Enhancement of Periodic Signals

with Application to Speech Signals

Jensen, Jesper Rindom

Publication date:
2012

Document Version
Accepted author manuscript, peer reviewed version

Link to publication from Aalborg University

Citation for published version (APA):
Jensen, J. R. (2012). *Enhancement of Periodic Signals: with Application to Speech Signals*.

Enhancement of Periodic Signals

- with Application to Speech Signals

Ph.D. Thesis

JESPER RINDOM JENSEN

Multimedia Information and Signal Processing
Department of Electronic Systems
Aalborg University
Niels Jernes 12, 9220 Aalborg Ø, Denmark

Enhancement of Periodic Signals - with Application to Speech Signals
Jesper Rindom Jensen

ISBN 978-87-92328-94-6

August 2012

Abstract

The topic of this thesis is the enhancement of noisy, periodic signals with application to speech signals. Generally speaking, enhancement methods can be divided into signal- and noise-driven methods. In this thesis, we focus on the signal-driven approach by employing relevant signal parameters for the enhancement of periodic signals. The enhancement problem consists of two major subproblems: the estimation of relevant parameters or statistics, and the actual noise reduction of the observed signal. We consider both of these subproblems.

First, we consider the problem of estimating signal parameters relevant to the enhancement of periodic signals. The fundamental frequency is one example of such a parameter. Furthermore, in multichannel scenarios, the direction-of-arrival of the periodic sources onto an array of sensors is another parameter of relevance. We propose methods for the estimation of the fundamental frequency that have benefits compared to other state-of-the-art estimation methods. For example, we consider improving the spectral resolution of existing subspace-based and optimal filtering based fundamental frequency estimators. Moreover, we propose fast implementations of the proposed optimal filtering based estimators, e.g., by exploiting matrix structures. This decreases the computational complexity by several orders of magnitude.

We also consider the joint estimation of the fundamental frequency and the direction-of-arrival. Joint estimation enables us to resolve multiple periodic sources that share the same fundamental frequency and have different directions-of-arrival and vice versa. This may not be possible if the parameters are estimated separately. Moreover, we stress the importance of estimating the parameters jointly in relation to the estimation accuracy. Both optimal filtering based and nonlinear least squares based joint estimators are proposed; the former is excellent for resolving closely spaced sources, while the latter is statistically efficient.

Then, we consider noise reduction methods based on the aforementioned parameter estimates. First, we propose several non-causal, time-domain filters for single-channel noise reduction. These non-causal filters can increase the noise reduction compared to their causal counterparts without increasing the distortion of the desired signal. We also show the link between some single-channel, signal- and noise-driven noise reduction filters; motivated by this, we suggest joint filtering schemes employing these two filter types for tackling the difficult problem of nonstationary noise reduction. It was shown that the suggested schemes outperform other widely used enhancement methods for nonstationary noise reduction in terms of perceptual scores. Finally, we propose an optimal filtering based method for multichannel periodic signal enhancement that is driven by fundamental frequency and direction-of-arrival estimates. This method was proven useful for enhancement of real-life, multichannel, periodic signals.

In summary, the importance of joint parameter estimation is clarified by our contributions to the relatively young research topic of joint fundamental frequency and direction-of-arrival estimation of multichannel periodic signals. Joint estimation is a key to obtain robust and accurate fundamental frequency and direction-of-arrival estimators. Moreover, our contributions on noise reduction reveals the applicability of signal-driven enhancement of single-channel and multichannel periodic signals. By utilizing information about relevant signal parameters such as the fundamental frequency and the direction-of-arrival, noise reduction can be conducted without relying fully on the noise statistics. As appearing from our results, this can be exploited to obtain robust methods for nonstationary noise reduction.

Resumé

Emnet for denne afhandling er støjreduktion af støjfyldte og periodiske signaler som for eksempel talesignaler. Generelt kan støjreduktionsmetoder opdeles i signal- og støjdrevne metoder. I denne afhandling fokuseres der på den signaldrevne tilgang ved at bruge relevante signalparametre i støjreduktionen af periodiske signaler. Støjreduktionsproblemet består af to overordnede delproblemer: Estimeringen af de relevante signalparametre eller statistikker og den egentlige støjreduktion af det observerede signal. Begge delproblemer behandles i denne afhandling.

Først behandles delproblemet vedrørende estimering af de relevante parametre til støjreduktion af periodiske signaler. Den fundamentale frekvens er ét eksempel på sådan en parameter. En anden relevant parameter i multikanals scenarier er ankomstvinklen af en periodisk kilde på et array af sensorer. Metoder foreslås til estimering af den fundamental frekvens. Disse har en række fordele sammenlignet med de nyeste eksisterende metoder. For eksempel foreslås forbedringer mht. den spektrale opløsning af underrums- og optimal filtreringsbaserede metoder til estimering af den fundamental frekvens. Ydermere foreslås hurtige implementationer af de foreslåede optimal filtreringsbaserede estimatorer. De hurtige implementationer udnytter eksempelvis matrixstrukturer, hvilket sænker den beregningsmæssige kompleksitet betydeligt.

Samtidig estimation af den fundamental frekvens of ankomstvinklen behandles også. Samtidig estimation gør det muligt at adskille flere periodiske kilder med den samme fundamental frekvens og forskellige ankomstvinkler og vice versa. Dette er ikke nødvendigvis muligt, hvis parametrene estimeres hver for sig. Desuden understreges vigtigheden af at estimere parametrene samtidigt i relation til estimationsnøjagtigheden. Dernæst foreslås både optimal filtreringsbaserede og ulineær mindste kvadraters baserede samtidige estimatorer. Den første metode er fremragende til at adskille tætplacerede kilder og den anden er statistisk effektiv.

Derefter betragtes støjeduktionsmetoder baseret på de førnævnte parameterestimater. Først foreslås en række nonkausale, tidsdomæne filtre til enkeltkanals støjreduktion. Disse nonkausale filtre kan øge støjreduktionen sammenlignet med deres tilsvarende kausale filtre uden at øge forvrængningen af det ønskede signal. Sammenhængen mellem nogle enkeltkanals signal- og støjdrevne støjreduktionsfiltre vises også. Denne sammenhæng motiverer anvendelsen af disse to filtertype samtidigt til reduktion af us-

tationær støj, som det udnyttes i de to efterfølgende foreslåede filtreringssystemer. Det er vist, at disse systemer udkonkurrerer andre ofte brugte metoder til reduktion af ustationær støj med hensyn til et perceptuelt mål. Endeligt foreslås en optimal filtreringsbaseret metode til støjreduktion af multikanals periodiske signaler. Denne metode er drevet af estimater af den fundamentale frekvens og af ankomstvinklen. Metoden har vist sig anvendelig til støjreduktion af virkelige, multikanals, periodiske signaler.

Bidragene til det relativt unge forskningsområde omkring samtidig estimering af den fundamentale frekvens og ankomstvinklen tydeliggør vigtigheden af samtidig estimering. Samtidig estimering er en nøgle til at opnå robuste og præcise estimatorer af den fundamentale frekvens og ankomstvinklen. Ydermere viser bidragene vedrørende støjreduktion anvendeligheden af signaldrevet støjreduktion af enkelt- og multikanals periodiske signaler. Ved at anvende information om relevante signalparametre såsom den fundamental frekvens og ankomstvinklen, kan der foretages støjreduktion uden at hvile fuldstændigt på støjstatistikkerne. Som det fremgår af resultaterne, kan dette udnyttes til at opnå robuste metoder til reduktion af ustationær støj.

List of Papers

The main body of this thesis consists of the following papers:

[A] J. R. Jensen, M. G. Christensen and S. H. Jensen, "Fundamental Frequency Estimation using Polynomial Rooting of a Subspace-Based Method". In *Proc. European Signal Processing Conference*, 2010.

[B] J. R. Jensen, M. G. Christensen and S. H. Jensen, "A Single Snapshot Optimal Filtering Method for Fundamental Frequency Estimation". In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2011.

[C] 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. on Signal Processing*, 2012, submitted.

[D] 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 Conference*, 2010.

[E] 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 Processing*, 2012, submitted.

[F] J. R. Jensen, J. Benesty, M. G. Christensen and S. H. Jensen, "Non-Causal Time-Domain Filters for Single-Channel Noise Reduction". *IEEE Trans. on Audio, Speech and Language Processing*, vol. 20(5), pp. 1526–1541, 2012.

[G] 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 Processing*, vol. 20(7), pp. 1948–1963, 2012.

[H] J. R. Jensen, J. Benesty, M. G. Christensen and S. H. Jensen, "Joint Filtering Scheme for Nonstationary Noise Reduction". Accepted for *Proc. European Signal Processing Conference*, 2012.

[I] J. R. Jensen, M. G. Christensen and S. H. Jensen, "An Optimal Spatio-Temporal Filter for Extraction and Enhancement of Multi-Channel Periodic Signals". In *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2010.

The following additional papers have been published by the author:

[1] 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 *Proc. European Signal Processing Conference*, 2008.

[2] J. K. Nielsen, J. R. Jensen, M. G. Christensen, S. H. Jensen and T. Larsen, "Waveform Approximating Residual Audio Coding with Perceptual Pre- and Post-Filtering". In *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2008.

[3] J. R. Jensen, M. G. Christensen, M. H. Jensen, S. H. Jensen and T. Larsen, "Multiple Description Spherical Quantization of Sinusoidal Parameters with Repetition Coding of the Amplitudes". In *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2009.

[4] J. R. Jensen, M. G. Christensen, M. H. Jensen, S. H. Jensen and T. Larsen, "Robust Parametric Audio Coding Using Multiple Description Coding". *IEEE Signal processing Letters*, vol. 16(12), pp. 1083–1086, 2009.

[5] J. R. Jensen, G.-O. Glentis, M. G. Christensen, A. Jakobsson and S. H. Jensen, "Computationally Efficient IAA-Based Estimation of the Fundamental Frequency". Accepted for *Proc. European Signal Processing Conference*, 2012.

Preface

This thesis is submitted to the Doctoral School of Engineering and Science at Aalborg University in partial fulfillment of the requirements for the degree of doctor of philosophy. The thesis consists of two parts: the first part is an introduction to the research area and the second part is a collection of papers that have been published in or submitted to peer-reviewed conferences or journals. The work was carried out in the period from August 2009 to July 2012 at the Department of Electronic Systems at Aalborg University.

During the past three years, several persons have made praiseworthy contributions to my work and life. First of all, I am grateful that my supervisor Søren Holdt Jensen gave me the opportunity to work on this research topic, and for giving me the freedom to choose the research directions of my interest. Also, I am very grateful to my co-supervisor Mads Græsbøll Christensen who contributed greatly to my work through all of our endless and fruitful discussions. Moreover, I would like to thank Mads for spurring my interest in joint fundamental frequency and direction-of-arrival estimation.

I would also like to thank Jacob Benesty for inviting me, as a visiting researcher, to the Institut National de la Recherche Scientifique - Énergie Matériaux Télécommunications in Montreal. Also, Jacob deserves my gratitude for all our fruitful discussions on speech enhancement from which I have benefited greatly. Furthermore, I am grateful for the collaboration I have had with George-Othon Glentis and Andreas Jakobsson. A special thanks goes to George for sharing his expertise on fast implementations.

The microphone array that I have been using for recording signals for some of the papers in this thesis was developed in collaboration with Ben Krøyer. Building this array would have been difficult without the countless hours Ben has spend on this, so for that I am very thankful. I would also like to acknowledge the members of the Multimedia and Information Signal Processing section at the Department of Electronic Systems, Aalborg University and all other people at Aalborg University who has contributed to my research or my social life. In this regard, special thanks goes to Jesper Kjær Nielsen with whom I have shared office during most of my years at Aalborg University. I would like to thank Jesper for all his contributions to my social life and for all our office discussions.

Last, but not least, I thank my friends, family and, most of all, my wife Helle for love and support the past three years.

Contents

Part I

Introduction

Introduction to Periodic Signal Enhancement

1 Introduction

If a function $f(t)$ repeats its values in regular intervals of length T, i.e.,

$$f(t) = f(t + T), \tag{1}$$

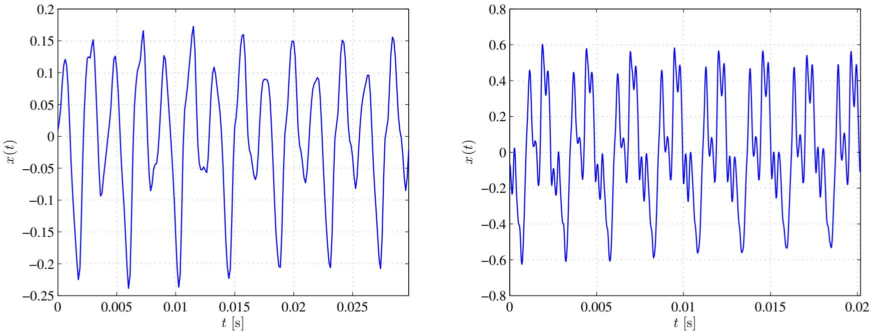
we say that it is a periodic function [146]. A signal $x(t)$ which can be exactly described by a periodic function is termed a periodic signal. If the period T of $x(t)$ changes slowly over time, we say that the signal $x(t)$ is quasiperiodic. That is, for a quasiperiodic signal $x(t)$, we have that

$$x(t) \approx x\left(t + T\right). \tag{2}$$

If a relatively small, confined time interval $t \in [t_1; t_2]$ is considered, in which the period is almost constant, the signal $x(t)$ can be treated as periodic signal.

Many real-life signals, both artificial and natural, are quasiperiodic. A few examples of such signals are electrocardiograms (ECGs) [137], voiced speech [58], sounds from musical instruments (guitar, violin, trumpet, etc.) [4], passive sonar signals from boats [144], radar returns from helicopters [171], vibration signals [5, 70], astronomical data (e.g., star observations) [152], seismological data [139], and infant cry [111]. In Fig. 1, we have depicted two examples of such real-life quasiperiodic signals. During the last century, countless applications have spawned employing quasiperiodic signals as the ones just described. Considering voiced speech and audio, for example, some applications are hearing-aids, teleconference systems, music information retrieval, diagnosis of illnesses, and surveillance systems.

A common desire in most applications utilizing quasiperiodic signals is that the raw quasiperiodic signal is available. Unfortunately, this is rarely the case due to the presence of noise. In hearing-aids, for example, the use case is often that the user wants to focus the attention on a particular quasiperiodic stimulus while ignoring a range of other



(a) (b)

Fig. 1: Excerpts from recordings of (a) female speech and (b) a violin, respectively.

stimuli commonly referred to as the cocktail party problem [28]. Hearing impaired persons, however, can not resolve multiple sound sources naturally as efficient as persons with normal hearing [85]. That is, background noise can have a devastating impact on the listening experience of hearing impaired people; the noise can cause both auditory fatigue [177], and general discomfort for the listener [48, 198]. Coding systems for, e.g., mobile phones and Voice over IP (VoIP) are other examples of applications involving quasiperiodic signals which are negatively affected by noise since noise will often cause the coding efficiency to decrease [132, 164, 185]. A third application example is automatic speech recognition (ASR). ASR systems are used for control in, e.g., aircraft cockpits and wheelchairs where environmental noise is inevitable. It is well-known that noise can increase the word error rate (WER) of such systems significantly [83]. While not mentioned here, several other noise critical applications exists (see, e.g., [12] for additional speech related examples).

In summary, many applications exist which utilize a quasiperiodic signal. As the quasiperiodic signal is often corrupted by noise, the performance of such applications will most likely be degraded if we do not apply any preprocessing. This fact has spurred decades of research in noise reduction (aka. enhancement) methods since the early 1960s where Schroeder filed a couple of patents on the analog implementation of the spectral magnitude subtraction method [167, 169]. While several types of naturally occurring quasiperiodic signals exist, we focus on enhancement of speech in this thesis. The remainder of this chapter gives an introduction to enhancement methods for both single- and multichannel periodic signals.

1.1 Signal Model

In this thesis, we consider both single- and multichannel quasiperiodic signals. To facilitate the development of enhancement methods for such signals, it is essential to have general and appropriate models.

Single-Channel Model

In the single-channel scenario, a single sensor, say a microphone, is used to pick up the discrete observed signal $y(n)$ which is constituted by a desired quasiperiodic signal $x(n)$ and a noise signal $v(n)$ as

$$y(n) = x(n) + v(n), \tag{3}$$

where n is the discrete time index. This rather simple additive noise model is efficient in describing real-life quasiperiodic signals in noise like noise corrupted voiced speech and audio [33]. In many applications, the desired signal is reflected by surfaces generating a multitude of echoes [52]. We can easily extend the model in (3) to encompass this scenario. However, convolutive noise is not the topic of this thesis, and will not be considered further herein.

As the desired signal is quasiperiodic, we can extend the model in (3). In the 19th century, Joseph Fourier showed that any periodic signal can be decomposed into a sum of a set of complex exponentials [67]. This knowledge enable us to rewrite (3) as

$$\begin{aligned} y(n) &= \sum_{l=1}^{L} A_l \cos(l\omega_0 n + \phi_l) + v(n), && (4) \\ &= \sum_{l=1}^{L} \left(a_l e^{jl\omega_0 n} + a_l^* e^{-jl\omega_0 n} \right) + v(n) \ , && (5) \end{aligned}$$

where L is the harmonic model order, $a_l = \frac{A_l}{2} e^{j\phi_l}$, $A_l > 0$ is the complex amplitude of the lth harmonic, ϕ_l is the phase of the lth harmonic, ω_0 is the fundamental frequency, and $(\cdot)^*$ denotes the complex conjugate. The model in (4) assumes exact periodicity even though some real-life, quasiperiodic signals contain inharmonicities [66, 162]. To account for inharmonicity, (4) can be adapted to model the underlying physical phenomenon causing the inharmonicity [59, 75, 77, 106], or a more general model with perturbed harmonics can be used [33, 40, 74].

As we will see later, an integral part of the enhancement methods considered in this thesis is fundamental frequency estimation. Often, the fundamental frequency is estimated from analytic signals while the desired signal is in fact real as in (4) [33]. It is straightforward, though, to convert real signals to analytic signals[1] by using the

[1] Note that analytic signals only exist for continuous-time real signals, so the analytic signals we refer to herein are analytic-like discrete signals obtained using the discrete Hilbert transform [126].

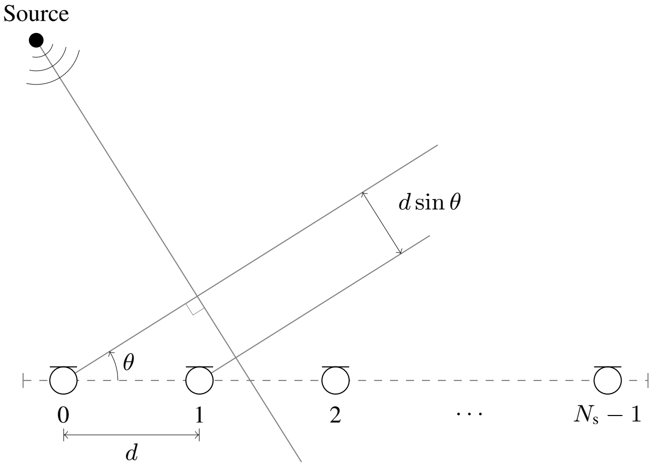


Fig. 2: Illustration of the uniform linear array structure.

Hilbert transform [81, 126]. Moreover, utilizing analytic signals can have the benefits of a lower computational complexity and a simpler notation of the estimator [33]. The complex, single-channel counterpart to (4) is given by

$$y(n) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n} + v(n), \tag{6}$$

with $\alpha_l = A_l e^{j\phi_l}$. Another important parameter is the model order L. If the model order is not chosen or estimated correctly, many fundamental frequency estimators would very likely yield erroneous estimates (aka. fundamental frequency halvings and doublings). To circumvent this problem, the model order can be estimated either separately from [181, 183] or jointly with [36, 141] the fundamental frequency.

Multichannel Model

The noise reduction capabilities of single-channel enhancement methods are rather limited since no spatial information is available. That is, the potential performance of enhancement methods can be increased by considering multiple channels [12]. In the multichannel scenario, we can model the observed signal $y_{n_\mathrm{s}}(n_\mathrm{t})$ at the n_sth sensor and at time instance n_t as [15]

$$\begin{aligned} y_{n_\mathrm{s}}(n_\mathrm{t}) &= x_{n_\mathrm{s}}(n_\mathrm{t}) + v_{n_\mathrm{s}}(n_\mathrm{t}) && (7)\\ &= \beta_{n_\mathrm{s}} s(n_\mathrm{t} - f_\mathrm{s}\tau_{n_\mathrm{s}}) + v_{n_\mathrm{s}}(n_\mathrm{t}), && (8) \end{aligned}$$

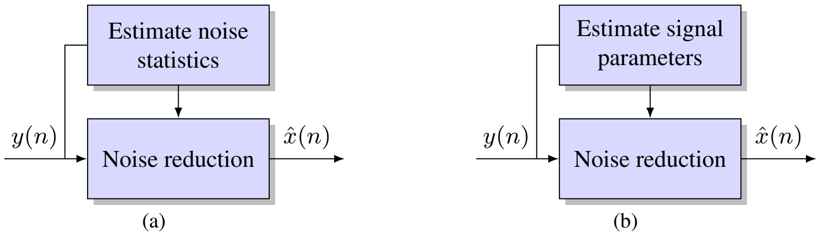


Fig. 3: Generic block diagrams for (a) noise- and (b) signal-driven enhancement methods.

where $x_{n_\mathrm{s}}(n_\mathrm{t})$ and $v_{n_\mathrm{s}}(n_\mathrm{t})$ denotes the desired signal and the noise, respectively, β_{n_s} is the attenuation of the source at the n_sth sensor, τ_{n_s} is the delay from sensor 0 to sensor n_s, and f_s is the sampling frequency. As the model in (3), this model is anechoic since convolutive noise is not considered.

