimates with the spectrogram of the trumpet signal. Moreover, we see that the proposed 'so' and 'bo' methods along with the 'an' and 'n' methods obtain DOA estimates closer to the true DOA than the 'sp' method at most time instances. Subsequently, a similar experiment was carried out where reverberation was added, i.e., the same simulation setup was used except that the reflection order was set to -1 (maximum), and the reverberation time was 0.5 s. With this setup, we obtained the results in Fig. 10. Again, all methods seem to provide pitch estimates close to the true pitch. The DOA estimates obtained using all methods are less accurate and biased in this scenario. In general, the proposed 'so' and 'bo' methods seem to perform similar to the 'n', 'an' methods in terms of accuracy, whereas the 'sp' method is generally outperformed.

V. Conclusion

In this paper, the problem of estimating the fundamental frequency as well as the direction-of-arrival of a desired, periodic signal has been considered, and some new methods based on spatio-temporal filtering have been proposed. The methods are based on optimal filter designs that leave periodic signals of a certain fundamental frequency from a certain direction-of-arrival unchanged while everything else is attenuated as much as possible. The resulting filters are adaptive if the statistics of the observed signal is estimated adaptively, and several incarnations of the ideas have been presented, including single filter and filterbank designs, simplifications based on the assumption that the observed noise signal is white and the filters being infinitely long. The application of the recently introduced iterative adaptive approach to estimation of the involved covariance matrix has also been proposed and investigated. This approach is capable of overcoming the usual limitations on the filter length relative to the number of samples available. That is, with this approach we can estimate the pitch and DOA using fewer samples which is preferable when processing nonstationary signals such as speech. In simulations, the proposed methods outperform state-of-the-art methods under adverse conditions, including the recently proposed maximum likelihood approach which is optimal for white, Gaussian noise and a single periodic signal. More specifically, the spatio-temporal filtering methods outperform the competing methods when multiple periodic signals are present at the same time, something that frequently happens in practice, cf. the well-known cocktail party problem. Finally, experiments on real data in form of a trumpet signal show the applicability of the proposed optimal filtering methods even in scenarios with slight reverberation.

References

[1] M. G. Christensen and A. Jakobsson, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

[2] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 7, pp. 1948–1963, Sep. 2012.

[3] V. C. Shields, Jr., "Separation of added speech signals by digital comb filtering," S. M. thesis, M.I.T., Cambridge, 1970.

[4] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 5, pp. 1124–1138, Oct. 1986.

[5] K. Nishi and S. Ando, "An optimal comb filter for time-varying harmonics extraction," *IEICE Trans. Fundamentals*, vol. E81-A, no. 8, pp. 1622–1627, Aug. 1998.

[6] M. G. Christensen, J. H. Jensen, A. Jakobsson, and S. H. Jensen, "On optimal filter designs for fundamental frequency estimation," *IEEE Signal Process. Lett.*, vol. 15, pp. 745–748, 2008.

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

[8] M. Brandstein and D. Ward, Eds., *Microphone Arrays - Signal Processing Techniques and Applications*. Springer-Verlag, 2001.

[9] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[10] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear least squares methods for joint DOA and pitch estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 923–933, May 2013.

[11] X. Qian and R. Kumaresan, "Joint estimation of time delay and pitch of voiced speech signals," *Rec. Asilomar Conf. Signals, Systems, and Computers*, vol. 1, pp. 735–739, Oct. 1995.

[12] M. Jian, A. C. Kot, and M. H. Er, "DOA estimation of speech source with microphone arrays," in *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 5, May 1998, pp. 293–296.

[13] L. Y. Ngan, Y. Wu, H. C. So, P. C. Ching, and S. W. Lee, "Joint time delay and pitch estimation for speaker localization," in *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 3, May 2003, pp. 722–725.

[14] 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 Process.*, vol. 2012, no. 1, pp. 1–11, Jan. 2012.

[15] M. Wohlmayr and M. Képesi, "Joint position-pitch extraction from multichannel audio," in *Proc. Interspeech*, Aug. 2007, pp. 1629–1632.