For multichannel, quasiperiodic signals, we have extended models for real signals

$$y_{n_\mathrm{s}}(n_\mathrm{t}) = \beta_{n_\mathrm{s}} \sum_{l=1}^{L} \left[a_l e^{jl\omega_0(n_\mathrm{t}-f_\mathrm{s}\tau_{n_\mathrm{s}})} + a_l^* e^{-jl\omega_0(n_\mathrm{t}-f_\mathrm{s}\tau_{n_\mathrm{s}})} \right] + v_{n_\mathrm{s}}(n_\mathrm{t}), \tag{9}$$

and complex, analytic signals

$$y_{n_\mathrm{s}}(n_\mathrm{t}) = \beta_{n_\mathrm{s}} \sum_{l=1}^{L} \alpha_l e^{jl\omega_0(n_\mathrm{t}-f_\mathrm{s}\tau_{n_\mathrm{s}})} + v_{n_\mathrm{s}}(n_\mathrm{t}). \tag{10}$$

The models in (9) and (10) are general and holds for different structures of sensor arrays. If the array structure is known, we can specify the models even more by modeling the time-delays τ_{n_s} and/or the attenuation factors β_{n_s}. An example of a commonly utilized array structure is the uniform linear array (ULA) illustrated in Fig. 2 [192]. The ULA structure is assumed and utilized in the remainder of this thesis. If we assume that the source is in the far field of the array, we know that the delay τ_{n_s} is given by

$$\tau_{n_\mathrm{s}} = n_\mathrm{s} \frac{d \sin\theta}{c}, \tag{11}$$

where d is the inter-element spacing of the ULA, and c is the wave propagation velocity.

1.2 The Enhancement Problem

The ultimate goal of enhancement methods is to recover a desired signal from a noisy single-channel or multichannel mixture as in (3) and (7), respectively. In practice, however, it is extremely difficult, if not impossible, to remove the noise completely. A common practical goal is therefore that the noise should be attenuated as much as possible while the distortion of the desired signal is insignificant. During the previous decades, numerous methods have been proposed that consider this enhancement problem. Generally speaking, we can categorize these methods as either being driven by estimates of the noise statistics or of parameters describing the desired signal. The two different approaches are illustrated in Fig. 3. As hinted by these block diagrams, we can divide the enhancement problem into two subproblems: first, we need to estimate either the noise statistics or relevant signal parameters. Then, on basis of these estimates, we need to conduct the actual enhancement of the observed signal. The two subproblems are considered individually in the following sections.

2 Statistics and Parameter Estimation

The first step in most enhancement methods is to estimate either the noise statistics or relevant parameters describing the desired signal. For periodic signals, relevant parameters are, e.g., the fundamental frequency and the direction-of-arrival (DOA). Followingly, we describe the estimation of these noise and desired signal related quantities.

2.1 Noise Estimation

The vast majority of enhancement methods for, e.g., speech are driven by a noise estimate (see, e.g., [8, 12, 48, 124, 195] and the references therein). Naturally, this has nourished research in noise estimation, resulting in innumerable proposed methods. In many of the enhancement methods, only an implicit estimate of the noise is necessary, for example, in form of a noise autocorrelation matrix estimate $\hat{\mathbf{R}}_{\mathbf{v}} \in \mathbb{C}^{M_\mathrm{s}M_\mathrm{t}\times M_\mathrm{s}M_\mathrm{t}}$ or a noise cross power spectral density (CPSD) matrix estimate $\hat{\mathbf{S}}_{\mathbf{v}}(f) \in \mathbb{C}^{N_\mathrm{s}\times N_\mathrm{s}}$ at the frequency f. The true matrices are defined as [114]

$$\begin{align} \mathbf{R}_{\mathbf{v}} &= \mathrm{E}\left\{\mathbf{v}_{n_\mathrm{s}}(n_\mathrm{t})\mathbf{v}_{n_\mathrm{s}}^H(n_\mathrm{t})\right\}, \tag{12} \\ \mathbf{S}_{\mathbf{v}}(f) &= \mathcal{F}\left\{\mathbf{R}_{\mathbf{v}_\mathrm{s}}(m_\mathrm{t})\right\}, \tag{13} \end{align}$$

where $\mathrm{E}\{\cdot\}$ denotes the mathematical expectation operator, $\mathcal{F}\{\cdot\}$ denotes the element-wise discrete Fourier transform (DFT), and

$$\begin{align} \mathbf{v}_{n_\mathrm{s}}(n_\mathrm{t}) &= \mathrm{vec}\left\{\begin{bmatrix} v_{n_\mathrm{s}}(n_\mathrm{t}) & \cdots & v_{n_\mathrm{s}}(n_\mathrm{t}-M_\mathrm{t}+1) \\ \vdots & \ddots & \vdots \\ v_{n_\mathrm{s}+M_\mathrm{s}-1}(n_\mathrm{t}) & \cdots & v_{n_\mathrm{s}+M_\mathrm{s}-1}(n_\mathrm{t}-M_\mathrm{t}+1) \end{bmatrix}\right\}, \tag{14} \\ \mathbf{R}_{\mathbf{v}_\mathrm{s}}(m_\mathrm{t}) &= \mathrm{E}\left\{\mathbf{v}_\mathrm{s}(n_\mathrm{t}+m_\mathrm{t})\mathbf{v}_\mathrm{s}^H(n_\mathrm{t})\right\}, \tag{15} \\ \mathbf{v}_\mathrm{s}(n_\mathrm{t}) &= \begin{bmatrix} v_0(n_t) & \cdots & v_{N_\mathrm{s}-1}(n_\mathrm{t}) \end{bmatrix}^T, \tag{16} \end{align}$$

with vec$\{\mathbf{X}\}$ denoting the vectorization or column-wise stacking of the matrix $\mathbf{X}$. Followingly, we describe different approaches for estimating these noise related quantities for the single-channel and multichannel signals, respectively.

Single-Channel Noise Estimation

For some types of signals contaminated by noise, the noisy signal can be divided into two types of segments: segments in which the desired signal is present and absent, respectively. If it is assumed that the noise is stationary, the noise can be estimated during the segments of desired signal absence. To accomplish this, we need to detect whether or not the desired signal is present. Such a detection has been extensively studied for, e.g., speech signals in form of voice activity detection (VAD). However, it is non-trivial to detect if the desired signal is present as features describing only the desired signal needs to be extracted. Early examples of such features used in VAD methods are energy-levels and zero-crossings [154], cepstral features [82], the Itakura distance measure [155], and a periodicity measure [188]. Despite of the intuitiveness of the activity detection approach, it often yields erroneous error estimates in practice as the noise is rarely stationary.

Followingly, we consider three other classes of noise estimators that can estimate the noise even during presence of the desired signal. The first is the minimum tracking approach [124]. In this approach, it is assumed that the power of a noisy speech signal in each frequency bin decays to the power level of the noise even in segments with speech activity. That is, the noise level in a frequency bin can be estimated by tracking the minimum spectral level in that particular bin. This approach was first considered in the minimum statistics method originally proposed by Martin in [128]. The minimum statistics method was later refined in [129] by introducing bias compensation. In [55], Doblinger proposed an alternative to the minimum statistics method, where the noise spectral level is instead tracked per sample. However, while the minimum statistics based methods are useful for tracking the noise level even during speech presence, they generally need long time windows to reduce speech leakage in the noise estimate; this effectively puts a limit on these methods' ability to track rapid changes in the noise level [63].

Another well-known class of noise estimators is the time-recursive averaging algorithms. In these algorithms, it is exploited that the signal-to-noise ratio (SNR) and the presence probability of the desired signal vary across the different frequency bands. For example, if either the SNR or the presence probability is low in a particular frequency band, it is reasonable to update the noise spectral level estimate in that band. Pioneering examples of time-recursive averaging algorithms based on the SNR level and the presence probability can be found in [121, 122] and [41, 42, 63, 175, 176], respectively. Without going into too much detail, the time-recursive approach to noise estimation can yield significantly better noise estimates than the minimum statistic based methods for abruptly changing noise level; this is achieved by using a recursively averaged minimum MSE (MMSE) of the noise power for determining speech presence probability as considered in [63].

Histogram-based noise estimation algorithms constitute a third class of noise estimators. These estimators are based on the observation that the most occurring value of energy values in a frequency band corresponds to the noise level. That is, the noise level in a particular frequency band can be estimated from the histogram of the spectral levels within that band. Generally, the histograms have two modes: 1) a mode corresponding to frames which could contain noise and where the desired signal is absent, and 2) a mode corresponding to frames where the desired signal and maybe noise are present [160]. When noise is added, the modes get close and eventually they merge into one mode. In the two-mode case, the former mode is the most frequently occurring, and it will typically correspond to the noise level. A few examples of noise estimators using histograms are found in [88, 131, 189]. In general the histogram-based estimators can be used for noise estimation even during speech presence, however, they are computationally expensive, require much memory resources, and have poor performances in low SNR conditions. Moreover, the signal segments used for generating the histograms are typically several hundreds of milliseconds long, which limits these estimators' ability to track nonstationary noise [41].

Multichannel Noise Estimation

If multiple sensors in the vicinity of each other are used for measuring the desired signal in noise, spatial information about the sources is also embedded in the measurements. That is, the noise present in the measurements can be reduced, not only in the time- and frequency-domains, but also in the spatial domain. To achieve this, however, the spatial characteristics of the noise or desired signal need to be known in terms of, e.g., the cross power spectral density (CPSD). Estimating the multichannel noise can be even more challenging than estimating the single-channel noise as the noise can also be spatially nonstationary [86].

Through, at least, the past 40 years, the multichannel noise characteristics have been the foundation of several multichannel noise reduction methods such as the minimum variance distortionless response (MVDR) beamformer [44] and, more recently, the multichannel quotient singular value decomposition (QSVD) based Wiener filter [56]. Despite the importance of knowing the multichannel noise characteristics in noise reduction methods, the estimation of the multichannel noise has apparently been considered less than estimation of single-channel noise. A common approach to the multichannel estimation problem is to use energy-based activity detection. In many methods based on this approach, a single-channel noise power spectral density (PSD) estimator is applied on each sensor signal, and the resulting estimate is used to detect the presence of the desired signal. When the desired signal is not present on a single channel, the noise PSD can be updated. If the desired signal is absent on two or more channels, the noise CPSD between these sensors can also be updated. Examples of methods employing this approach are found in [18, 68, 157, 207]. However, like the single-channel activity detection based noise estimators, the multichannel counterparts are vulnerable to nonstationary noise which is frequently encountered in practice.

Another and more recent group of multichannel noise estimators are based on making assumptions on the type of noise field. These methods do not need activity detection, however, their practical applicability are restricted by the noise assumptions that are not always realistic. An example of a noise type that enables multichannel noise estimation without activity detection is the diffuse noise field as exploited by the methods proposed in [125, 156]. The method in [125] utilizes the fact that the noise CPSD matrix for a diffuse noise field can be decomposed into the noise PSD and a matrix completely determined by the noise coherence function. Under the assumed noise field condition, the noise coherence function is known [80]. The other method in [156] explicitly exploits the noise field assumption by only computing the real part of the noise CPSD matrix.

Recently, a more general multichannel noise estimation method was proposed in [86] that can be applied to the noise estimation on both spatially and temporally nonstationary noise fields. In this method, the diagonal terms of the CPSD matrix are estimated using traditional single-channel noise PSD estimators as previously described, while the off-diagonal terms are updated recursively from the DFT coefficients related to the observed data. It should be emphasized, however, that the method in [86] assumes that the propagation of the desired signal is known which might not be the case in practice.

2.2 Fundamental Frequency Estimation

While most enhancement methods for, e.g., periodic signals are driven by a noise estimate, it is also feasible to design enhancement methods driven by parameters or statistics related to the desired signal [48]. An example of a parameter that is applicable in enhancement of periodic signals is the fundamental frequency [98]. Besides being implicitly applicable for enhancement, the fundamental frequency has been proven useful for signal compression [30, 133, 149], signal modification [71, 151], music transcription [22, 105], tuning of musical instruments [33], etc. The usefulness of the fundamental frequency in various applications has resulted in extensive research in fundamental frequency estimation methods.

Single-Channel Fundamental Frequency Estimation

Traditionally, fundamental frequency estimation has been considered as a single-channel estimation problem. Most of the classical fundamental frequency estimators do not assume a model for the periodic signal(s), i.e., we can classify these as non-parametric methods. A popular approach has been to compare the observed signal with a delayed version of the self-same signal by a similarity measure. The rationale behind doing this is that, when the delay corresponds to the reciprocal of the fundamental frequency, the similarity measure should be maximized since the desired signal is periodic. Some of the first methods utilizing this approach used the autocorrelation function (ACF) [153] and the average magnitude difference function (AMDF) [161] as similarity measures. More recent variants of methods using the delay approach can be found in [47, 134, 186]. Another subclass of non-parametric fundamental frequency estimators is based on peak detection. These methods exploits the fact that the peaks of, e.g., the time-series representation [76] or the cepstrum [142] of the observed signal should appear in fixed intervals, where the length of the intervals can be mapped to a fundamental frequency estimate. A third approach to non-parametric fundamental frequency estimation is based on the harmonic product spectrum. In methods based on this approach, the spectrum at the fundamental frequency and multiples thereof are multiplied for different candidate fundamental frequencies [143, 168]. Then, the fundamental frequency estimate is obtained from the maximizer of the so-called harmonic product spectrum (HPS). For an overview of the above and other non-parametric fundamental frequency estimators, see, e.g., [87].

While the non-parametric methods are intuitively sound, they are often relying on several heuristics and suffer from poor resolution. To tackle these issues, research in parametric fundamental frequency estimators has attracted considerable attention in the recent years. In general, the parametric estimators can be divided into three groups of methods [33]:

- statistical methods,
- subspace methods,
- filtering methods.

In the statistical methods, the likelihood or probability of the fundamental frequency is maximized possibly under some noise assumptions (e.g., the noise being white and Gaussian). Examples of maximum likelihood (ML) and maximum *a posteriori* (MAP) probability approaches can be found in [33, 39]. Moreover, examples of other Bayesian approaches the MAP approach can be found in [23, 46, 74]. The statistical methods do often provide efficient estimates, however, they are rather computationally demanding. This has motivated research in other groups of parametric methods such as the subspace methods.

The subspace methods utilize the fact that the space spanned by the observed signal covariance matrix can be divided into two subspaces spanning the signal and the noise subspaces, respectively. The properties of these subspaces can then be exploited for various estimation and identification tasks [108, 190, 191]. To perform the division into subspaces, we can take the eigenvalue decomposition (EVD) of the observed signal covariance matrix and group the eigenvectors and eigenvalues as

$$\begin{aligned}\mathbf{R}_{\mathbf{y}} &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H && (17)\\ &= \begin{bmatrix}\mathbf{S} & \mathbf{G}\end{bmatrix}\left(\begin{bmatrix}\mathbf{\Psi} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} + \sigma^2\mathbf{I}\right)\begin{bmatrix}\mathbf{S}^H\\ \mathbf{G}^H\end{bmatrix}, && (18)\end{aligned}$$

where $\mathbf{R}_{\mathbf{y}} = \mathrm{E}\{\mathbf{y}(n_{\mathrm{t}})\mathbf{y}^H(n_{\mathrm{t}})\}$, $\mathbf{U}$ contains the eigenvectors of $\mathbf{R}_{\mathbf{y}}$, $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues of $\mathbf{R}_{\mathbf{y}}$, $\mathbf{S}$ contains the eigenvectors of $\mathbf{R}_{\mathbf{y}}$ spanning the signal subspace, $\mathbf{G}$ contains the eigenvectors of $\mathbf{R}_{\mathbf{y}}$ spanning the noise subspace, $\mathbf{\Psi}$ contains the eigenvalues related to the signal subspace, $\mathbf{0}$ is a vector containing zeros, σ^2 is the noise variance, and

$$\mathbf{y}(n_{\mathrm{t}}) = \begin{bmatrix}y(n_{\mathrm{t}}) & \cdots & y(n_{\mathrm{t}} - M_{\mathrm{t}} + 1)\end{bmatrix}^T. \quad (19)$$

Recently, fundamental frequency estimators were proposed based on this subspace division, e.g., by exploiting orthogonality of the signal and noise subspaces [36, 38] and the shift-invariance of the signal subspace [35]. One of the disadvantages of the subspace method in [36, 38] is that the cost-function is multimodal and therefore needs to be evaluated over a frequency grid. To avoid this search, we proposed a new subspace-based estimator in paper A based on rooting of the cost-function considered in [38]. Moreover, as we showed, the rooting based estimates often have higher resolution than the estimates obtained using the method [38]. Another disadvantage of the subspace methods is that the noise needs to be white. In practice, this assumption rarely holds, in which case the observed signal needs to pre-whitened [33].

A third group of parametric fundamental frequency estimators is the filtering methods. The concept behind these methods is to design a filter that passes a periodic signal undistorted and apply it on the observed signal. This could, for example, be an M_{t}th order finite impulse response (FIR) filter, resulting in the filter output

$$z(n_{\mathrm{t}}) = \mathbf{h}^H\mathbf{y}(n_{\mathrm{t}}), \quad (20)$$

where the filter vector is defined as

$$\mathbf{h} = \begin{bmatrix}h_0 & \cdots & h_{M_{\mathrm{t}}-1}\end{bmatrix}^H. \quad (21)$$

More specifically, the filter $\mathbf{h}$ is designed such that it passes the harmonics of the periodic signal while the noise is attenuated. Some of the first methods utilizing this approach were based on comb filtering [136, 138], i.e., the filters are designed independent on the noise statistics. As a result of that, these methods are mainly applicable when the noise is white. To loosen up this implicit noise assumption, some optimal filtering based methods were proposed recently [32, 37]. In these methods, the filters are designed to pass the desired, periodic signal undistorted, while minimizing the filter output power. These filtering methods are generally not statistically efficient, but they are excellent for resolving closely spaced sources and are robust against various kinds of noise since no noise assumptions are needed.

The lack of statistical efficiency is a common issue for fundamental frequency estimators based on the data covariance matrix, since data partitioning is needed to obtain a full-rank covariance matrix estimate in form of the sample covariance matrix. This will effectively reduce the spectral resolution of these fundamental frequency estimators. To mitigate this issue, we consider the use of a method for obtaining a full-rank covariance matrix estimate without data partitioning in conjunction with an optimal filtering-based fundamental frequency estimator in paper B. By doing this, we can obtain a fundamental frequency estimator with a significantly better spectral resolution compared to if the sample covariance matrix is used. The computational complexity is also a problem for some of the filtering-based fundamental frequency estimators, since their cost-functions are multimodal with narrow peaks requiring a fine search grid. However, the complexity can be lowered significantly by exploiting matrix structures and using time-recursive updates as we propose in paper C. For an overview of the mentioned and other parametric methods, we refer to [33, 39].

Multichannel Fundamental Frequency Estimation

As hinted previously, fundamental frequency estimation from noisy, single-channel and periodic signals has been a popular research topic for decades. However, the multichannel estimation problem has received far less attention. Of course, one could argue that the multichannel fundamental frequency estimation problem can be considered as a single-channel estimation problem by applying beamforming on the multi-channel observed signal. It can be shown, however, that such a cascaded procedure can degrade the precision of the fundamental frequency estimate. Followingly, we mention a few examples of existing multichannel fundamental frequency estimators.

Quite a few heuristically motivated fundamental frequency estimators for multichannel signals have been proposed recently. For example, in [72], Gerkmann et al. proposed two multichannel fundamental frequency estimators. Both estimators were based on a preprocessing step followed by cepstrum based fundamental frequency estimation as in [142]. For preprocessing, they proposed to either average the cepstrum coefficients across all channels, or to find the cepstrum coefficients from the output of a delay-and-sum (DSB) beamformer. Another method proposed by Armani and Omologo in [3], was based on the autocorrelation approach [153]. In this method, the ACF for each channel was normalized before, eventually, being weighted and summed. However, it was not explained in [3] how the weights should be determined. The fundamental frequency estimate was then obtained by maximizing the sum of weighted ACFs with respect to the fundamental period τ_0[2]. Later, Flego and Omologo proposed yet another multichannel fundamental frequency estimator based on the maximization of the so-called *multi-microphone periodicity function* (MPF) [64, 65]. Their method

[2] We define the reciprocal of the fundamental frequency as the fundamental period.

can be interpreted as an autocorrelation approach, being different from the aforementioned autocorrelation approaches in the choice of weights; in the MPF-based method, the weight are found using the Cauchy-Schwarz inequality [187]. It was shown that the MPF based method in [64, 65] outperforms the weighted ACF approach in [3] in terms of gross error rate (GER), whereas Gerkmanns cepstrum based methods [72] were never compared to the others.

As mentioned previously, the just described estimators are based on several heuristics and will therefore most likely not be statistically efficient. This is in contrast to two recently proposed and statistically motivated multichannel fundamental frequency estimators. First, Chan et al. proposed a weighted least squares (WLS) approach in [24]. Their method consists of two steps: first, a set of unconstrained frequencies are estimated using an iterative WLS method, and, then, the fundamental frequency is estimated from the unconstrained frequencies, again, using a WLS approach. Then, Christensen proposed a ML based method in [31]. This method yields an ML estimate of the fundamental frequency when the noise on each channel is white Gaussian even when the noise variances for the different channels are different. The methods in [24, 31] are statistically efficient when their respective assumptions are met. However, they have not been thoroughly evaluated in other scenarios.

2.3 Direction-of-Arrival Estimation

Another signal and noise related parameter that is essential in various multichannel enhancement methods is the DOA onto the array of sensors [11, 145]. If the DOA of the signal is known, the observed signal can be spatially preprocessed to attenuate signal components impinging from all other directions. Moreover, if the DOA of the noise is known it can be canceled out by explicitly placing a null in that direction. Apart from noise reduction, knowledge of the DOA is important in many other applications such as automated camera steering [200], wafer-mask rotational alignment in very-large-scale integration (VLSI), and autonomous vehicles [1].

Decades of research have resulted in numerous methods for DOA estimation; generally, these can be classified as either narrowband or broadband methods. The definition of a narrowband signal is that it has a bandwidth which is small compared to its center frequency and vice versa for broadband signals. According to the models in (9) and (10), DOA estimation of a periodic signal can be considered as either the problem of estimating the DOA of L narrowband signals, or the DOA of a broadband signal. Followingly, we review some of the popular approaches to narrowband and broadband DOA estimation, respectively.

Narrowband DOA Estimation

For narrowband signals, the DOA estimation problem boils down to the estimation of the so-called spatial frequency ω_{s} defined as [181]

$$\omega_{\mathrm{s}} = \omega_{\mathrm{c}} \frac{d \sin \theta}{c}, \tag{22}$$

where ω_{c} is the center frequency. That is, the narrowband DOA estimation problem basically resembles the problem of estimating the frequency of a sinusoid when ω_{c}, d and c are known. Since narrowband DOA estimation resembles frequency estimation when only the DOA is unknown, frequency estimation approaches such as statistical, subspace and filtering methods (aka. beamforming in the array signal processing literature) as described in Section 2.2 can be utilized. In fact, the subspace methods were originally developed for DOA estimation.

The most well known and frequently used approach for DOA estimation may be to apply the ML principle. This approach is based on modeling both the desired signal and the noise. In the array signal processing literature, two different models for the desired signal have been considered: a deterministic model and a stochastic model. This has resulted in two ML-based methodologies for DOA estimation, namely deterministic ML (DML) [108, 109, 192] and stochastic ML (SML) [16, 96] methods, respectively. It has been shown that the SML methods have a better large sample accuracy compared to the corresponding DML methods, especially when the number of sensors is small, the SNR is low, and in scenarios with highly correlated signals [108, 147].

While the ML-based methods provides good DOA estimates, they suffer from a high computational complexity due to high dimensional searches. Therefore, alternative approaches have been considered. Another common approach to DOA estimation is the subspace approach as described for fundamental frequency estimation. To summarize, these methods are based on exploiting certain properties of the so-called signal and noise subspaces. Examples of popular methods utilizing the subspace approach are the multiple signal classification (MUSIC) [166], the estimation of signal parameters by rotational invariance techniques (ESPRIT) [163], the minimum norm (Min-Norm) [110], and the weighted subspace fitting (WSF) [196] methods. Note that the ESPRIT method is only applicable to some array structures such as the ULA. The MUSIC, Min-Norm, and ESPRIT methods have all been shown to have a good statistical performance, i.e., the variance of the estimators are close to the Cramér-Rao bound (CRB), while the WSF method is statistically efficient. On a side note, it has been shown that the MUSIC method is a large sample realization of the DML method [182], and that the WSF method has the same asymptotic properties as the SML method [147].

The last group of narrowband DOA estimators considered here is beamforming methods. The concept in these methods is to steer a spatial filter in different directions while measuring the output power of the filter. The DOA estimate is then obtained by maximizing the output power with respect to the steering direction. Many beamforming methods exist [192], with the delay-and-sum beamformer (DSB) (aka. conventional beamforming) [193] and Capon's beamformer (aka. the minimum variance distortionless response (MVDR) beamformer) [21] being two widely used methods. The DSB is a data independent beamformer, i.e., its spatial response is designed without knowledge of the observed, desired or noise signal. Due to this, the DSB generally has poor noise reduction capabilities unless the noise is spatially white. Capon's beamformer, on the other hand, is designed to pass the desired signal undistorted while attenuating the noise as much as possible resulting in better noise reduction compared to the DSB. In general, however, the beamforming methods do not provide as accurate DOA estimates as the statistical and subspace methods [108], but filtering methods have proven useful for resolving closely spaced sources [39].

Broadband DOA Estimation

The narrowband DOA estimators are only usable in applications where the desired signal is indeed narrowband such as in radar and communication. However, in many applications such as enhancement of multichannel periodic signals, the desired signal is most likely broadband. Here, we consider the problem of estimating the DOA of a broadband signal which can be tackled in several different ways. Most existing solutions to the broadband DOA estimation problem can be loosely divided into three groups [50, 52]: beamforming methods, high-resolution spectral estimation methods, and time difference of arrival (TDOA) methods. Note that classifying broadband DOA estimators can be misleading as some estimators belong to more than one of the mentioned classes.

Many of the beamforming methods derived for narrowband signals have also been generalized to be applicable on broadband signals. The generalization is realized by narrowband filtering the observed signal, e.g., using the DFT [108], and then by applying narrowband beamformers on the individual frequency bins. Finally, the DOA is estimated by maximizing the output power, aka. the steered response power (SRP), of the broadband beamformer[3] [52]. The output power is obtained by accumulating the output powers of all narrowband beamformers for each candidate DOA. For broadband beamforming based DOA estimation, the DSB has been commonly used. The DSB can be implemented either in the frequency domain by applying phase shifts in the individual frequency bins corresponding to the DOA, or in the time domain by introducing different delays on each sensor to steer the array.

The DSB can also be written as a sum of cross-correlations between sensor pairs which resembles the so-called SRP method without filtering of the individual sensor signals [52, 53]. The general SRP method relies on a filter-and-sum beamformer (FSB), where a filter is applied to each sensor signal before beamforming. The output of the

[3] Here, a broadband beamformer is defined as a bank of narrowband beamformers

FSB steered in the direction θ at the frequency ω can be written as

$$Y(\omega, \theta) = \sum_{n_\mathrm{s}=0}^{N_\mathrm{s}-1} G_{n_\mathrm{s}}(\omega) X_{n_\mathrm{s}}(\omega) e^{j\omega f_\mathrm{s} \tau_{n_\mathrm{s}}(\theta)} , \tag{23}$$

where $G_{n_\mathrm{s}}(\omega)$ denotes the filter response at frequency ω for the n_sth sensor. A common filter choice is the phase transform (PHAT) [51, 107], i.e., $G_{n_\mathrm{s}}(\omega) = |X_{n_\mathrm{s}}(\omega)|^{-1}$. Basically, this filter whitens the signal in the desired direction. The SRP method combined with PHAT (SRP-PHAT) has been proven useful for broadband DOA estimation in moderate reverberation conditions [52].

A different approach to broadband DOA estimation has been to generalize existing high-resolution, narrowband DOA estimators to the broadband scenario. For example, the narrowband subspace approach has been considered for DOA estimation of broadband signals by dividing the broadband signal into several narrowband signals. In the first methods based on this approach, the DOA was estimated for all the narrowband components, and then combined to yield the DOA estimate of the broadband signal [184]. However, these methods are not applicable in scenarios with a low SNR and/or correlated sources. Later, a method with superior estimation performance compared to the incoherent methods (e.g., [201]) has been devised which processes the narrow bands coherently. The WSF method has also been extended to broadband scenarios; this was considered in [20]. However, the broadband WSF method is difficult to use in reverberant and time-varying environments as it is very sensitive to steering errors [50].

TDOA-based DOA estimation is a third strategy for localization of broadband sources. In methods based on this approach, the DOA is estimated in a two-step procedure; first, the time delays between signals obtained from sensor pairs are estimated, and, then, the estimated time delays are mapped to a DOA estimate. Examples of methods for mapping the TDOA estimates to a DOA estimate are found in [19, 165, 174] and the references therein. The cornerstone in the TDOA approach is the estimation of the time differences. Commonly, these differences have been estimated through maximization of the generalized cross-correlation (GCC) function [107]. The GCC function for the sensors p and q is defined as

$$R_{pq}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Psi_{pq}(\omega) X_p(\omega) X_q^*(\omega) e^{j\omega\tau} d\omega, \tag{24}$$

where $\Psi_{pq}(\omega)$ is a weighting function. We note that the SRP and GCC methods are closely related. In the SRP method, the GCC function between all sensor pairs are summed and the DOA is estimated by maximizing the sum corresponding to the steered response power. In the GCC method, on the other hand, the GCC function is used only for estimating the time differences between the sensor pairs. Different choices of weighting functions have been considered for the GCC method including SNR-based weighting and the PHAT [107]. The TDOA-based DOA estimators are typically outperformed by the SRP methods due to the suboptimal two-step procedure, however, they are less computationally demanding [52]. In summary, many of the broadband DOA estimators have proven useful for speech signals, but they will most likely yield suboptimal DOA estimates, since they do not exploit the underlying model for, e.g., periodic signals.

2.4 Joint Parameter Estimation

Traditionally, fundamental frequency and DOA estimation has been considered as two separate estimation problems. However, there is a number of benefits by estimating these parameters jointly. In a two-source scenario, for instance, the DOA of the sources can not be resolved by only DOA estimation if they share the same DOA. The same problem arises for fundamental frequency estimation of two sources sharing the same fundamental frequency. If the DOA and the fundamental frequency are estimated jointly, it may be possible to resolve such overlapping sources as long as they are separated in one dimension. Moreover, estimating the fundamental frequency and DOA separately will most likely yield estimates with lower accuracy compared to joint estimation methods [99]. These observations have in the recent years inspired researchers to work on the joint fundamental frequency and DOA estimation methods.

The joint frequency and DOA estimator can generally be divided into two groups: methods that jointly estimate the frequency and DOA of single sinusoid, and methods that jointly estimate the fundamental frequency and DOA of a number of harmonically related sinusoids. First, we consider examples of methods for jointly estimating the frequency and DOA of a single sinusoidal source. In [197], a subspace method was proposed that is based on state-space modeling of the observed signal. In this method, the frequency is estimated first whereupon a beamforming like estimate of the DOA is obtained. Another subspace method for joint frequency and DOA estimation was proposed in [112, 113]. This method is an extension of the ESPRIT method for direction finding. In scenarios with white Gaussian noise, the method in [112, 113] outperforms the method in [197] in terms of root mean squared error (RMSE) and vice versa in colored noise scenarios [197]. Later, it has been shown that the MUSIC method can be extended to joint frequency and DOA estimation by using MUSIC iteratively in time and space [120]. This extension showed better estimation accuracy than the ESPRIT-based method in [112, 113]. A different approach was taken in [97] where two-dimensional (2-D) MVDR beamforming was considered. Instead of first applying beamforming and then temporal filtering, a two-dimensional filter is applied on the signal. The frequency and DOA can also be estimated jointly using a multi-stage Wiener filter (MWF) approach [172]. This was realized by using the MWF to obtain a signal subspace estimate, which is then used to obtain the frequency and DOA estimates similar to the approach in [197]. It was shown that the MWF approach outperforms the methods mentioned above (except for the 2-D MVDR method which was not considered in the comparison)

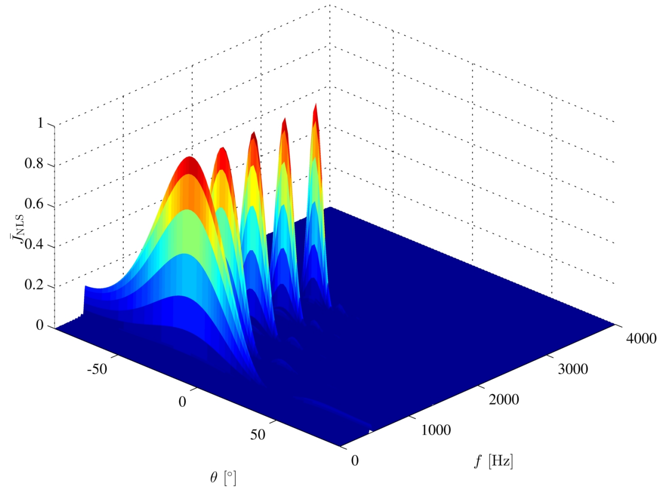


Fig. 4: Pseudo spectrum of a synthetic, multichannel, periodic signal constituted by five harmonics of unit amplitude with $\theta = -23^{\circ}$ and $f_0 = 450$ Hz.

in colored noise in terms of RMSE.

As mentioned previously, many real-life quasiperiodic signals are constituted by a number of harmonically related sinusoids. In Fig. 4, we have illustrated the structure of such signals in form of a pseudo spectrum[4] of a synthetic, multichannel, periodic signal with five harmonics of unit amplitude. For such signals, it is often desired to estimate both the fundamental frequency and the DOA of the quasiperiodic signal, and not only the frequencies and DOAs of the individual harmonics. Recently, quite a few methods for this estimation problem have been proposed. For example, a ML-based estimator was proposed in [150] that jointly estimates the fundamental frequency and TDOA. The method proposed by Qian and Kumaresan is a generalization of the ML fundamental frequency estimator proposed in [203]. Moreover, a few subspace based methods have been considered. In [117], it was proposed to find a joint time delay and fundamental frequency estimate by using the eigenvectors of matrix that was derived from the covariance matrix of the received signals. This subspace method is outperformed by another subspace method proposed in [140] in terms of estimation accuracy. This other subspace method is based on a state-space realization where the fundamental frequency and time delay are estimated from the transition and observation matrices. However, the subspace methods in [117, 140] are only applicable in scenarios with two sensors as opposed to the subspace method by Zhang et al. proposed in [206]. This method can be seen as a generalization of the single-channel fundamental frequency estimator in [36, 38] to multiple channels. The RMSE of the estimates obtained using the method

[4] The pseudo spectrum was obtained using the nonlinear least squares method proposed in paper E.

in [206] was close to the CRB and lower than the RMSE of weighted least squares (WLS) estimator.

The concept of comparing the observed signal with a delayed version of the observed signal in terms of correlation has also been applied to joint time delay and fundamental frequency estimation. This approach has been considered by Képesi et al. in [104, 204]. They introduce the position-pitch plane that can be interpreted as a spatio-temporal ACF, and the time delays and fundamental frequencies of one or more periodic sources are then found from the maxima of this plane. This approach, however, requires several heuristics to obtain good estimates and it is not statistically efficient. Recently, we proposed another joint DOA and fundamental frequency estimator in paper D based on LCMV filtering [54, 69]. The main advantage of this approach is that it can resolve closely space sources as the optimal filters for fundamental frequency estimation. Similarly to the single-channel filtering method in [32, 37], the proposed beamforming- or filtering-based method is based on designing a filter that passes the spatio-temporal periodic signal undistorted while suppressing the noise as much as possible. The DOA and the fundamental frequency are then estimated by maximizing the filter output power for a set of parameter candidates. Despite the high-resolution of this method, it is also statistically inefficient. Therefore, in paper E, we have also proposed a nonlinear least squares (NLS) method for joint estimation that yields maximum likelihood (ML) estimates for periodic signals under three assumptions: the noise should be white Gaussian, the source should be in the far field of the array, and the environment should be anechoic. Even when the assumptions do not hold, the estimator still yields approximately ML estimates.

2.5 Estimation Bounds

When estimating signal parameters such as the fundamental frequency and DOA it is useful to be able to place a lower bound on the variance of the parameter estimate obtained using an unbiased estimator. The bound can be used for determining whether an estimator is the minimum variance unbiased (MVU) estimator or just to benchmark different estimators. Moreover, the bound can reveal if a desired accuracy is obtainable in a given scenario. A few examples of different estimation bounds are the Barankin bound [130], the Seidman bound [170], the Ziv-Zakai bound [208], and the Cramér-Rao bound [45, 158]. Out of these bounds, the CRB is probably the most commonly used as it is the easiest to determine, although other bounds have been proved to be tighter [102]. Therefore, we only consider the CRB in the remainder of this section.

In the derivation of the CRB it is assumed that the probability density function (pdf) $p(\mathbf{x};\boldsymbol{\theta})$ of the observed signal $\mathbf{x}$ satisfies the so-called regularity conditions, i.e.,

$$\mathrm{E}\left\{\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\} = \mathbf{0} \quad \forall \boldsymbol{\theta}, \tag{25}$$

where $\boldsymbol{\theta}$ is a vector containing the unknown signal parameters. Then, the covariance matrix $\mathbf{C}_{\hat{\boldsymbol{\theta}}}$ of any unbiased estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ satisfies [102]

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} - \mathbf{I}^{-1}(\boldsymbol{\theta}) \geq \mathbf{0}, \tag{26}$$

with $\geq \mathbf{0}$ denoting that the matrix is positive semidefinite, and $I(\boldsymbol{\theta})$ is the Fisher information matrix (FIM). The (p, q)th element of the FIM is given by

$$[\mathbf{I}(\boldsymbol{\theta})]_{pq} = -\mathrm{E}\left[\frac{\partial^2 \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial\theta_p \partial\theta_q}\right]. \tag{27}$$

That is, from (26) we can see that the variance of the estimator of the pth parameter $[\boldsymbol{\theta}]_p$ is lower bounded by

$$\mathrm{var}\left([\hat{\boldsymbol{\theta}}]_p\right) \geq [\mathbf{I}^{-1}(\boldsymbol{\theta})]_{pp}. \tag{28}$$

The CRBs for a particular estimation problem can then be found by modeling the pdf $p(\mathbf{x}; \boldsymbol{\theta})$ and by using (27)-(28). The exact CRBs obtained using (28) are most likely given by complex expressions, but for large sample sizes we can often find simpler asymptotic expressions for the CRBs that are easier to interpret (see, e.g., [36, 180]).

In paper E, we derived the exact and asymptotic CRBs for the joint DOA and fundamental frequency estimation problem for ULAs, white Gaussian noise, sources being in the far field of the array, and anechoic environments. In this case, and for a single periodic source, the asymptotic CRBs are given by

$$\mathrm{CRB}(\omega_0) \approx \frac{6}{N_\mathrm{t}^3 N_\mathrm{s}} \mathrm{PSNR}^{-1}, \tag{29}$$

$$\mathrm{CRB}(\theta) \approx \left[\left(\frac{c}{\omega_0 f_\mathrm{s} d \cos\theta}\right)^2 \frac{6}{N_\mathrm{t} N_\mathrm{s}^3} + \left(\frac{\tan\theta}{\omega_0}\right)^2 \frac{6}{N_\mathrm{t}^3 N_\mathrm{s}}\right] \mathrm{PSNR}^{-1}, \tag{30}$$

where the pseudo SNR (PSNR) is defined as

$$\mathrm{PSNR} = \frac{\sum_{l=1}^{L} l^2 A_l^2}{\sigma^2}. \tag{31}$$

These asymptotic expressions hold when the number of samples and the number of sensors are large. From these bounds, we can see that it is advantageous to estimate the DOA and the fundamental frequency jointly. First, taking the harmonic structure into account decreases the CRB for the DOA, since it depends on the PSNR. Moreover, the CRB of the pitch decreases linearly as a function of the number of sensors.

2.6 Summary

In this section, we have considered the estimation of the statistics and parameters needed in most reduction methods for periodic signals. As it appears from these considerations, both estimation of the noise statistics and estimation of signal parameters such as the fundamental frequency and the direction-of-arrival have been popular research topics for several decades. Although several methods for estimating these quantities have been proposed, there are still unsolved problems.

For example, many fundamental frequency estimators are based on the covariance matrix of the observed signal. As the covariance matrix is unknown in practice it is typically replaced by the sample covariance matrix. Doing this, however, has a detrimental impact on the spectral resolution of such estimators, since this covariance matrix estimate requires data partitioning. Another example of an unsolved problem was how to obtain joint and statistically efficient estimates of the fundamental frequency and the DOA of periodic signals such as speech. In the papers A–E, we have proposed solutions for these and other important research problems. For a more detailed overview of these contributions, we refer to Section 4.

3 Noise Reduction

Equipped with an estimate of either the noise statistics or relevant signal parameters such as the fundamental frequency and DOA, the second step in most enhancement methods is to attenuate or, ideally, to remove the noise from the observed noisy signal. In this section, we consider the noise reduction problem, and describe different solutions to it for both single-channel and multichannel scenarios. Moreover, we describe several performance measures that can be used to quantify the performance of noise reduction methods.

3.1 Single-Channel Noise Reduction

The research area of noise reduction, and single-channel noise reduction in particular, is well established. That is, a multitude of different techniques for combating this problem have been proposed.