[16] M. Képesi, L. Ottowitz, and T. Habib, "Joint position-pitch estimation for multiple speaker scenarios," May 2008, pp. 85–88.

[17] J. Dmochowski, J. Benesty, and S. Affes, "Linearly constrained minimum variance source localization and spectral estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 8, pp. 1490–1502, Nov. 2008.
[18] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Joint DOA and fundamental frequency estimation methods based on 2-d filtering," in *Proc. European Signal Processing Conf.*, Aug. 2010, pp. 2091–2095.
[19] Z. Zhou, M. G. Christensen, J. R. Jensen, and H. C. So, "Joint DOA and fundamental frequency estimation based on relaxed iterative adaptive approach and optimal filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013.
[20] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.
[21] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.
[22] T. Yardibi, J. Li, P. Stoica, M. Xue, and A. B. Baggeroer, "Source localization and sensing: A nonparametric iterative adaptive approach based on weighted least squares," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 1, pp. 425–443, Jan. 2010.
[23] W. Roberts, P. Stoica, J. Li, T. Yardibi, and F. A. Sadjadi, "Iterative adaptive approaches to MIMO radar imaging," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 1, pp. 5–20, Feb. 2010.
[24] G.-O. Glentis and A. Jakobsson, "Efficient implementation of iterative adaptive approach spectral estimation techniques," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4154–4167, Sep. 2011.
[25] ——, "Superfast approximative implementation of the IAA spectral estimate," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 472–478, Jan. 2012.
[26] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Statistically efficient methods for pitch and DOA estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013.
[27] P. Stoica, A. Jakobsson, and J. Li, "Matched-filter bank interpretation of some spectral estimators," *Elsevier Signal Process.*, vol. 66, no. 1, pp. 45–59, 1998.
[28] J. Li and P. Stoica, "An adaptive filtering approach to spectral estimation and SAR imaging," *IEEE Trans. Signal Process.*, vol. 44, no. 6, pp. 1469–1484, Jun. 1996.
[29] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 409–412.
[30] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays - Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Springer-Verlag, 2001, ch. 8, pp. 157–180.
[31] E. A. P. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Tech. Rep., 2010, ver. 2.0.20100920. [Online]. Available: http://home.tiscali.nl/ehabets/rir_generator.html

Jesper Rindom Jensen (S'09–M'12) was born in Ringkøbing, Denmark in August 1984. He received the M.Sc. degree *cum laude* for completing the elite candidate education in 2009 from Aalborg University in Denmark. In 2012, he received the Ph.D. degree from Aalborg University.

Currently, he is a Postdoctoral Researcher at the Department of Architecture, Design & Media Technology at Aalborg University in Denmark, where he is also a member of the Audio Analysis Lab. He has been a Visiting Researcher at the University of Quebec, INRS-EMT, in Montreal, Quebec, Canada, and at the Friedrich-Alexander Universität Erlangen-Nürnberg in Erlangen, Germany. He has published more than 30 papers in peer-reviewed conference proceedings and journals. Among others, his research interests are digital signal processing, microphone array signal processing, and joint audio-visual signal processing with application to, e.g., enhancement, separation, localization, and tracking of speech and audio sources. In particular, he is interested in parametric analysis, modeling and extraction of such signals. Dr. Jensen has received an individual postdoc grant from the Danish Independent Research Council as well as several travel grants from private foundations.

Mads Græsbøll Christensen (S'00–M'05–SM'11) received the M.Sc. and Ph.D. degrees in 2002 and 2005, respectively, from Aalborg University (AAU) in Denmark, where he is also currently employed at the Dept. of Architecture, Design & Media Technology as Professor in Audio Processing and is head and founder of the Audio Analysis Lab.

He was formerly with the Dept. of Electronic Systems at AAU and has been held visiting positions at Philips Research Labs, ENST, UCSB, and Columbia University. He has published more than 100 papers in peer-reviewed conference proceedings and journals as well as 2 research monographs. His research interests include signal processing theory and methods with application to speech and audio, in particular parametric analysis, modeling, enhancement, separation, and coding.