Spectral-Subtractive Methods

For the single-channel case, one of the first popular algorithms is the spectral-subtractive algorithms proposed in [17] by Boll. In this algorithm, it is assumed that the noise additive and that the noise amplitude spectrum can be estimated by other means. Then, the desired signal spectrum $X(\omega)$ can be estimated by subtracting the estimated noise amplitude spectrum $|\hat{V}(\omega)|$ from the observed signal spectrum $Y(\omega)$ as

$$\hat{X}(\omega) = \Big[|Y(\omega)| - |\hat{V}(\omega)|\Big] e^{j\phi_y(\omega)}, \tag{32}$$

where $\phi_y(\omega)$ is the phase spectrum of the observed signal $y(n_t)$. Note that it is implicitly assumed by (32) that the phase of the noise can be replaced by the phase of the observed signal which is a reasonable assumption when the SNR in all frequency bands is larger than 8 dB [124, 194]. However, inaccuracies in the noise amplitude spectrum estimate may result in negative spectral components in the estimated desired signal spectrum. This issue was originally resolved by half-way rectification [17], i.e.,

$$\hat{X}(\omega) = \begin{cases} \left[|Y(\omega)| - |\hat{V}(\omega)|\right] e^{j\phi_y(\omega)}, & \text{for } |Y(\omega)| > |\hat{V}(\omega)| \\ 0, & \text{otherwise} \end{cases}. \tag{33}$$

When an estimate of the desired signal spectrum is obtained, the desired signal can be estimated by applying the inverse DFT.

Unfortunately, the simple and intuitively sound spectral-subtractive method in [17] suffers from so-called musical noise in the estimated desired signal [14]. To alleviate this, several modifications to the method have been proposed. Examples of such modifications are to overestimate the noise power spectrum and use spectral flooring [14], to use frequency dependent oversubtraction [101, 123, 173], to apply adaptive gain averaging [79], or to use perceptual weighting [199]. The improvements obtained by these modifications come at the cost of the introduction of additional heuristics that may be difficult to control in practice though.

Filtering Methods

Another well-known class of noise reduction methods is the filtering methods. Many of these methods are based on optimal filtering techniques such as Wiener filtering [202]. Let us assume that the order M_{t} filter $\mathbf{h}$ under consideration has a finite impulse response such that the filter output is given by

$$z(n_{\text{t}}) = \mathbf{h}^H \mathbf{y}(n_{\text{t}}). \tag{34}$$

Then, the Wiener filter is obtained by minimizing the minimum mean-squared error between filter output and the desired signal $x(n_{\text{t}})$. Considering time-domain Wiener filtering for noise reduction, this filter design problem can be written as,

$$\mathbf{h}_{\text{W}} = \arg\min_{\mathbf{h}} \text{E}\left\{|x(n_{\text{t}}) - z(n_{\text{t}})|^2\right\}. \tag{35}$$

Although the Wiener filter is optimal in the MSE sense, its noise reduction capabilities partly come at the cost of distortion of the desired signal. Therefore, different extensions of the Wiener filter have been proposed in, e.g., [26, 119] that enables control of the noise reduction and the distortion of the desired signal. Another extension of the Wiener filter was proposed in [118] that designs the filter iteratively without an estimate of the desired signal or noise spectrum. The Wiener filter and the variants thereof mentioned above are all derived for stationary signals. When the signals are nonstationary, the Wiener filters can be extended to handle such signals by means of Kalman filtering (see, e.g., [148]).

While the observed signal is indeed captured in the time-domain, the Wiener filtering methods for noise reduction are often derived and implemented in the frequency domain. Conducting the filtering operation in the frequency domain has advantages such as computationally more efficient implementations and easier monitoring and analysis of the performance [7]. Recently, the noise reduction Wiener filter was also derived in the Karhunen-Loève expansion (KLE) domain, and in a generalized transform domain [7, 8]. In these domains, the speech and noise may be better separated. Moreover, when the filters are derived in the KLE domain, there is no aliasing problem [7].

Recently, a new class of noise reduction filters were proposed based on a orthogonal decomposition of the desired signal defined as

$$\mathbf{x}(n_\mathrm{t}) = \begin{bmatrix} x(n_\mathrm{t}) & \cdots & x(n_\mathrm{t} - M_\mathrm{t} + 1) \end{bmatrix}^T = \boldsymbol{\rho}_{\mathbf{x}x} x(n_\mathrm{t}) + \mathbf{x}_\mathrm{i}(n_\mathrm{t}), \tag{36}$$

where $\boldsymbol{\rho}_{\mathbf{x}x} = \mathrm{E}\{\mathbf{x}(n_\mathrm{t})x(n_\mathrm{t})\}/\mathrm{E}\{x^2(n_\mathrm{t})\}$ is the normalized correlation vector with respect to $x(n_\mathrm{t})$ and $\mathbf{x}_i(n_\mathrm{t})$ is the interference signal vector being defined similarly to $\mathbf{y}(n_\mathrm{t})$ and $\mathbf{x}(n_\mathrm{t})$. That is, by applying this decomposition, we get an extra noise term in form of $\mathbf{x}_i(n_\mathrm{t})$. Noise reduction filters can then be designed that attenuates both the interference term $\mathbf{x}_i(n_\mathrm{t})$ and the noise $\mathbf{v}(n_\mathrm{t})$. Some well-known filtering techniques such as the maximum (max) SNR, Wiener, MVDR, trade-off and LCMV filters were recently rederived in this framework in both the time [6, 27] and frequency [9, 10] domains. The time-domain versions of these filters were derived to be causal. However, as we showed in paper F, the output SNR of the filters can be improved without increasing the distortion by allowing non-causality in the filter design.

The aforementioned filters for noise reduction are typically derived designed using an estimate of the noise statistics. The noise statistics, are difficult to estimate during presence of the desired signal as described previously, which make these filtering method vulnerable to nonstationary noise. However, when the underlying model of the desired signal is known, this can be exploited in the filter design to avoid the need for an explicit estimate of the noise statistics. When the desired signal is periodic, for example, the harmonic structure can be taken into account. An example of this is in [138], where Nehorai and Porat proposed an infinite impulse response (IIR) comb filtering method for enhancing periodic signals. The comb filter is designed to pass the harmonic components undistorted while attenuating other frequency component. The filter design is independent of the observed signal and is therefore implicitly designed under a white noise assumption. More recently, optimal filters were designed for enhancement of periodic signals without any noise assumptions [34]. These filter designs can be seen as extensions of the MVDR [21] and amplitude and phase estimation (APES) [116] filters.

The robustness that the above fundamental frequency driven filters have against nonstationary noise can not be obtained without paying a price though. In reality, signals such as voiced speech can only be approximately modeled as quasiperiodic signals. That is, when using these filters in practice, distortion of the desired signal will happen. In paper G, however, we propose the joint use of two filters driven by the noise statistics and the fundamental frequency, respectively, for speech enhancement. By doing so, the complementary advantages of both filters can be obtained.

Subspace Methods

As in the subspace methods for parameter estimation, the subspace methods for noise reduction are based on a decomposition of the space spanned by the observed signal covariance matrix into a signal and a noise subspace. This decomposition can be realized by applying orthogonal matrix factorizations such as the singular value decomposition (SVD) or the eigenvalue decomposition (EVD), and by then grouping the singular vectors and values or eigenvectors and -values. The SVD-based subspace approach to enhancement was proposed in [49] for white noise scenarios. In this approach, the observed signal is organized in a matrix, e.g., in a Toeplitz or Hankel structure; then, the SVD is applied on this matrix. An enhanced version of the desired signal is then obtained in the transform domain, e.g., by discarding the least significant singular vectors and values. As the noise in most practical scenarios is colored, several extensions for the SVD-based noise reduction method have been proposed to loosen the white noise assumption. This can be accomplished by applying prewhitening or by embedding prewhitening in the method as considered in [84, 100].

As mentioned previously, the EVD can also be used to divide the space spanned by the observed signal into signal and noise subspaces. This can be performed by applying the EVD on the covariance matrix of the observed signal as considered by Ephraim and Van Trees in [62]. Again, an estimate of the desired signal is then obtained in the transform domain by, for example, identifying the eigenvectors spanning the signal subspace, and by then using these eigenvectors to project the observed signal onto the the signal subspace. As the SVD-based approach in [49], the EVD-based noise reduction method in [62] was derived under a white Gaussian noise assumption. Examples of extensions of the method to colored noise scenarios can be found in, e.g., [115, 135, 159].

Statistical-Model based Methods

The last approach to single-channel noise reduction mentioned here is the statistical-model based. In these methods, the amplitude spectrum is estimated using nonlinear estimators on basis of the observed signal and the pdfs of the noise and desired signal DFT coefficients. One of the first of such statistical-model based noise reduction methods was proposed by McAulay and Malpass in [131]. They proposed to use a ML estimate of the spectral amplitudes on basis of a two-state model of the observed signal and soft-decision filtering based on the presence probability of the desired signal. Two other examples of well-known statistical-model based noise reduction methods are the MMSE and log-MMSE methods in [60, 61]. In the MMSE method in [60], the noise reduction is based on minimizing the MSE between the estimated short time spectral amplitude of the desired signal and its true value on basis of modeling the desired signal and noise spectral components as independent Gaussian random variables. This procedure was modified slightly in the log-MMSE method in [61] by minimizing instead the MSE of the log spectral amplitudes. This modification is motivated by the human perception of sounds.

3.2 Multichannel Noise Reduction

In many applications, signal observations from multiple sensors are available. That is, the just described single-channel noise reduction methods are not directly applicable in these applications if only a single enhanced output is desired. While multichannel noise reduction has not received as much research attention as single-channel noise reduction, quite a number of approaches for the multichannel problem have been considered. Generally speaking, two different types of multichannel noise reduction methods have been considered since the spatial and temporal filtering can be treated either separately or jointly.

Separate Spatial and Temporal Filtering

In separate spatial and temporal filtering methods, the noise can, e.g., be reduced by first performing spatial filtering. Spatial filtering can be conducted by using some of the well-known beamforming techniques also described for DOA estimation. Most original beamforming techniques, however, were derived for narrowband signals, but they can be extended to noise reduction of broadband signals by decomposing the observed signal into subbands and by then applying narrowband beamformers on the subband signals [11]. Examples of narrowband beamformers or spatial filters that can be applied in the subbands are the DSB [193], the maximum SNR filter [2, 11], and Capon's beamformer [21].

Unfortunately, extending the well-known narrowband beamforming methods to noise reduction of multichannel periodic broadband signals is problematic. The reason is that the beampatterns for the narrowband beamformers for the different frequency bands will be different; in general, the beamwidth of these beamformers will decrease for an increasing frequency. The effect of this is that the sources impinging from different DOAs than the look direction will be lowpass filtered, which results in disturbing artifacts, e.g., when enhancing audio and speech [11]. This problem can be dealt with by designing a response-invariant beamformer, i.e., a broadband beamformer having the same beamwidth for all frequencies. We can achieve this by means of a proper sensor placement, for example, by using harmonically nested subarrays [103, 127]. The subarrays are used for beamforming in the different frequency bands. To make the beamformers frequency invariant for their respective frequency band, the nested array structuring can be combined with FSB [29, 69].

The previously mentioned single-channel noise reduction methods can then be applied on the beamformer output as considered in [95] to achieve even further noise reduction. Alternatively, a single-channel noise reduction method can be applied on the signals observed on all sensors, and then a beamformer could be applied on the outputs of these filters as considered in [84]. Conducting the spatial and temporal filtering in two steps may not be optimal though, as the linear transformation in the first step can influence the maximum achievable noise reduction in the next step.

Joint Spatial and Temporal Filtering

As mentioned for parameter estimation of multichannel signals, it can be beneficial to process the observed signal jointly in time and space in multichannel scenarios. In the recent years, it has been shown how well-known filtering methods can be extended for such joint processing. For example, Doclo and Moonen considered time-domain multichannel Wiener filtering (MWF) for noise reduction in [56]. In their method, they used the generalized SVD (GSVD) to implement the multichannel Wiener filter. Later, the MWF was combined with the general sidelobe canceler (GSC) [78] in a generalized scheme in [179]. In this scheme, also termed the spatially pre-processed speech distortion weighted MWF (SP-SDW-MWF), the GSC was applied for spatial preprocessing and an adaptive noise canceler was designed to enable trading off noise reduction for less distortion of the desired signal. Originally, the SP-SDW-MWF was derived in the time-domain, but it can also be derived in the frequency domain as shown in [57]. The frequency domain parametric MWF (PMWF) [178] is another example of a multichannel filter that enables a trade off between speech distortion and noise reduction. It can be shown that the frequency domain MVDR filter for multichannel noise reduction presented in [11] is a special case of PMWF [178].

Another multichannel noise reduction method that shows better performance than the MWF in many scenarios in terms of both SNR and signal distortion is the spatio-temporal prediction approach [11]. It is well-known that temporal prediction implicitely plays a fundamental role in enhancement, so therefore spatial prediction was also considered. Theoretically, the spatio-temporal prediction approach should have lower output SNR than the Wiener filter, but, on the other hand, it may introduce less distortion of the desired signal. The Kalman filter for single-channel noise reduction of speech [73, 148], has also been extended to the multichannel scenario [11]. While the multichannel Kalman filter performs both dereverberation and noise reduction, it may be impractical in many scenarios as it requires knowledge of the signal parameters in form of autoregressive (AR)parameters and of the impulse responses from the source to the microphones. The last group of multichannel filtering methods for noise reduction considered here is the recently proposed orthogonal decomposition based filters. These are basically generalizations of the corresponding orthogonal decomposition based single-channel filters to the multichannel case. Examples of noise reduction filters derived in this framework are the max SNR filter, the Wiener filter, the MVDR filter, the LCMV filter, and the trade-off filter [6].

The aforementioned multichannel filtering methods for noise reduction typically use an estimate of the noise statistics of the filter design. As mentioned for the single-channel noise reduction problem, however, this makes the methods vulnerable against nonstationary noise. We have therefore proposed a fundamental frequency driven filter design for multichannel signals in paper I. The proposed filter can be seen as an extension of the APES filter [116].

3.3 Performance Measures

When designing and evaluating both single-channel and multichannel methods, it is essential to have relevant performance measures. In this section, we introduce a number of both objective and subjective performance measured related to noise reduction. For more details on these and other performance measures, we refer the interested reader to [8, 11, 13, 26, 48, 124].

Objective Noise Reduction Measures

One of the most important and well-known measures for noise reduction is the SNR [8]. Before any processing of the observed signal, the SNR is defined as the ratio between the power of the desired signal and the noise, i.e.,

$$\mathrm{iSNR} = \frac{\sigma_x^2}{\sigma_v^2}, \tag{37}$$

with σ_x^2 and σ_v^2 being the signal and noise variances, respectively. This SNR is also commonly referred to as the fullband input SNR (iSNR). The iSNR can also be defined for a frequency subband ω as

$$\mathrm{iSNR}(\omega) = \frac{\phi_x(\omega)}{\phi_v(\omega)}, \tag{38}$$

where $\phi_a(\omega)$ is the PSD of the signal $a(n_\mathrm{t})$ at frequency ω. The goal in many noise reduction methods is to improve the iSNR, i.e., the so-called output SNR (oSNR) should be greater than the iSNR. The oSNR is defined as the ratio between the variances of the desired signal and the noise after noise reduction

$$\mathrm{oSNR} = \frac{\sigma_{x,\mathrm{nr}}^2}{\sigma_{v,\mathrm{nr}}^2}. \tag{39}$$

Similarly to the iSNR, the oSNR can also be defined in frequency subbands. Another important measure for noise reduction is the noise reduction factor ξ_nr [26]. The noise reduction factor is defined as the ratio between the variance of the noise before and after filtering. That is,

$$\xi_{\mathrm{nr}} = \frac{\sigma_v^2}{\sigma_{v,\mathrm{nr}}^2}. \tag{40}$$

We note from this expression that the noise reduction factor should be greater than 1 if noise reduction is desired. For the orthogonal decomposition based noise reduction methods, the abovementioned noise reduction measures are defined in a slightly different way as the desired signal vector contains an interference term [6].

Objective Distortion Measures

When it comes to measuring the distortion of the desired signal, we have at least two widely used measures. First, we have the signal distortion index introduced in [26]. This index is given by the ratio between the power of the difference between the filtered and unfiltered desired signals, and the power of the desired signal. Mathematically speaking, we can write this as

$$\nu_{\mathrm{sd}} = \frac{\mathrm{E}\left\{|x_{\mathrm{f}}(n_{\mathrm{t}}) - x(n_{\mathrm{t}})|^2\right\}}{\sigma_x^2}, \tag{41}$$

where $x_{\mathrm{f}}(n_{\mathrm{t}})$ is the filtered desired signal at time instance n_{t}. The signal distortion index can be generalized to support subband measuring and scenarios where more than one sample of the desired signal are estimated simultaneously from the observed signal [8]. The distortion of the desired signal can also be measured using the signal-reduction factor. This factor is defined similarly to the noise-reduction factor, i.e., it is the ratio between the power of the desired signal before and after noise reduction. We can write this equivalently as

$$\xi_{\mathrm{sr}} = \frac{\sigma_x^2}{\sigma_{x,\mathrm{nr}}^2}. \tag{42}$$

Like the previously mentioned noise reduction measure, these distortion measures can be extended to support the orthogonal decomposition based noise reduction methods. If no distortion of the desired signal is desired, the signal-reduction factor should be equal to 1. This is not a sufficient condition, however, as the desired signal may still be distorted in frequency subbands. In paper G, we proposed an alternative distortion measure for periodic signals, namely the harmonic distortion measure. This measure is defined as the sum of the absolute differences between the powers of the harmonics before and after filtering, i.e.,

$$\xi_{\mathrm{hd}} = 2\sum_{l=1}^{L} |P_l - P_{\mathrm{f},l}|, \tag{43}$$

where $P_l = |\alpha_l|^2$ and $P_{\mathrm{f},l}$ is the power of the lth harmonic after filtering. When the harmonic distortion equals 0, none of the harmonics are distorted.

Objective Hybrid Measures

Besides the aforementioned noise reduction and distortion measures, there also exist hybrid objective measures that measures the overall performance, i.e., both noise reduction and distortion. An example of such a measure is the log-spectral distance (LSD) that can be used to estimate the distance between the spectra of, e.g., a desired signal and an estimated desired signal [195]. This distance is defined as

$$\mathrm{LSD} = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\log_{10}\frac{P(\omega)}{\hat{P}(\omega)}\right]^2 d\omega}. \tag{44}$$

Another widely used distance measure is the Itakura-Saito distance (ISD) that is more correlated with the perceptual quality compared to the LSD [25, 89, 90]. The ISD is given by

$$\mathrm{ISD} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\frac{P(\omega)}{\hat{P}(\omega)} - \log\frac{P(\omega)}{\hat{P}(\omega)} - 1\right] d\omega. \tag{45}$$

For speech and audio applications, two other commonly used objective measures are the perceptual evaluation of speech quality (PESQ) [94] measure and the perceptual evaluation of audio quality (PEAQ) measure [91]. The PESQ and PEAQ measures are objective measures that reflects the perceptual quality of speech and audio, respectively, and they were both selected as recommendations from the International Telecommunication Union (ITU).

Subjective measures

The objective measures just described have the advantage that they can be computed directly from the signals. To get subjective measures, on the other hand, it is necessary to conduct experiments with test subjects. An example is in noise reduction of speech and audio, where the perceptual performance can be measured using listening tests in terms of quality and intelligibility. The quality of a speech signal can be measured using various kinds of listening tests. In general, these tests can be grouped into two categories: tests that are based on relative preference, and tests based on assigning a numerical value to the quality of a stimuli. In the relative preference tests, a reference signal is compared with a test signal to measure the subjective difference between these signals. The difference is quantified by methods such as the degradation category rating (DCR) method [43], the A/B comparison [124], or the comparison category rating (CCR) method [124].

The relative preference based methods are excellent for detecting even subtle differences between a reference signal and a test signal. When the number of test signals is large, however, it is preferable to use numerical rating methods as the listening test sessions should not be too long [92]. In these methods, the signals are given an individual score that should reflect their quality. An example of such a test is the MUSHRA test [93] in which the quality of the test signals is measured using the mean opinion score (MOS).

Intelligibility is another subjective measure that can be obtained from listening tests. It can be measured in several different ways, but, in general, three classes of intelligibility measurement methods exists [124]: recognition of syllables made up of meaningless combinations of speech, recognition of single meaningful words, or recognition of meaningful sentences. Using either of these methodologies, the recognition rate is measured which reflects the intelligibility.

3.4 Summary

We considered noise reduction of periodic signals in this section. The noise reduction methods considered are based on either an estimate of the noise statistics or estimates of relevant signals parameters, and the estimation of these quantities was considered in the previous section. Considering speech, as an example of a real-life periodic signal, it is clear that enhancement is a well-established research area. In general, the enhancement methods for speech can be grouped into two classes: those that are based on an estimate of the noise statistics and those that are based on the desired signal, i.e., in the form of relevant signal parameter estimates.

The first class of methods are basically applicable to enhancement of any kind of desired signal as they are fully relying on the noise statistics. The noise statistics, however, are difficult to estimate when the desired signal is present though. These methods are therefore vulnerable against nonstationary noise. The second class of methods, on the other hand, are robust against such noise as they rely on estimates of signal parameters such as the fundamental frequency and the DOA. The price to pay for this robustness is an increased distortion of the desired signals, since the assumed signal model will not be exact in practice. Therefore, in the papers G–H, we proposed joint filtering schemes for reduction of nonstationary noise in speech signals. In these schemes, we use filters based on a noise statistics estimate and a fundamental frequency estimate, respectively, to obtain the complementary advantages of these two different approaches. Moreover, it is well known that in many enhancements methods it is only possible to improve the noise reduction performance by increasing the distortion of the desired signal. In paper F, however, we propose a novel set of non-causal, time-domain filters that can improve the noise reduction performance compared to their causal counterparts without necessarily increasing the distortion. A more detailed overview of our contributions within the research field of noise reduction can be found in the following section.

4 Contributions

In summary, the overall topic of this thesis is enhancement of periodic signals with focus on speech signals as hinted by the title. This research area covers both estimation of either the noise or relevant signal parameters, and noise reduction as depicted in Fig. 3. The main body of this thesis is constituted by the papers A–I that are all related to the abovementioned research topic. The papers A–E deal with the estimation of parameters of a periodic signal. In particular, paper A–C consider the estimation of the fundamental frequency of a single-channel periodic signal, and paper D–E are on the joint estimation of the fundamental frequency and the direction-of-arrival of a multichannel periodic signal. Given knowledge about the fundamental frequency and the direction-of-arrival, we considered noise reduction of periodic signals in the papers G–I. In the papers F–G, we considered single-channel signals, whereas multichannel signals were considered in paper I. Followingly, we provide a more detailed description of the contributions from the individual papers.

Paper A The first paper presented in this thesis considers the estimation of the fundamental frequency. The paper presents a novel fundamental frequency estimator based on rooting of the harmonic MUSIC (HMUSIC) algorithm proposed in [36, 38]. Compared to the HMUSIC method, the proposed method does not require a search grid and, in many scenarios, it has a better spectral resolution.

Paper B We proposed another fundamental frequency in this paper. The proposed estimator is obtained by using the harmonic LCMV filter proposed in [39] in conjunction with the iterative adaptive approach (IAA) [205]. In this paper, the IAA is only used for estimating the covariance matrix of the observed signal. One of the major benefits of estimating the covariance matrix using the IAA is that it enables estimation from only a single signal snapshot while the covariance matrix estimate is still full rank as opposed to when using the sample covariance matrix estimate. As a results of that, the spectral resolution of the proposed method is improved compared when the sample covariance matrix estimate is used in conjunction with LCMV-based fundamental frequency estimator.

Paper C While the fundamental frequency estimator proposed in paper B provides a good spectral resolution, it also has a relatively high computational complexity. To alleviate this issue, we exploited the inherently low displacement rank of the necessary products of Toeplitz-like matrices. By doing this, we reduced the computational complexity of the method in paper B by several orders of magnitude. We also propose an approximative implementation using the preconditioned conjugates gradient method and a Quasi-Newton approach, and, finally, we propose a set of time-recursive implementations; these initiatives lower the computational complexity even further. The difference between the estimates obtained using the direct and fast implementations, respectively, was shown to be negligible, and the time-recursive implementations were shown to be able to track the fundamental frequency of both synthetic and real-life signals.

Paper D In this paper, we proposed two joint fundamental frequency and DOA estimators. Estimating these parameters jointly is advantageous, as it enable us to, for example, resolve sources with overlapping fundamental frequency as long as their DOAs are distinct. The first estimator proposed is based on a filterbank of periodogram-based filters, and it is therefore implicitely derived under a white noise assumption. The other estimator is a spatio-temporal HLCMV filter, i.e., it is signal dependent and useful for any additive noise scenario. At the cost of a higher computational complexity, the latter estimator shows superior estimation performance.

Paper E This paper expands further on joint fundamental frequency and DOA estimation. In this paper, we provide expression for the exact as well as the asymptotic CRBs of the joint estimation problem at hand, and we describe why it is beneficial to estimate the parameters jointly. We also propose a non-linear least squares (NLS) estimator and an approximate NLS (aNLS) estimator for the joint estimation problem. The proposed estimators are applicable on real-life signals, robust against reverberation, and they outperform several existing multichannel fundamental frequency estimators and broadband DOA estimators in terms of estimation performance.

Paper F Here, we turn the focus to noise reduction. In this paper, a novel set of orthogonal decomposition based filters for noise reduction is proposed; these filters can be regarded as generalizations of the time-domain filters in [6] to incorporate non-causality into the filter design. It is shown that the introduction of non-causality can be beneficial from a noise reduction point of view. Besides the filter designs, we proposed some performance measures for the non-causal filters. Moreover, we showed how some of the filters can be updated recursively which, eventually, proves that output SNR always increases if we increase the filter length (provided that the signals are stationary).

Paper G Following the trail of paper F, we consider causal, time-domain filters for noise reduction in this paper. First, the relationship between two novel, time-domain, noise reduction filters is investigated. This investigation reveals that the orthogonal decomposition based MVDR filter is asymptotically equivalent with the harmonic decomposition based LCMV filter. Therefore, as the filters have complementary advantages and disadvantages, we propose a joint filtering scheme employing both filters. Experiments show that the proposed noise reduction scheme outperforms existing noise reduction methods in terms of PESQ scores.

Paper H In this paper, we propose another joint filtering scheme for reduction of nonstationary noise in speech signals. The difference between the scheme considered in this paper and in paper G is the choice of filter for noise reduction. In both papers, a harmonic decomposition based LCMV filter is used for estimating the noise statistics. Then, an orthogonal decomposition based Wiener filter is used for reduction of the noise in this paper, whereas an orthogonal decomposition based MVDR filter was used in paper H. Out of these two joint filtering schemes, the scheme of this paper has the best perceptual performance in terms of PESQ scores.

Paper I The noise reduction filtering methods in the papers F–H were only derived for single-channel signals. However, in this paper, we proposed a novel pitch-based filtering method for noise reduction of multichannel signals. The proposed filter design was inspired by the APES filter that was extended to enhancement of multichannel periodic signals. The experimental results indicate that the proposed filter has better noise reduction properties than the spatio-temporal HLCMV filter, and that the filter can be applied for noise reduction of real-life, multichannel, periodic signals.

Followingly, we present the major general conclusions that can be drawn on basis of the contributions described above. Regarding estimation of the fundamental frequency and DOA, it is the opinion of the author that these parameters should be estimated jointly when both are needed for a number of reasons. First of all, joint estimation enables us to resolve the fundamental frequencies and DOAs if only one of these parameters are sufficiently spaced in a multisource scenario. Furthermore, as we considered in paper E, estimating the parameters separately using a cascade procedure can deteriorate the estimation performance. Finally, estimating the parameters jointly enables us to treat the DOA estimation problem as a narrowband estimation problem, as a otherwise broadband periodic signal can be decomposed into a number of narrowband signals. The second major aspect covered in our contributions is noise reduction of single-channel and multichannel periodic signals. We have showed examples of how information about the fundamental frequency can be incorporated into noise reduction methods without relying fully on the harmonic model of periodic signals.

It is the opinion of the author that continued research in the exploitation of fundamental frequency information in noise reduction of, e.g., speech and audio signals is a key to obtain a robust solution to the difficult problem of nonstationary noise reduction. To achieve such a solution, the author believes that future work on, e.g, improving the robustness of the proposed methods is of great importance. An example of relevant future work in this regard is model selection. Many real-life signals have missing harmonics, so it is important to incorporate such information into fundamental frequency based enhancement methods to optimize the noise reduction performance. Furthermore, the author believes that it is important to investigate how fundamental frequency based enhancement methods can be robustly applied on speech, since only the voiced parts of speech is harmonic. Finally, the author finds it plausible that the harmonic model driven approach to noise reduction considered in this thesis can be applied for tackling the difficult and pertinent problem of dereverberation.

References

[1] H. K. Aghajan and T. Kailath, "Sensor array processing techniques for super resolution multi-line-fitting and straight edge detection," *IEEE Trans. Image Process.*, vol. 2, no. 4, pp. 454–465, Oct. 1993.

[2] S. P. Applebaum, "Adaptive arrays," *IEEE Trans. Antennas Propag.*, vol. 24, no. 5, pp. 585–598, Sep. 1976.

[3] L. Armani and M. Omologo, "Weighted autocorrelation-based f0 estimation for distant-talking interaction with a distributed microphone network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 2004, pp. 113–116.

[4] J. W. Beauchamp, "Time-variant spectra of violin tones," *J. Acoust. Soc. Am.*, vol. 56, no. 3, pp. 995–1004, Sep. 1974.

[5] M. E. H. Benbouzid, "A review of induction motors signature analysis as a medium for faults detection," *IEEE Trans. Ind. Electron.*, vol. 47, no. 5, pp. 984–993, Oct. 2000.

[6] J. Benesty and J. Chen, *Optimal Time-Domain Noise Reduction Filters – A Theoretical Study*, 1st ed. Springer, 2011, no. VII.

[7] J. Benesty, J. Chen, and Y. Huang, "Speech enhancement in the Karhunen-Loève expansion domain," *Synthesis Lectures on Speech and Audio Processing*, vol. 7, no. 1, pp. 1–112, 2011.

[8] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Springer-Verlag, 2009.

[9] J. Benesty and Y. Huang, "A perspective on single-channel frequency-domain speech enhancement," *Synthesis Lectures on Speech and Audio Processing*, vol. 7, no. 2, pp. 1–109, 2011.

[10] ——, "A single-channel noise reduction MVDR filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2011, pp. 273–276.

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

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

[13] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*. Springer-Verlag, 2008.

[14] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, 1979, pp. 208–211.

[15] 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.

[16] J. F. Böhme, “Separated estimation of wave parameters and spectral parameters by maximum likelihood,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 11, Apr. 1986, pp. 2819–2822.

[17] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.

[18] R. L. Bouquin-Jeannès, A. A. Azirani, and G. Faucon, “Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator,” *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 484–487, Sep. 1997.

[19] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, “A closed-form location estimator for use with room environment microphone arrays,” *IEEE Trans. Speech Audio Process.*, vol. 5, no. 1, pp. 45–50, Jan. 1997.

[20] J. A. Cadzow, “Multiple source location - the signal subspace approach,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 7, pp. 1110–1125, Jul. 1990.

[21] J. Capon, “High-resolution frequency-wavenumber spectrum analysis,” *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[22] A. T. Cemgil, “Bayesian music transcription,” Ph.D. dissertation, Nijmegen University, 2004.

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

[24] F. K. W. Chan, H. C. So, W. H. Lau, and C. F. Chan, “Efficient approach for sinusoidal frequency estimation of gapped data,” *IEEE Signal Process. Lett.*, vol. 17, no. 6, pp. 611–614, Jun. 2010.

[25] J. Chen, J. Benesty, Y. Huang, and E. J. Diethorn, “Fundamentals of noise reduction,” in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer-Verlag, 2008, ch. 43, pp. 843–871.

[26] J. Chen, J. Benesty, Y. Huang, and S. Doclo, “New insights into the noise reduction Wiener filter,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

[27] J. Chen, J. Benesty, Y. Huang, and T. Gaensler, “On single-channel noise reduction in the time domain,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2011, pp. 277–280.

[28] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, Sep. 1953.

[29] T. Chou, “Frequency-independent beamformer with low response error,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, May 1995, pp. 2995–2998.

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

[31] ——, “Multi-channel maximum likelihood pitch estimation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012.

[32] M. G. Christensen, J. L. Højvang, A. Jakobsson, and S. H. Jensen, “Joint fundamental frequency and order estimation using optimal filtering,” *EURASIP J. on Applied Signal Processing*, vol. 2011, no. 1, pp. 1–18, Jun. 2011.

[33] M. G. Christensen and A. Jakobsson, “Multi-pitch estimation,” *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[34] ——, “Optimal filter designs for separating and enhancing periodic signals,” *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

[35] 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*, Nov. 2007, pp. 631–635.

[36] ——, “Joint high-resolution fundamental frequency and order estimation,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 5, pp. 1635–1644, Jul. 2007.

[37] 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.

[38] M. G. Christensen, S. H. Jensen, S. V. Andersen, and A. Jakobsson, “Subspace-based fundamental frequency estimation,” in *Proc. European Signal Processing Conf.*, Sep. 2004, pp. 637–640.

[39] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, “Multi-pitch estimation,” *Elsevier Signal Process.*, vol. 88, no. 4, pp. 972–983, Apr. 2008.

[40] M. G. Christensen, P. Vera-Candeas, S. D. Somasundaram, and A. Jakobsson, “Robust subspace-based fundamental frequency estimation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2008, pp. 101–104.

[41] 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, Sep. 2003.

[42] 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, Jan. 2002.

[43] P. Combescure, A. Le Guyader, and A. Gilloire, “Quality evaluation of 32 kbit/s coded speech by means of degradation category ratings,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 7, May 1982, pp. 988–991.

[44] H. Cox, “Resolving power and sensitivity to mismatch of optimum array processors,” *J. Acoust. Soc. Am.*, vol. 54, no. 3, pp. 771–785, Sep. 1973.

[45] H. Cramér, *Mathematical Methods of Statistics*. Princeton University Press, 1999.

[46] 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, Apr. 2006.

[47] 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.

[48] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. Institute of Electrical and Electronics Engineers, 2000.

[49] M. Dendrinos, S. Bakamidis, and G. Carayannis, “Speech enhancement from noise: A regenerative approach,” *Speech Communication*, vol. 10, no. 1, pp. 45 – 57, 1991.

[50] E. D. Di Claudio and R. Parisi, “Multi-source localization strategies,” in *Microphone Arrays - Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Springer-Verlag, 2001, ch. 9, pp. 181–201.

[51] 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.

[52] 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.

[53] J. Dmochowski, J. Benesty, and S. Affes, “Direction of arrival estimation using the parameterized spatial correlation matrix,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 4, pp. 1327–1339, May 2007.

[54] ——, “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.

[55] G. Doblinger, “Computationally efficient speech enhancement by spectral minima tracking in subbands,” in *Proc. Eurospeech*, Sep. 1995, pp. 1513–1516.

[56] 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, Sep. 2002.

[57] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, “Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction,” *Speech Commun.*, vol. 49, no. 7–8, pp. 636–656, Jul. 2007.

[58] H. Dudley, “The carrier nature of speech,” *Bell Syst. Tech. J.*, vol. 19, no. 4, pp. 495–515, Oct. 1940.

[59] V. Emiya, B. David, and R. Badeau, “A parametric method for pitch estimation of piano tones,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 2007, pp. 249–252.

[60] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[61] ——, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

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

[63] J. S. Erkelens and R. Heusdens, "Tracking of nonstationary noise based on data-driven recursive noise power estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 6, pp. 1112–1123, Aug. 2008.

[64] F. Flego, "Fundamental frequency estimation techniques for multi-microphone speech input," Ph.D. dissertation, University of Trento, Mar. 2006.

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

[66] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, 2nd ed. Springer Science+Business Media, Inc., 1998.

[67] J. B. J. Fourier, *Œvres de Fourier*, M. G. Darboux, Ed. Gauthier-Villars (Paris), 1890, vol. 2.

[68] J. Freudenberger, S. Stenzel, and B. Venditti, "A noise PSD and cross-PSD estimation for two-microphone speech enhancement systems," in *Proc. IEEE Workshop Statist. Signal Process.*, Aug. 2009, pp. 709–712.

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

[70] C. W. Garland, J. W. Nibler, and D. P. Shoemaker, *Experiments in Physical Chemistry*. McGraw-Hill Higher Education, 2008.

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

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

[73] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Process.*, vol. 39, no. 8, pp. 1732–1742, Aug. 1991.

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

[75] ——, "Bayesian computational models for inharmonicity in musical instruments," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.*, Oct. 2005, pp. 283–286.

[76] B. Gold and L. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Am.*, vol. 46, no. 2B, pp. 442–448, Aug. 1969.

[77] 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.

[78] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[79] H. Gustafsson, S. E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 799–807, Nov. 2001.

[80] E. A. P. Habets and S. Gannot, “Generating sensor signals in isotropic noise fields,” *J. Acoust. Soc. Am.*, vol. 122, no. 6, pp. 3464–3470, Dec. 2007.

[81] S. Hahn, *Hilbert Transforms in Signal Processing*. Artech House, Inc., 1996.

[82] J. A. Haigh and J. S. Mason, “Robust voice activity detection using cepstral features,” in *Proc. IEEE TENCON*, vol. 3, Oct. 1993, pp. 321–324.

[83] J. H. L. Hansen and M. A. Clements, “Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress,” *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 407–415, Sep. 1995.

[84] P. S. K. Hansen, “Signal subspace methods for speech enhancement,” Ph.D. dissertation, Techn. Univ. Denmark, Lyngby, Denmark, 1997.

[85] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, “The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer,” *J. Acoust. Soc. Am.*, vol. 115, no. 2, pp. 833–843, Jan. 2004.

[86] R. C. Hendriks and T. Gerkmann, “Noise correlation matrix estimation for multi-microphone speech enhancement,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 1, pp. 223–233, Jan. 2012.

[87] W. Hess, *Pitch Determination of Speech Signals - Algorithms and Devices*. Springer-Verlag, 1983.

[88] H. G. Hirsch and C. Ehrlicher, “Noise estimation techniques for robust speech recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 1995, pp. 153–156.

[89] B. Iser, W. Minker, and G. Schmidt, *Bandwidth Extension of Speech Signals*. Springer Science+Business Media, 2008.

[90] F. Itakura and S. Saito, “A statistical method for estimation of speech spectral density and formant frequencies,” *Electron. Commun. Jpn.*, vol. 52A, pp. 36–43, 1970.

[91] ITU-R, “Method for objective measurements of perceived audio quality,” no. BS.1387-1, pp. 1–100, Nov. 2001.

[92] ——, “General methods for the subjective assessment of sound quality,” no. BS.1284-1, pp. 1–13, Dec. 2003.

[93] ——, “Method for the subjective assessment of intermediate quality level of coding systems,” no. BS.1534-1, pp. 1–18, Jan. 2003.

[94] ITU-T, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” no. P.862, pp. 1–30, Feb. 2001.

[95] F. Jabloun and B. Champagne, “A multi-microphone signal subspace approach for speech enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, 2001, pp. 205–208.

[96] A. G. Jaffer, “Maximum likelihood direction finding of stochastic sources: a separable solution,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, Apr. 1988, pp. 2893–2896.

[97] A. Jakobsson, S. L. Jr. Marple, and P. Stoica, "Computationally efficient two-dimensional Capon spectrum analysis," *IEEE Trans. Signal Process.*, vol. 48, no. 9, pp. 2651–2661, Sep. 2000.

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

[99] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Joint direction-of-arrival and fundamental frequency estimation," *IEEE Trans. Audio, Speech, and Language Process.*, pp. 1–11, 2012, submitted.

[100] 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.

[101] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2002.

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

[103] W. L. Kellerman, "Acoustic echo cancellation for beamforming microphone arrays," in *Microphone Arrays - Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Springer-Verlag, 2001, ch. 13, pp. 281–306.

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

[105] A. Klapuri and M. Davy, *Signal Processing Methods for Music Transcription*. Springer Science+Business Media LLC, 2006.

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

[107] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[108] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996.

[109] R. Kumaresan and A. K. Shaw, "High resolution bearing estimation without eigen decomposition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 10, Apr. 1985, pp. 576–579.

[110] R. Kumaresan and D. W. Tufts, "Estimating the angles of arrival of multiple plane waves," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 19, no. 1, pp. 134–139, Jan. 1983.

[111] D. Lederman, E. Zmora, S. Hauschildt, A. Stellzig-Eisenhauer, and K. Wermke, "Classification of cries of infants with cleft-palate using parallel hidden markov models," *Med. Biol. Eng. Comput.*, vol. 46, pp. 965–975, Mar. 2008.

[112] A. N. Lemma, A.-J. van der Veen, and E. F. Deprettere, “Joint angle-frequency estimation using multi-resolution ESPRIT,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, May 1998, pp. 1957–1960.

[113] ——, “Analysis of joint angle-frequency estimation using ESPRIT,” *IEEE Trans. Signal Process.*, vol. 51, no. 5, pp. 1264–1283, May 2003.

[114] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*, 2nd ed. Addison-Wesley Publishing Company, Inc., 1994.

[115] H. Lev-Ari and Y. Ephraim, “Extension of the signal subspace speech enhancement approach to colored noise,” *IEEE Signal Process. Lett.*, vol. 10, no. 4, pp. 104–106, Apr. 2003.

[116] 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.

[117] G. Liao, H. C. So, and P. C. Ching, “Joint time delay and frequency estimation of multiple sinusoids,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, May 2001, pp. 3121–3124.

[118] J. S. Lim and A. V. Oppenheim, “All-pole modeling of degraded speech,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 3, pp. 197–210, Jun. 1978.

[119] J. Lim and A. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proc. IEEE*, vol. 67, no. 12, pp. 1586 – 1604, dec. 1979.

[120] J.-D. Lin, W.-H. Fang, Y.-Y. Wang, and J.-T. Chen, “FSF MUSIC for joint DOA and frequency estimation and its performance analysis,” *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4529–4542, Dec. 2006.