Prof. Christensen has received several awards, including the Spar Nord Foundation's Research Prize, a Danish Independent Research Council Young Researcher's Award, and the Statoil Prize, as well as grants from the Danish Independent Research Council and the Villum Foundation's Young Investigator Programme. He is an Associate Editor for IEEE/ACM Trans. on Audio, Speech, and Language Processing and has previously served as an Associate Editor for IEEE Signal Processing Letters.

Jacob Benesty was born in 1963. He received a Master degree in microwaves from Pierre & Marie Curie University, France, in 1987, and a Ph.D. degree in control and signal processing from Orsay University, France, in April 1991. During his Ph.D. (from Nov. 1989 to Apr. 1991), he worked on adaptive filters and fast algorithms at the Centre National d'Etudes des Telecomunications (CNET), Paris, France. From January 1994 to July 1995, he worked at Telecom Paris University on multichannel adaptive filters and acoustic echo cancellation. From October 1995 to May 2003, he was first a Consultant and then a Member of the Technical Staff at Bell Laboratories, Murray Hill, NJ, USA. In May 2003, he joined the University of Quebec, INRS-EMT, in Montreal, Quebec, Canada, as a Professor. He is also a Visiting Professor at the Technion, in Haifa, Israel, and an Adjunct Professor at Aalborg University, in Denmark and at Northwestern Polytechnical University, in Xi'an, Shaanxi, China.

His research interests are in signal processing, acoustic signal processing, and multimedia communications. He is the inventor of many important technologies. In particular, he was the lead researcher at Bell Labs who conceived and designed the world-first real-time hands-free full-duplex stereophonic teleconferencing system. Also, he conceived and designed the world-first PC-based multi-party hands-free full-duplex stereo conferencing system over IP networks.

He was the co-chair of the 1999 International Workshop on Acoustic Echo and Noise Control and the general co-chair of the 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. He is the recipient, with Morgan and Sondhi, of the IEEE Signal Processing Society 2001 Best Paper Award. He is the recipient, with Chen, Huang, and Doclo, of the IEEE Signal Processing Society 2008 Best Paper Award. He is also the co-author of a paper for which Huang received the IEEE Signal Processing Society 2002 Young Author Best Paper Award. In 2010, he received the "Gheorghe Cartianu Award" from the Romanian Academy. In 2011, he received the Best Paper Award from the IEEE WASPAA for a paper that he co-authored with Chen.

Søren Holdt Jensen (S'87–M'88–SM'00) received the M.Sc. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 1988, and the Ph.D. degree in signal processing from the Technical University of Denmark, Lyngby, Denmark, in 1995. Before joining the Department of Electronic Systems of Aalborg University, he was with the Telecommunications Laboratory of Telecom Denmark, Ltd, Copenhagen, Denmark; the Electronics Institute of the Technical University of Denmark; the Scientific Computing Group of Danish Computing Center for Research and Education (UNI•C), Lyngby; the Electrical Engineering Department of Katholieke Universiteit Leuven, Leuven, Belgium; and the Center for PersonKommunikation (CPK) of Aalborg University. He is Full Professor and heading a research team working in the area of numerical algorithms, optimization, and signal processing for speech and audio processing, image and video processing, multimedia technologies, and digital communications. Prof. Jensen was an Associate Editor for the IEEE Transactions on Signal Processing, Elsevier Signal Processing and EURASIP Journal on Advances in Signal Processing, and is currently Associate Editor for the IEEE/ACM Transactions on Audio, Speech and Language Processing. He is a recipient of an European Community Marie Curie Fellowship, former Chairman of the IEEE Denmark Section and the IEEE Denmark Section's Signal Processing Chapter. He is member of the Danish Academy of Technical Sciences and was in January 2011 appointed as member of the Danish Council for Independent Research—Technology and Production Sciences by the Danish Minister for Science, Technology and Innovation.

ISSN: 1603-6204
ISBN: 978-87-85000-00-2