[121] L. Lin, W. H. Holmes, and E. Ambikairajah, “Adaptive noise estimation algorithm for speech enhancement,” *Electronics Lett.*, vol. 39, no. 9, pp. 754–755, May 2003.

[122] ——, “Subband noise estimation for speech enhancement using a perceptual Wiener filter,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 2003, pp. 80–83.

[123] P. Lockwood and J. Boudy, “Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars,” *Speech Commun.*, vol. 11, no. 2–3, pp. 215–228, Jun. 1992.

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

[125] T. Lotter and P. Vary, “Dual-channel speech enhancement by superdirective beamforming,” *EURASIP J. on Applied Signal Processing*, vol. 2006, no. 1, pp. 1–14, Jan. 2006.

[126] 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.

[127] C. Marro, Y. Mahieux, and K. U. Simmer, “Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering,” *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp. 240–259, May 1998.

[128] R. Martin, “Spectral subtraction based on minimum statistics,” in *Proc. European Signal Processing Conf.*, Sep. 1994, pp. 1182–1185.

[129] ——, “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.

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

[131] R. J. McAulay and M. L. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[132] 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, Aug. 1986.

[133] ——, “Sinusoidal coding,” in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 4, pp. 121–173.

[134] 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.

[135] U. Mittal and N. Phamdo, “Signal/noise KLT based approach for enhancing speech degraded by colored noise,” *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 159–167, Mar. 2000.

[136] J. A. 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.

[137] V. K. Murthy, L. J. Haywood, J. Richardson, R. Kalaba, S. Salzberg, G. Harvey, and D. Vereeke, “Analysis of power spectral densities of electrocardiograms,” *Mathematical Biosciences*, vol. 12, no. 1–2, pp. 41–51, Oct. 1971.

[138] 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.

[139] J. Neuberg, R. Luckett, B. Baptie, and K. Olsen, “Models of tremor and low-frequency earthquake swarms on Montserrat,” *J. Volcanol. Geotherm. Res.*, vol. 101, no. 1–2, pp. 83–104, Aug. 2000.

[140] 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.

[141] J. K. Nielsen, M. G. Christensen, and S. H. Jensen, “An approximate Bayesian fundamental frequency estimator,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012.

[142] A. M. Noll, “Cepstrum pitch determination,” *J. Acoust. Soc. Am.*, vol. 41, no. 2, pp. 293–309, Feb. 1967.

[143] 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.

[144] G. L. Ogden, L. M. Zurk, M. E. Jones, and M. E. Peterson, “Extraction of small boat harmonic signatures from passive sonar,” *J. Acoust. Soc. Am.*, vol. 129, no. 6, pp. 3768–3776, Jun. 2011.

[145] M. Omologo, P. Svaizer, and M. Matassoni, “Environmental conditions and acoustic transduction in hands-free speech recognition,” *Speech Commun.*, vol. 25, no. 1–3, pp. 75–95, Aug. 1998.

[146] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 2nd ed. Prentice Hall, Inc., 1999.

[147] B. Ottersten, M. Viberg, and T. Kailath, “Analysis of subspace fitting and ml techniques for parameter estimation from sensor array data,” *IEEE Trans. Signal Process.*, vol. 40, no. 3, pp. 590–600, Mar. 1992.

[148] K. Paliwal and A. Basu, “A speech enhancement method based on Kalman filtering,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 12, Apr. 1987, pp. 177–180.

[149] H. Purnhagen and N. Meine, “HILN - the MPEG-4 parametric audio coding tools,” in *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 3, May 2000, pp. 201–204.

[150] 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.

[151] T. F. Quatieri and R. J. McAulay, “Speech transformations based on a sinusoidal representation,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 6, pp. 1449–1464, Dec. 1986.

[152] B. G. Quinn and P. J. Thomson, “Estimating the frequency of a periodic function,” *Biometrika*, vol. 78, no. 1, pp. 65–74, Mar. 1991.

[153] L. R. Rabiner, “On the use of autocorrelation analysis for pitch detection,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 1, pp. 24–33, Feb. 1977.

[154] L. R. Rabiner and M. R. Sambur, “An algorithm for determining the endpoints of isolated utterances,” *Bell Syst. Tech. J.*, vol. 54, no. 2, pp. 297–315, Feb. 1975.

[155] ——, “Voiced-unvoiced-silence detection using the Itakura LPC distance measure,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 1977, pp. 323–326.

[156] M. Rahmani, A. Akbari, B. Ayad, and B. Lithgow, “Noise cross PSD estimation using phase information in diffuse noise field,” *Signal Process.*, vol. 89, no. 5, pp. 703–709, May 2009.

[157] M. Rahmani, A. Akbari, B. Ayad, M. Mazoochi, and M. S. Moin, “A modified coherence based method for dual microphone speech enhancement,” in *Proc. IEEE Int. Conf. Signal Process. Commun.*, Nov. 2007, pp. 225–228.

[158] C. R. Rao, “Information and the accuracy attainable in the estimation of statistical parameters,” *Bull. Calcutta Math. Soc.*, vol. 37, no. 3, pp. 81–91, 1945.

[159] A. Rezayee and S. Gazor, “An adaptive KLT approach for speech enhancement,” *IEEE Trans. Speech Audio Process.*, vol. 9, no. 2, pp. 87–95, Feb. 2001.

[160] C. Ris and S. Dupont, “Assessing local noise level estimation methods: Application to noise robust ASR,” *Speech Commun.*, vol. 34, no. 1–2, pp. 141–158, Apr. 2001.

[161] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, “Average magnitude difference function pitch extractor,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 22, no. 5, pp. 353–362, Oct. 1974.

[162] T. D. Rossing, F. R. Moore, and P. A. Wheeler, *The Science of Sound*, 3rd ed. Addison Wesley, 2002.

[163] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[164] M. R. Sambur and N. S. Jayant, "LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 6, pp. 488–494, Dec. 1976.

[165] R. O. Schmidt, "A new approach to geometry of range difference location," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-8, no. 6, pp. 821–835, Nov. 1972.

[166] ——, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[167] M. R. Schroeder, "Apparatus for suppressing noise and distortion in communication signals," US Patent 3,180,936, Apr. 27, 1965.

[168] ——, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," *J. Acoust. Soc. Am.*, vol. 43, no. 4, pp. 829–834, Apr. 1968.

[169] ——, "Processing of communications signals to reduce effects of noise," US Patent 3,403,224, Sep. 24, 1968.

[170] L. P. Seidman, "Performance limitations and error calculations for parameter estimation," *Proc. IEEE*, vol. 58, no. 5, pp. 644–652, May 1970.

[171] P. Setlur, F. Ahmad, and M. Amin, "Helicopter radar return analysis: Estimation and blade number selection," *Signal Process.*, vol. 91, no. 6, pp. 1409–1424, Jun. 2011.

[172] T. Shu and X. Liu, "Robust and computationally efficient signal-dependent method for joint DOA and frequency estimation," *EURASIP J. on Advances in Signal Processing*, vol. 2008, no. 1, pp. 1–16, Apr. 2008.

[173] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 328–337, Jul. 1998.

[174] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 12, pp. 1661–1669, Dec. 1987.

[175] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 1998, pp. 365–368.

[176] 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 Applied Signal Processing*, vol. 2005, no. 18, pp. 2954–2964, 2005.

[177] C. Sorin and C. Thouin-Daniel, "Effects of auditory fatigue on speech intelligibility and lexical decision in noise," *J. Acoust. Soc. Am.*, vol. 74, no. 2, pp. 456–466, Aug. 1983.

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

[179] A. Spriet, M. Moonen, and J. Wouters, “Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction,” *Elsevier Signal Process.*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[180] 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, Aug. 1997.

[181] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Education, Inc., 2005.

[182] 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.

[183] 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.

[184] G. Su and M. Morf, “The signal subspace approach for multiple wide-band emitter location,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 6, pp. 1502–1522, Dec. 1983.

[185] J. Tierney, “A study of LPC analysis of speech in additive noise,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 389–397, Aug. 1980.

[186] T. Tolonen and M. Karjalainen, “A computationally efficient multipitch analysis model,” *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 708–716, Nov. 2000.

[187] L. N. Trefethen and D. Bau III, *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, 1997.

[188] R. Tucker, “Voice activity detection using a periodicity measure,” *Proc. IEE*, vol. 139, no. 4, pp. 377–380, Aug. 1992.

[189] D. Van Compernolle, “Noise adaptation in a hidden Markov model speech recognition system,” *Comput. Speech Language*, vol. 3, no. 2, pp. 151–167, Apr. 1989.

[190] A.-J. van der Veen, E. F. Deprettere, and A. L. Swindlehurst, “Subspace-based signal analysis using singular value decomposition,” *Proc. IEEE*, vol. 81, no. 9, pp. 1277–1308, Sep. 1993.

[191] P. van Overschee and B. de Moor, *Subspace Identification for Linear Systems: Theory, Implementation, Applications*. Kluwer Academic Publishers, 1996.

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

[193] B. D. Van Veen and K. M. Buckley, “Beamforming: a versatile approach to spatial filtering,” *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[194] P. Vary, “Noise suppression by spectral magnitude estimation –mechanism and theoretical limits,” *Signal Process.*, vol. 8, no. 4, pp. 387–400, Jul. 1985.

[195] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. John Wiley & Sons Ltd, 2006.

[196] 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, Nov. 1991.

[197] M. Viberg and P. Stoica, “A computationally efficient method for joint direction finding and frequency estimation in colored noise,” in *Rec. Asilomar Conf. Signals, Systems, and Computers*, vol. 2, Nov. 1998, pp. 1547–1551.

[198] E. Villchur, “Signal processing to improve speech intelligibility in perceptive deafness,” *J. Acoust. Soc. Am.*, vol. 53, no. 6, pp. 1646–1657, Aug. 1973.

[199] N. Virag, “Single channel speech enhancement based on masking properties of the human auditory system,” *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 126–137, Mar. 1999.

[200] H. Wang and P. Chu, “Voice source localization for automatic camera pointing system in videoconferencing,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 1997, pp. 187–190.

[201] H. Wang and M. Kaveh, “Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 4, pp. 823–831, Aug. 1985.

[202] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. M.I.T. Press, 1949.

[203] J. D. Wise, J. R. Caprio, and T. W. Parks, “Maximum likelihood pitch estimation,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 5, pp. 418–423, Oct. 1976.

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

[205] 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.

[206] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, “Joint DOA and multi-pitch estimation based on subspace techniques,” *EURASIP J. on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–11, Jan. 2012.

[207] X. Zhang and Y. Jia, “A soft decision based noise cross power spectral density estimation for two-microphone speech enhancement systems,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Mar. 2005, pp. 813–816.

[208] J. Ziv and M. Zakai, “Some lower bounds on signal parameter estimation,” *IEEE Trans. Inf. Theory*, vol. 15, no. 3, pp. 386–391, May 1969.

Part II

Papers

Paper A

Fundamental Frequency Estimation using Polynomial Rooting of a Subspace-Based Method

Jesper Rindom Jensen, Mads Græsbøll Christensen and
Søren Holdt Jensen

The paper has been published in the
Proc. European Signal Processing Conference, 2010.

Abstract

We consider the problem of estimating the fundamental frequency of periodic signals such as audio and speech. A novel estimation method based on polynomial rooting of the harmonic MUltiple SIgnal Classification (HMUSIC) is presented. By applying polynomial rooting, we obtain two significant improvements compared to HMUSIC. First, by using the proposed method we can obtain an estimate of the fundamental frequency without doing a grid search like in HMUSIC. This is due to that the fundamental frequency is estimated as the argument of the root lying closest to the unit circle. Second, we obtain a higher spectral resolution compared to HMUSIC which is a property of polynomial rooting methods. Our simulation results show that the proposed method is applicable to real-life signals, and that we in most cases obtain a higher spectral resolution than HMUSIC.

1 Introduction

In many signal processing applications, it is of great importance to estimate the fundamental frequency. A specific example is in audio and speech processing. For example, the fundamental frequency is needed in parametric coding of audio and speech using a harmonic sinusoidal model. Also, many music information retrieval applications, such as automatic music transcription and musical genre classification, rely on the knowledge of the fundamental frequency. Within the last couple of decades, the problem of estimating the fundamental frequency has attracted considerable attention. This has resulted in numerous different fundamental frequency estimators. For a few examples of such estimators, we refer to [1–5].

Following, we define the fundamental frequency estimation problem. Consider a harmonic signal buried in white Gaussian noise $w(n)$, for $n = 0, \dots, N-1$,

$$x(n) = \sum_{l=1}^{L} \alpha_l e^{j\omega_0 ln} + w(n) , \tag{A.1}$$

where L is the model order and $\alpha_l = A_l e^{j\phi_l}$ is the complex amplitude of the lth sinusoid with $A_l > 0$ and ϕ_l being the real amplitude and the phase, respectively. In this paper, we will assume that the model order is known, hence, the problem at hand is to estimate the unknown fundamental frequency ω_0. While not considered in this paper, we refer the reader to [6] for few examples on how the model order could be estimated. In many existing methods for fundamental frequency estimation, the estimator is based on a grid search over a set of candidate fundamental frequencies [7]. This can be problematic for several reasons. For example, it can be hard to choose the resolution of the grid since the width of the peaks in the cost-function relies on, the sample size, the method, the signal-to-noise ratio (SNR), the source spacing (in multi-source scenarios), etc. Another issue is the compuational complexity. Naturally, the computational complexity depends on the resolution of the grid. That is, if the peaks are narrow or if high-resolution is required, it is necessary to use a fine grid which of course increases the computational complexity. The problem of choosing the right grid can, to some extend, be relieved by introducing a gradient search. To alleviate the abovementioned issues, we consider the problem of obtaining an estimate of the fundamental frequency without having to do a grid search.

It was shown in [8] that the MUltiple SIgnal Classification (MUSIC) estimation criterion [9, 10] can be used to obtain a high-resolution estimate of the fundamental frequency. The resulting estimator, refered to as Harmonic MUSIC (HMUSIC), was shown to have a good statistical performance. In this paper, we propose an estimator which is a relaxation of the HMUSIC cost-function from the unit circle onto the whole complex plane. That is, the proposed estimator evaluates the HMUSIC cost-function using a polynomial rooting method which can be seen as a generalization of the original root MUSIC method [11]. Using polynomial rooting has two signficant advantages. First, it gives an increased spectral resolution in multi-source scenarios and, second, it will give an estimate of the fundamental frequency without using a grid search. For more on the performance of the MUSIC and root MUSIC algorithms see, e.g., [12, 13]. Through simulations we investigate the performance of the proposed method on real-life signals. Also, using synthetic data, we evaluate the proposed estimator in Monte-Carlo simulations, and we compare the result with both the performance of the HMUSIC estimator and the Cramér-Rao Lower Bound (CRLB).

The rest of the paper is organized as follows. In Section 2, we make a brief introduction to the HMUSIC estimation method and we describe the proposed method. In Section 3, we evalute the performance of the proposed using both qualitative and quantitative measurements. Finally, Section 4 concludes on our work.

2 Proposed Methods

In this section, we present the fundamental theory behind the HMUSIC estimator [8] and we present the proposed estimator. Consider a signal of the form (A.1) from which we take M consecutive samples. The samples is then used to form a signal vector

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

where $(\cdot)^T$ denotes the transpose. If we then assume that the phases of the harmonics are independent and uniformly distributed in the interval $(-\pi;\pi]$, we can write the covariance matrix $\mathbf{R} \in \mathbb{C}^{M\times M}$ as [14]

$$\begin{aligned} \mathbf{R} &= \mathrm{E}\left\{\mathbf{x}(n)\mathbf{x}^H(n)\right\} && \text{(A.3)} \\ &= \mathbf{A}\mathbf{P}\mathbf{A}^H + \sigma_w^2\mathbf{I} \,, && \text{(A.4)} \end{aligned}$$

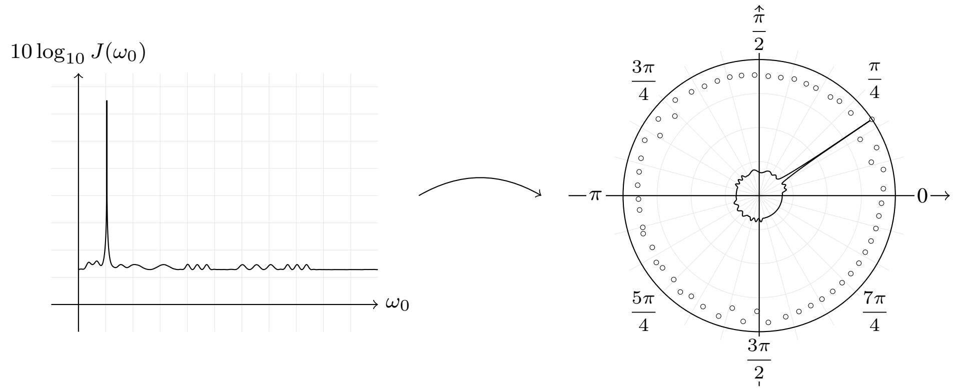


Fig. A.1: An example of a HMUSIC cost-function transformed into polar coordinates. The point $(0, 0)$ in the right-hand plot corresponds to $J(\omega_0) = 0$ while the whole unit circle corresponds to $J(\omega_0) = \infty$. Note that $\circ$ denotes a root of $p(z)$.

where $\mathrm{E}\{\cdot\}$ and $(\cdot)^H$ denotes the expectation and the conjugate transpose, respectively, σ_w^2 is the noise variance and $\mathbf{I}$ is the $M \times M$ identity matrix. The matrix $\mathbf{P}$ is a diagonal matrix containing the squared real amplitudes, i.e.,

$$\mathbf{P} = \text{diag}\left(\begin{bmatrix} A_1^2 & \cdots & A_L^2 \end{bmatrix}\right) , \tag{A.5}$$

and $\mathbf{A} \in \mathbb{C}^{M \times L}$ is a full-rank Vandermonde matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}(\omega_0) & \cdots & \mathbf{a}(L\omega_0) \end{bmatrix} , \tag{A.6}$$

with $\mathbf{a}(\omega) = \begin{bmatrix} 1 & e^{-j\omega} & \cdots & e^{-j\omega(M-1)} \end{bmatrix}^T$. Note that since we assume a harmonic model, the Vandermonde matrix $\mathbf{A}$ is only dependend on a single frequency, namely the fundamental frequency. Let us then define

$$\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H , \tag{A.7}$$

as the eigenvalue decomposition (EVD) of the covariance matrix. The matrix $\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_M \end{bmatrix}$ then contains the M orthonormal eigenvectors of $\mathbf{R}$ and $\mathbf{\Lambda}$ is a diagonal matrix containing the corresponding eigenvalues, λ_k. Note that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$. It is well known that the L most significant eigenvectors will span the signal subspace while the noise subspace is spanned by the $M - L$ least significant eigenvectors. That is, the noise subspace is spanned by $\mathbf{G}$ defined as

$$\mathbf{G} = \begin{bmatrix} \mathbf{u}_{L+1} & \cdots & \mathbf{u}_M \end{bmatrix} . \tag{A.8}$$

We know that range$(\mathbf{A})$ = range$(\mathbf{S})$ where $\mathbf{S} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_L \end{bmatrix}$ spans the signal subspace. Also, we know that the signal subspace is orthogonal to the noise subspace which allows us to write

$$\mathbf{A}^H \mathbf{G} = \mathbf{0} \, . \tag{A.9}$$

The covariance matrix, however, is most often not available in practice. Therefore, we will replace the covariance matrix in the above expression by the sample covariance matrix defined as

$$\hat{\mathbf{R}} = \frac{1}{N-M+1} \sum_{n=M-1}^{N-1} \mathbf{x}(n)\mathbf{x}^H(n) \, . \tag{A.10}$$

Due to estimation errors, $\mathbf{A}$ will not be exactly orthogonal to $\mathbf{G}$. Therefore, in HMUSIC, the fundamental frequency is found by

$$\hat{\omega}_0 = \arg \max_{\omega_0 \in \Omega_0} \frac{1}{\|\mathbf{A}^H \mathbf{G}\|_F^2} \tag{A.11}$$

$$= \arg \max_{\omega_0 \in \Omega_0} \underbrace{\frac{1}{\text{Tr}\left\{\mathbf{A}^H \mathbf{G}\mathbf{G}^H \mathbf{A}\right\}}}_{J(\omega_0)} \, , \tag{A.12}$$

with Tr$\{\cdot\}$ and $\|\cdot\|_F$ denoting the trace and the Frobenius norm, respectively, and Ω_0 is the set of candidate fundamental frequencies. Notice, that the HMUSIC criterion can be seen as an approximation to the angle between subspaces [15]. The minimization is done over the set Ω_0, i.e., the resolution of the estimate depends on the cardinality of Ω_0. The resolution can, however, be refined by performing a gradient search after a coarse estimate has been obtained.

Instead, we will now present how the cost-function can be evaluated using a rooting method. This has both the advantage of obtaining a solution without doing a grid search and an increased spectral resolution. Let us define a new matrix $\mathbf{C} = \mathbf{G}\mathbf{G}^H$ and rewrite the cost-function $J(\omega_0)$ by using the definition of the trace

$$J(\omega_0) = \frac{1}{\text{Tr}\{\mathbf{A}^H \mathbf{C} \mathbf{A}\}} \tag{A.13}$$

$$= \frac{1}{\sum_{l=1}^{L} \mathbf{a}^H(l\omega_0) \mathbf{C} \mathbf{a}(l\omega_0)} \, . \tag{A.14}$$

As mentioned previously, the expression in the denominator will have no solutions when equated with zero. However, if we instead replace $e^{j\omega}$ in $\mathbf{a}(\omega)$ with the variable $z = |z|e^{j \arg(z)}$, we can expect that denominator will have some solutions when equated with zero. That is, we can write

$$\frac{1}{J(z)} = \sum_{l=1}^{L} \mathbf{a}^T(z^{-l})\mathbf{C}\mathbf{a}(z^l) \tag{A.15}$$

$$= \sum_{l=1}^{L} \sum_{k=-(M-1)}^{M-1} c_k z^{lk} \tag{A.16}$$

$$= \sum_{l=1}^{L} p_l(z) = p(z) = 0, \tag{A.17}$$

where $p_l(z)$ is the lth polynomial and c_k is the sum of entries of $\mathbf{C}$ along the kth diagonal, i.e.,

$$c_k = \sum_{m-n=l} [\mathbf{C}]_{mn} \ . \tag{A.18}$$

The expression in (A.17) will only be zero when all of the individual polynomials $p_l(z)$ for $l = 1, \ldots, L$ is equal to zero. This can be proven by the fact that $\mathbf{C} = \mathbf{G}\mathbf{G}^H$ is Hermitian and thereby positive semi-definite. Positive semi-definiteness implies that

$$\mathbf{x}^H\mathbf{C}\mathbf{x} \geq 0 \ , \qquad \forall \mathbf{x} \ , \tag{A.19}$$

which proves our statement. Therefore, we can conclude that $p(z)$ has a root close on the unit circle only when $\hat{\omega}_0$ approaches ω_0. This will only be fulfilled when $M \to \infty$ which implies that $N \to \infty$

$$\lim_{N\to\infty} p(z) = 0\Big|_{z=e^{j\omega_0}} \Leftrightarrow \lim_{N\to\infty} J(z) = \infty\Big|_{z=e^{j\omega_0}} \ . \tag{A.20}$$

In reality, the roots of the polynomial will not lie exactly on the unit circle since we have a limited number of samples. Instead, if N and M are sufficiently large, we can assume that the root lying closest to the unit circle will correspond to the largest peak of the HMUSIC pseudospectra. This is also illustrated in Fig. A.1 which shows an example of a HMUSIC cost-function and its related roots. The fundamental frequency can therefore be estimated as the angle of the root $\hat{r}$ being closest to the unit circle, i.e.,

$$\hat{\omega}_0 = \angle\hat{r} \ . \tag{A.21}$$

Notice, however, that the roots come in complex conjugate pairs so we only consider the roots within the unit circle.

3 Experimental Results

This section contains the experimental results obtained during evaluation of the proposed method. First, we investigate the performance of the proposed method on a

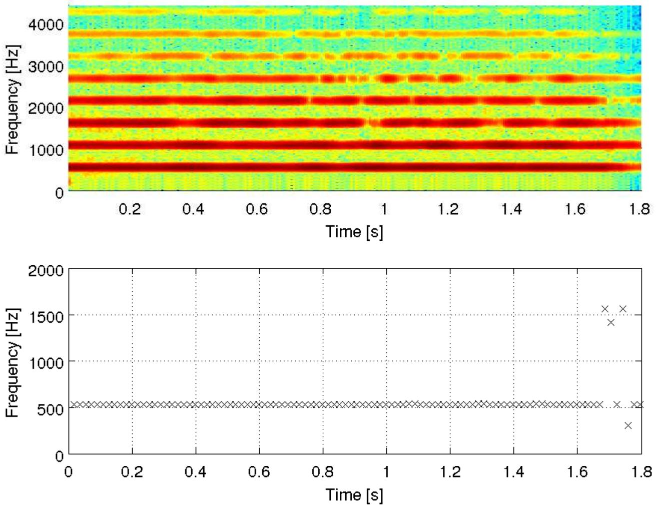


Fig. A.2: A spectrogram of a trumpet signal sampled at 8,820 Hz (top) and fundamental frequency estimates obtained using root HMUSIC (bottom).

real-life signal. The signal used in this experiment, was a trumpet signal sampled at 8,820 Hz. In Fig. A.2, the spectrogram of the trumpet signal is shown. We divided the trumpet signal into blocks of length $N = 160$ overlapping each other by 50 $\%$. The fundamental frequency was estimated from each block with $M = 65$ and by assuming that $L = 7$. The results are depicted in Fig. A.2. In the end of the signal, the proposed estimator seems to give erroneous estimates, however, it can also be seen that the model order in this part of the signal is rather five than seven. Except from the parts where there is a missmatch between the assumed model order and the true model order, the proposed estimator obtains estimates close to the true fundamental frequency. This verifies that the proposed estimator is applicable to real-life signals.

Also, we have conducted a series of Monte-Carlo simulations evaluating the statistical performance of the proposed method compared to both the original HMUSIC estimator and the CRLB [16]. In the first of these simulations, we evaluated the estimation performance with respect to the choice of M for N being fixed to 80. The signal used in this simulation, was a synthetic signal composed by $L = 3$ harmonically related complex sinusoids each with unit amplitudes with a fundamental frequency of 189.44

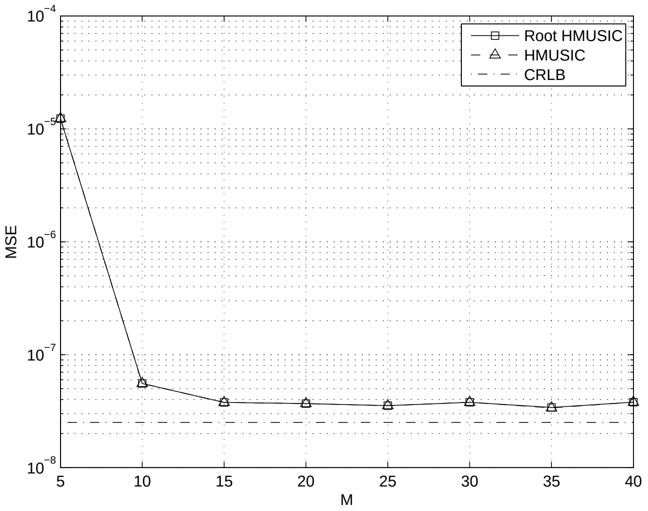


Fig. A.3: Plot of the asymptotic CRLB and the MSE of root HMUSIC and HMUSIC as a function of M.

Hz. Complex white Gaussian noise was added to the signal such that the SNR

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_{l=1}^{L} A_l^2}{\sigma_w^2} \ , \tag{A.22}$$

was 20 dB. Furthermore, the signal was sampled at $f_s = 2$ kHz. We then conducted 500 Monte-Carlo trials for each different M where we estimated the fundamental frequency. Also, for each different M we calculated the mean squared estimation error (MSE) defined as

$$\mathrm{MSE} = \frac{1}{S}\sum_{s=1}^{S}\left(\omega_0 - \hat{\omega}_0^{(s)}\right)^2 \ , \tag{A.23}$$

with ω_0 and $\hat{\omega}_0^{(s)}$ being the true fundamental frequency and its estimate in the sth Monte-Carlo trial, respectively, and S is the number of Monte-Carlo trials. The resulting MSEs for both root HMUSIC and MUSIC from this Monte-Carlo simulation are shown in Fig. A.3 together with the CRLB. We calculated the CRLB by using the asymptotic expression in [8]

$$\mathrm{CRLB}(\omega_0) = \frac{6\sigma_w^2}{N(N^2-1)\sum_{l=1}^{L} A_l^2 l^2} \ . \tag{A.24}$$

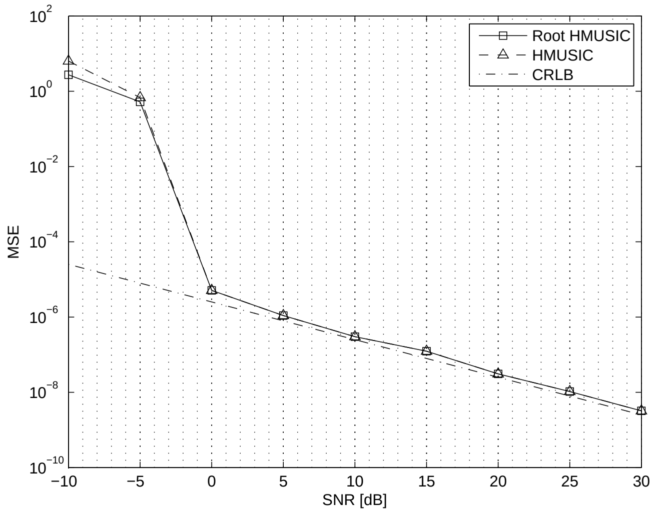


Fig. A.4: Plot of the asymptotic CRLB and the MSE of root HMUSIC and HMUSIC as a function of the SNR.

The first observation from the first Monte-Carlo simulation is, that both root HMUSIC and HMUSIC shows similar performance independently on the choice of M. Below $M = 10$ we see some thresholding behaviour for both methods. Also, we note that both methods are following but not reaching the CRLB as it was also reported in [8]. In another Monte-Carlo simulation, we evaluated the performance of root HMUSIC and HMUSIC with respect to the SNR. The parameters N, ω_0, L and f_s had the same values as in the previous Monte-Carlo simulation while M was fixed to $\lfloor \frac{N}{3} \rfloor$. We then ran 500 Monte-Carlo trials for each different SNR, and the results are depicted in Fig. A.4 in terms of the MSE. We note that for high SNRs, the two methods show the same performance while for low SNRs root HMUSIC seems to perform slightly better than HMUSIC. Thresholding behaviour is observed around an SNR of 0 dB for this particular setup.

Also, we evaluated the performance with respect to the fundamental frequency. In this Monte-Carlo simulation, $N = 60$ samples with a sampling frequency of $f_s = 2$ kHz of a synthetic signal having $L = 3$ sinusoids with unit amplitudes was used. Complex white Gaussian noise was added such that the SNR was 40 dB. Again, M was chosen to $\lfloor \frac{N}{3} \rfloor$. We ran 500 trials for each different fundamental frequency, and the results are depicted in Fig. A.5. Notice that for low fundamental frequencies, the proposed method shows a better performance compared to HMUSIC. This is also expected, since it has been reported that rooting methods have a better spectral resolution

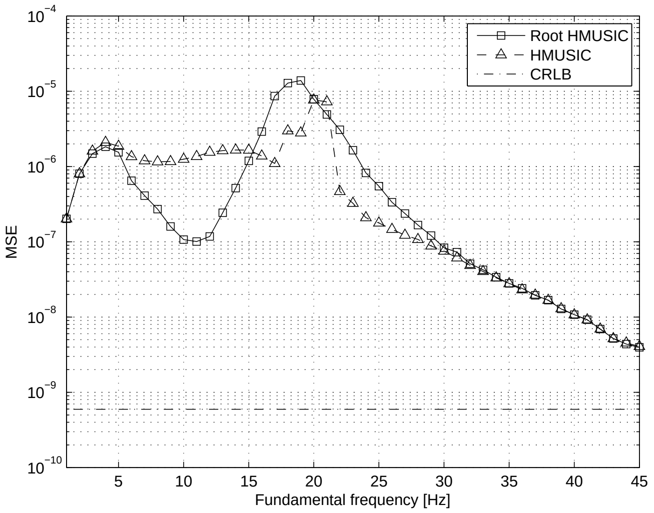


Fig. A.5: Plot of the asymptotic CRLB and the MSE of root HMUSIC and HMUSIC as a function of the fundamental frequency.

than spectral methods [11]. In the final Monte-Carlo simulation, we evaluated the performance of both root HMUSIC and HMUSIC in a two-source scenario. The sample length in this experiment was $N = 120$, M was 40 and the sampling frequency was $f_s = 2$ kHz. We generated the signal such that it was composed by two harmonic signals each with $L = 2$. The fundamental frequency of one of the harmonic signals was fixed to 114.79 Hz while the fundamental frequency of the other harmonic signal was varied. Furthermore, the SNR with respect to one harmonic signal was set to 40 dB. We ran 500 trials for each different fundamental frequency of the second harmonic source, and the outcome of this Monte-Carlo simulations is shown in Fig. A.6. Using this particular setup, it can be seen that at low frequency spacings (< 30 Hz), both methods show the same poor performance since they cannot resolve the sources. However, for higher frequency spacings (> 30 Hz), the proposed method shows a better performance compared to HMUSIC. In this simulation, the performance of both methods are relatively far away from the CRLB which is partly explained by the fact, that the CRLB is derived for a single source scenario with white noise.

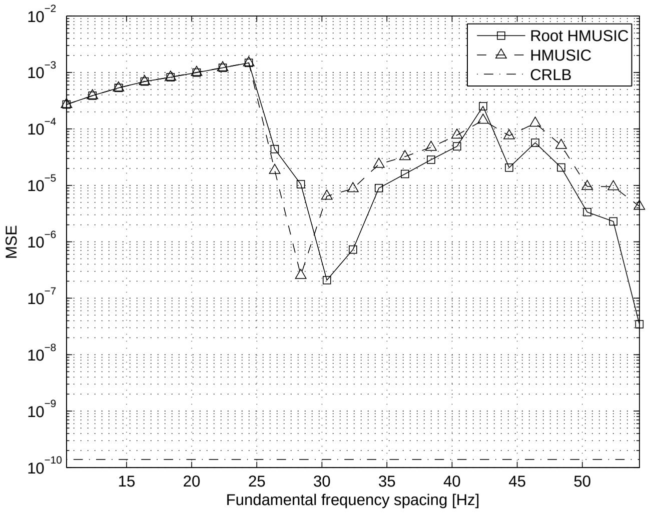


Fig. A.6: Plot of the asymptotic CRLB for the one source in white Gaussian noise scenario and the MSE of root HMUSIC and HMUSIC as a function of the fundamental frequency frequency spacing in a two-source scenario.

4 Conclusion

In this paper, we considered the fundamental frequency estimation problem. We proposed a new estimation method which is based on polynomial rooting of the known HMUSIC estimator. This has two significant advantages: 1) using the proposed method we obtain an estimate of the fundamental frequency without having to do a grid search and 2) using polynomial rooting we obtain a better spectral resolution compared to HMUSIC. We evaluated the proposed method using simulations. First, we showed that the proposed method is applicable to real-life signals, by using the root HMUSIC to correctly estimate the fundamental frequency. Second, we performed a series of statistical measurements on the proposed method. These simulations showed, that in many cases root HMUSIC will have a similar performance as HMUSIC. However, in multi-source scenarios with closely-spaced sources, the simulations showed that for most fundamental frequency spacings the proposed root HMUSIC method outperforms HMUSIC. This was also expected due to the properties of polynomial rooting methods. Like the HMUSIC method, the root HMUSIC method follows, but do not reach, the CRLB in good conditions.

References

[1] H. Li, P. Stoica, and J. Li, “Computationally efficient parameter estimation for harmonic sinusoidal signals,” *Elsevier Signal Process.*, vol. 80(9), pp. 1937–1944, 2000.

[2] K. W. Chan and H. C. So, “Accurate frequency estimation for real harmonic sinusoids,” *IEEE Signal Process. Lett.*, vol. 11, no. 7, pp. 609–612, 2004.

[3] 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.

[4] V. Emiya, B. David, and R. Badeau, “A parametric method for pitch estimation of piano tones,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 2007, pp. 249–252.

[5] S. Godsill and M. Davy, “Bayesian harmonic models for musical pitch estimation and analysis,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2002, pp. 1769–1772.

[6] 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.

[7] M. G. Christensen and A. Jakobsson, “Multi-pitch estimation,” *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[8] 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, no. 5, pp. 1635–1644, Jul. 2007.

[9] R. O. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[10] G. Bienvenu, “Influence of the spatial coherence of the background noise on high resolution passive methods,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, Apr. 1979, pp. 306–309.

[11] A. Barabell, “Improving the resolution performance of eigenstructure-based direction-finding algorithms,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 8, Apr 1983, pp. 336–339.

[12] B. D. Rao and K. V. S. Hari, “Performance analysis of root-MUSIC,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 1939 –1949, Dec 1989.

[13] 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.

[14] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Education, Inc., 2005.

[15] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Sinusoidal order estimation using angles between subspaces," *EURASIP J. on Advances in Signal Processing*, vol. vol. 2009, pp. 1–11, 2009.

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

Paper B

A Single Snapshot Optimal Filtering Method for Fundamental Frequency Estimation

Jesper Rindom Jensen, Mads Græsbøll Christensen and
Søren Holdt Jensen

The paper has been published in the
Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing, 2011.

Abstract

Recently, optimal linearly constrained minimum variance (LCMV) filtering methods have been applied for fundamental frequency estimation. Like many other fundamental frequency estimators, these methods utilize the inverse covariance matrix. Therefore, the covariance matrix needs to be invertible which is typically ensured by using the sample covariance matrix involving data partitioning. The partitioning adversely affects the spectral resolution. We propose a novel optimal filtering method which utilizes the LCMV principle in conjunction with the iterative adaptive approach (IAA). The IAA enables us to estimate the covariance matrix from a single snapshot, i.e., without data partitioning. The experimental results show, that the performance of the proposed method is comparable or better than that of other competing methods in terms of spectral resolution.

1 Introduction

There exists a multitude of signal processing applications in which the fundamental frequency is an essential parameter. A few examples are, e.g., parametric coding of audio and speech, automatic music transcription, musical genre classification, tuning of musical instruments, separation and enhancement of audio and speech sources, etc. Due to the importance of knowing the fundamental frequency, numerous of approaches and methods have been proposed for estimating this parameter. For a few examples of such estimators see, e.g., [1–7] and the references therein.

We will now introduce the problem of fundamental frequency estimation. The reasoning behind describing audio and speech signals by the fundamental frequency, among other parameters, is that audio and speech signals are quasi-periodic. That is, for a limited amount of signal samples, we can safely assume that for $n = 0, \dots, N-1$

$$x(n) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n} + w(n) \, , \tag{B.1}$$

where L is the number of harmonics, $\alpha_l = A_l e^{j\phi_l}$ with $A_l > 0$ and ϕ_l denoting the real amplitude and the phase of the lth harmonic, ω_0 is the fundamental frequency and $w(n)$ is complex noise. We assume that the model order L is known, hence, the fundamental frequency estimation problem is to estimate ω_0 from (B.1). While not considered in this paper, the model order assumption can easily be avoided by using a model order estimator [8, 9] or even by doing the model order and fundamental frequency estimation jointly [7].

Many of the aforementioned fundamental frequency estimators (e.g., optimal filtering techniques and subspace-based methods) utilizes the covariance matrix inverse [7], hence, in such estimators the covariance matrix must be invertible. In consequence of that, the covariance matrix must be full-rank. Typically, this is ensured by using the sample covariance matrix

$$\hat{\mathbf{R}} = \frac{1}{N-M+1} \sum_{n=M-1}^{N-1} \mathbf{x}(n)\mathbf{x}^H(n) \; , \tag{B.2}$$

where $\mathbf{x}(n) = \begin{bmatrix} x(n) & \cdots & x(n-M+1) \end{bmatrix}^T$ and $M < \frac{N}{2} + 1$. It is well-known that the spectral resolution depends on the sample length. That is, the resolution is decreased by the data partitioning embedded in (B.2).

Recently, however, the iterative adaptive approach (IAA) was proposed [10, 11], which can be used for covariance matrix and spectrum estimation. There is no data partitioning in this method, i.e., the covariance matrix is estimated iteratively from only a single snapshot. In this paper, we will propose to use a covariance matrix estimate, obtained by using the IAA, in conjunction with an optimal filtering method for fundamental frequency estimation. Note that the IAA could be used in conjunction with other covariance based fundamental frequency estimators as well. Since our method operates on a single snapshot of data, we can expect that our proposed optimal filtering method has a higher spectral resolution compared to the optimal filtering method in [7].

The remainder of the paper is organized as follows. In Section 2, we briefly review the optimal filtering method for fundamental frequency estimation and propose to use it in conjunction with the IAA. In Section 3, we present some experimental results obtained from quantitative experiments. Finally, in Section 4 we conclude on our work.

2 Optimal Filtering Method Utilizing the Iterative Adaptive Approach

2.1 Fundamental Frequency Estimation using Optimal Filtering

First, we will briefly review the concept of using an optimal filtering method for fundamental frequency estimation. This concept was introduced in [12] and is based on an optimal harmonic LCMV (hLCMV) filter. Consider M time-reversed samples from (B.1) in vector format

$$\mathbf{x}(n) = \begin{bmatrix} x(n) & x(n-1) & \cdots & x(n-M+1) \end{bmatrix}^T \; , \tag{B.3}$$

for $n = M-1, \ldots, N-1$. We introduce the FIR filter $\mathbf{h} = \begin{bmatrix} h(0) & \cdots & h(M-1) \end{bmatrix}^H$, from which the output is given by

$$y(n) = \mathbf{h}^H \mathbf{x}(n) \; . \tag{B.4}$$

The output power of the filter is the defined as

$$\mathrm{E}\{|y(n)|^2\} = \mathbf{h}^H \mathbf{R} \mathbf{h} \; , \tag{B.5}$$

where $\mathbf{R} = \mathrm{E}\{\mathbf{x}(n)\mathbf{x}^H(n)\}$. The optimal filter response is found, by using the LCMV principle. That is, we design the filter to have a unit gain at the harmonic frequencies while having maximum noise suppression

$$\min_{\mathbf{h}} \mathbf{h}^H\mathbf{R}\mathbf{h} \quad \text{s.t. } \mathbf{h}^H\mathbf{z}(l\omega_0) = 1 \ , \qquad \text{for } l = 1, \dots, L, \tag{B.6}$$

where $\mathbf{z}(\omega) = \begin{bmatrix} 1 & e^{-j\omega_0} & \cdots & e^{-j(M-1)\omega_0} \end{bmatrix}^T$. The well-know solution to this optimization problem is

$$\hat{\mathbf{h}} = \mathbf{R}^{-1}\mathbf{Z}(\omega_0)\left(\mathbf{Z}(\omega_0)^H\mathbf{R}^{-1}\mathbf{Z}(\omega_0)\right)^{-1}\mathbf{1} \ , \tag{B.7}$$

with $\mathbf{Z}(\omega_0) = \begin{bmatrix} \mathbf{z}(\omega_0) & \cdots & \mathbf{z}(L\omega_0) \end{bmatrix}$. We can then obtain an estimate of the fundamental frequency by inserting (B.7) into (B.5) and maximize the output power as

$$\hat{\omega}_0 = \arg\max_{\omega_0} \mathbf{1}^H\left(\mathbf{Z}^H(\omega_0)\mathbf{R}^{-1}\mathbf{Z}(\omega_0)\right)^{-1}\mathbf{1} \ . \tag{B.8}$$

The covariance matrix $\mathbf{R}$ is replaced by (B.2). Recall, that for $\mathbf{R}$ to be invertible, it is required that $M < \frac{N}{2} + 1$. In this paper, we propose instead to use a covariance matrix estimate obtained by using the iterative adaptive approach. In this method, the covariance matrix can be estimated from a single snapshot, i.e., we can obtain an $N \times N$ covariance matrix estimate.

2.2 Covariance Matrix Estimation using the Iterative Adaptive Approach

The iterative adaptive approach (IAA), proposed in [11], is a method for estimating the spectral amplitudes. In the estimation procedure, a WLS cost-function [13] is minimized

$$\hat{\alpha}_k = \arg\min_{\alpha_k} \left(\mathbf{x}(n) - \alpha_k\mathbf{z}(\omega_k)\right)^H \mathbf{Q}^{-1}(\omega_k)\left(\mathbf{x}(n) - \alpha_k\mathbf{z}(\omega_k)\right) \tag{B.9}$$

where $\mathbf{Q}(\omega_k)$ is the noise covariance matrix defined as

$$\mathbf{Q}(\omega_k) = \mathbf{R} - |\alpha_k|^2\mathbf{z}(\omega_k)\mathbf{z}^H(\omega_k) \ . \tag{B.10}$$

In the IAA, the covariance matrix is approximated by the well-known covariance matrix model [9]

$$\tilde{\mathbf{R}} = \bar{\mathbf{Z}}(\boldsymbol{\omega})\hat{\mathbf{P}}\bar{\mathbf{Z}}^H(\boldsymbol{\omega}) \ , \tag{B.11}$$

where $\boldsymbol{\omega} = \begin{bmatrix} 0 & 2\pi\frac{1}{K} & \cdots & 2\pi\frac{K-1}{K} \end{bmatrix}$ is the K-point frequency grid. The matrices $\bar{\mathbf{Z}}(\boldsymbol{\omega})$ and $\hat{\mathbf{P}}$ are defined as

$$\bar{\mathbf{Z}}(\boldsymbol{\omega}) = \begin{bmatrix} \mathbf{z}(\boldsymbol{\omega}(0)) & \cdots & \mathbf{z}(\boldsymbol{\omega}(K-1)) \end{bmatrix} \tag{B.12}$$

$$\hat{\mathbf{P}} = \operatorname{diag}\left\{ \begin{bmatrix} |\hat{\alpha}_0|^2 & \cdots & |\hat{\alpha}_{K-1}|^2 \end{bmatrix}^T \right\} , \tag{B.13}$$

where $|\hat{\alpha}_k|^2 = \hat{P}_k$. Minimizing (B.9) with respect to α_k yields

$$\hat{\alpha}_k = \frac{\mathbf{z}^H(\omega_k)\mathbf{Q}^{-1}(\omega_k)\mathbf{x}(n)}{\mathbf{z}^H(\omega_k)\mathbf{Q}^{-1}(\omega_k)\mathbf{z}(\omega_k)} . \tag{B.14}$$

By using the matrix inversion lemma it turns out that we can simplify (B.14) as

$$\hat{\alpha}_k = \frac{\mathbf{z}^H(\omega_k)\tilde{\mathbf{R}}^{-1}\mathbf{x}(n)}{\mathbf{z}^H(\omega_k)\tilde{\mathbf{R}}^{-1}\mathbf{z}(\omega_k)} . \tag{B.15}$$

Note, however, that to estimate the covariance matrix using (B.11), we need an estimate of the spectral amplitudes (B.15) and vice versa. The estimation is therefore performed iteratively initialized by the periodogram estimate. For most applications, 15 iterations is enough [11].

2.3 Proposed Optimal Filtering Method Utilizing the Iterative Adaptive Approach

In the proposed filtering method, we use the filter design in (B.7) where we replace the covariance matrix with the estimate in (B.11). This result in the optimal harmonic IAA (hIAA) filter

$$\tilde{\mathbf{h}} = \tilde{\mathbf{R}}^{-1}\mathbf{Z}(\omega_0)\left(\mathbf{Z}(\omega_0)^H\tilde{\mathbf{R}}^{-1}\mathbf{Z}(\omega_0)\right)^{-1}\mathbf{1} . \tag{B.16}$$

We could also use the noise covariance matrix $\mathbf{Q}$ instead of $\mathbf{R}$ in (B.16) which is intuitively more correct, i.e.,

$$\tilde{\tilde{\mathbf{h}}} = \tilde{\mathbf{Q}}^{-1}(\omega_0)\mathbf{Z}(\omega_0)\left(\mathbf{Z}(\omega_0)^H\tilde{\mathbf{Q}}^{-1}(\omega_0)\mathbf{Z}(\omega_0)\right)^{-1}\mathbf{1} . \tag{B.17}$$

We can write the IAA-based noise covariance matrix estimate as

$$\tilde{\mathbf{Q}}(\omega_0) = \tilde{\mathbf{R}} - \mathbf{Z}(\omega_0)\hat{\mathbf{P}}_s\mathbf{Z}^H(\omega_0) , \tag{B.18}$$

where $\hat{\mathbf{P}}_s$ is a diagonal matrix containing the estimated powers of the harmonics. By making use of the matrix inversion lemma, it can then be shown that

$$\tilde{\mathbf{h}} = \tilde{\tilde{\mathbf{h}}} . \tag{B.19}$$

```
$\hat{\alpha}_k = \dfrac{\mathbf{z}^H(\boldsymbol{\omega}(k))\mathbf{x}(n)}{N}$ ,        $k = 0, \ldots, K-1$
repeat
$\tilde{\mathbf{R}} = \bar{\mathbf{Z}}(\boldsymbol{\omega})\hat{\mathbf{P}}\bar{\mathbf{Z}}^H(\boldsymbol{\omega})$
  for $k = 0, \ldots, K-1$
    $\hat{\alpha}_k = \dfrac{\mathbf{z}^H(\boldsymbol{\omega}(k))\tilde{\mathbf{R}}^{-1}\mathbf{x}(n)}{\mathbf{z}^H(\boldsymbol{\omega}(k))\tilde{\mathbf{R}}^{-1}\mathbf{z}(\boldsymbol{\omega}(k))}$
    $\hat{P}_k = |\hat{\alpha}_k|^2$
  end
until (convergence)
$\hat{\omega}_0 = \arg\max_{\omega_0} \mathbf{1}^H \left(\mathbf{Z}^H(\omega_0)\tilde{\mathbf{R}}^{-1}\mathbf{Z}(\omega_0)\right)^{-1} \mathbf{1}$
```

Table B.1: The optimal filtering method for ω_0 estimation utilizing the IAA

Since the two filter designs are identical for the problem at hand, we will just use the design in (B.16) which is simpler. In Table B.1, it is shown how we can use the optimal hIAA filter to estimate the fundamental frequency.

As it can be seen, the estimate is obtained by maximizing the expected filter output power over a set of candidate frequencies. If a fine estimate is required, a relatively coarse set of candidate frequencies can be chosen whereupon the coarse fundamental frequency estimate is refined using a gradient search. The gradient, needed in that respect, is given by

$$g_{\omega_0} = -2\text{Re}\{\mathbf{1}^H(\mathbf{Z}^H\tilde{\mathbf{R}}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^H\tilde{\mathbf{R}}\mathbf{Y}(\mathbf{Z}^H\tilde{\mathbf{R}}^{-1}\mathbf{Z})^{-1}\mathbf{1}\} , \tag{B.20}$$

where $[\mathbf{Y}]_{pq} = \left[\frac{\partial}{\partial\omega}\mathbf{Z}\right]_{pq} = -j(p-1)qe^{-j\omega_0 q(p-1)}$.

3 Experimental Results

In this section, we describe the experimental evaluation of the proposed method. Note that in all simulations we estimate the fundamental frequency over a relatively coarse grid and refine the estimate using (B.20) in a steepest-descent algorithm with exact line search. First, we investigated how to choose the frequency grid size when estimating the covariance matrix using (B.11). To investigate this, we performed a series of Monte-Carlo simulations where we varied the frequency grid size. For each grid size we conducted 500 Monte-Carlo simulations. To evaluate the average error of doing the discretization in (B.11), we chose a random fundamental frequency in all simulations for a certain grid size. The random fundamental frequency was sampled from a uniform distribution $\mathcal{U}(0.4, 0.5)$. The model order was set to $L = 3$, the sample length

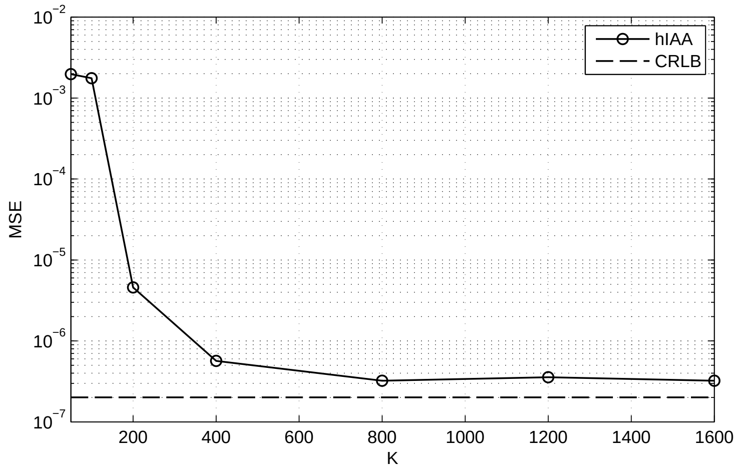


Fig. B.1: Fundamental frequency estimation MSE as a function of the grid size used in estimation of the covariance matrix with $N = 40$.

was $N = 40$ and the SNR, defined as

$$\text{SNR} = 10 \log_{10} \frac{\sum_{l=1}^{L} |\alpha_l|^2}{\sigma_w^2} , \tag{B.21}$$

was 20 dB (σ_w^2 is the noise variance). The results from this series of simulations are shown in Fig. B.1. From the results it can be seen that for this particular setup, a grid size of $K \approx 600$ frequency points is enough. Note also, that the MSE is following but not reaching the Cramér-Rao lower bound (CRLB). This is common, however, for the inverse covariance based methods [7]. The depicted CRLB is the asymptotic CRLB [14]

$$\text{CRLB}(\omega_0) \approx \frac{6\sigma_w^2}{N^3 \sum_{l=1}^{L} A_l^2 l^2} . \tag{B.22}$$

The same simulations were conducted when $N = 80$ and the results from these simulations are depicted in Fig. B.2. For the case with $N = 80$, $K \approx 1000$ is enough. The important thing to note is, that when we increase the number of samples N we also need to increase the number of frequency grid points K, to achieve the maximum possible performance.

We also compared the proposed method with the harmonic WLS (hWLS) [1], the harmonic LCMV (hLCMV) [7], the harmonic approximate NLS (hANLS) [7], and the harmonic MUSIC (hMUSIC) methods [7]. For example, we compared the methods for different sample lengths. For each sample length we conducted 500 Monte-Carlo

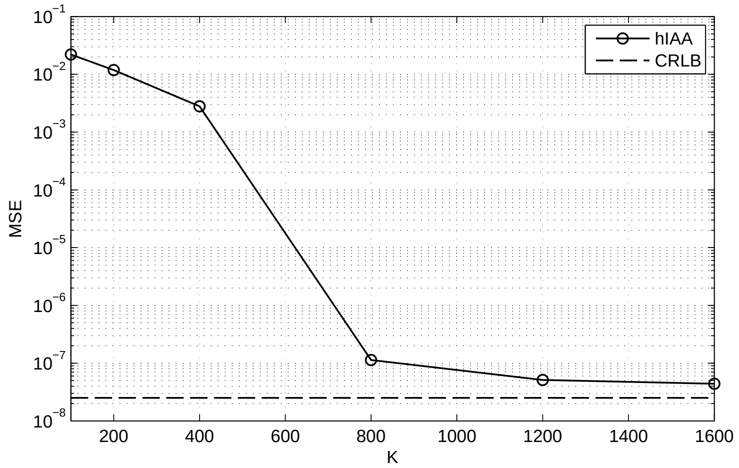


Fig. B.2: Fundamental frequency estimation MSE as a function of the grid size used in estimation of the covariance matrix with $N = 80$.

simulations and in each simulation ω_0 was sampled randomly from $\mathcal{U}(0.42, 0.43)$. The remaining setup was: $L = 3$, SNR = 20 dB and $K = 2000$. The results from this series of simulations are shown in Fig. B.3. First, we note that the hANLS method shows an erratic behaviour for these small sample lengths and is thereby outperformed by the other methods. The hIAA method outperforms the hLCMV method for all Ns, which is also expected since it has more degrees of freedom in the filter. Finally, we note that hMUSIC and hWLS performs best for $N < 25$ while for $N \geq 25$ hIAA, hWLS and hMUSIC show the same performance. Also we note, that for high Ns all methods seem to closely follow the CRLB. Then we compared the methods for different values of the fundamental frequency. A series of Monte-Carlo simulations were conducted with 500 simulations for each fundamental frequency. In each simulation K was sampled randomly from $\mathcal{U}_d(2000, 3000)$ ($\mathcal{U}_d(x_1, x_2)$ is the discrete uniform distribution taking integer values in the interval from x_1 to x_2). The remaining set up was: $N = 35$, $L = 3$ and SNR = 20 dB. The results from this experiment are depicted in Fig. B.4. Again we note that hANLS is unreliable for the given setup. The hIAA shows an improvement compared to hLCMV for $\omega_0 < 0.4$. For low fundamental frequencies ($\omega_0 < 0.3$), hIAA and hWLS outperforms the other methods, while for $\omega_0 > 0.4$ all methods except hANLS show the same performance. The results indicate that the proposed method (along with hWLS and hMUSIC) has a better spectral resolution than hLCMV. Finally, we compared hIAA, hLCMV, hANLS and hMUSIC in a scenario with two harmonic sources. The two sources both had $L = 3$ harmonics each with unit amplitudes. We then conducted a series of Monte-Carlo simulations for different spacings of the fundamental frequencies of the two sources (500 simulations for each

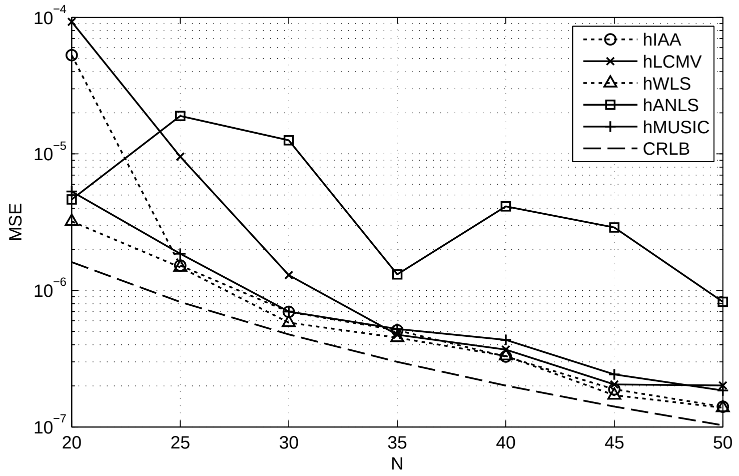


Fig. B.3: Fundamental frequency estimation MSE as a function of the sample length N.

frequency spacing). In each simulation, the number of samples was $N = 80$ and the SNR was 40 dB. The results from these simulations are shown in Fig. B.5. For $\Delta > 0.05$ the proposed method clearly outperforms the other methods.

4 Conclusion

In this paper, we proposed a new optimal filtering method for estimating the fundamental frequency of a (quasi-)periodic signal. The proposed method is an optimal LCMV filtering method which operates on single data snapshot. This is possible, because we estimate the covariance matrix using the iterative adaptive approach (IAA). By filtering on a single data snapshot rather than having to partition the data vector as in the filtering methods in [7], we obtain a better spectral resolution. The claim on increased spectral resolution was supported by the simulation results. The results showed that for small numbers of samples, low fundamental frequencies, and small frequency spacings in a two-source scenario, the proposed method clearly outperforms the optimal LCMV filtering method in [7]. This was also expected since the proposed method is an improvement of this method. Furthermore, for small number of samples and low frequencies, the proposed methods performance is comparable with that of the harmonic MUSIC and harmonic WLS methods. In a two-source scenario, it outperforms all the methods in the comparison above the frequency spacing threshold.

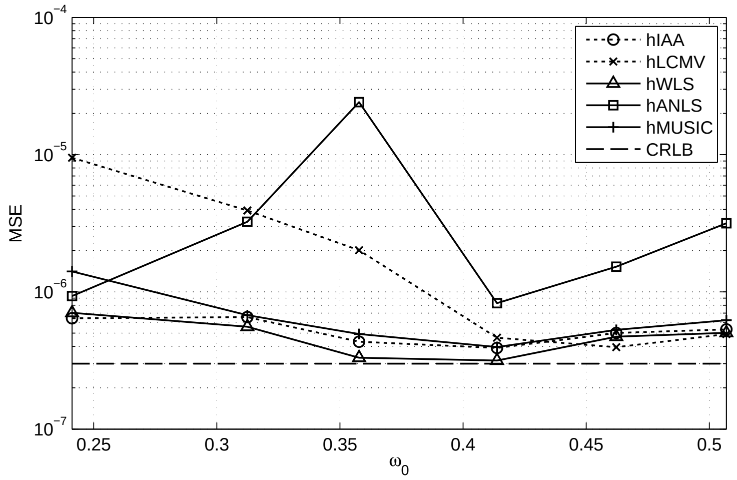


Fig. B.4: Fundamental frequency estimation MSE as a function of the fundamental frequency ω_0.

References

[1] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Elsevier Signal Process.*, vol. 80(9), pp. 1937–1944, 2000.

[2] K. W. Chan and H. C. So, "Accurate frequency estimation for real harmonic sinusoids," *IEEE Signal Process. Lett.*, vol. 11, no. 7, pp. 609–612, 2004.

[3] 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.

[4] V. Emiya, B. David, and R. Badeau, "A parametric method for pitch estimation of piano tones," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, Apr. 2007, pp. 249–252.

[5] S. Godsill and M. Davy, "Bayesian harmonic models for musical pitch estimation and analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, May 2002, pp. 1769–1772.

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

[7] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

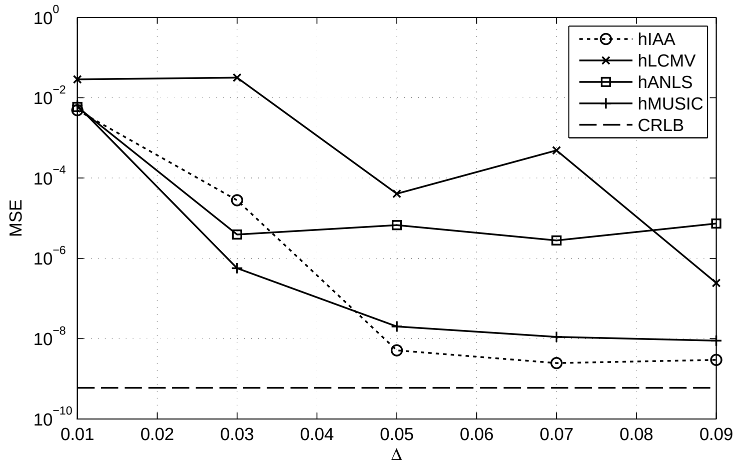


Fig. B.5: Fundamental frequency estimation MSE as a function of the fundamental frequency spacing Δ in a two-source scenario.

[8] 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.

[9] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Education, Inc., 2005.

[10] L. Du, T. Yardibi, J. Li, and P. Stoica, "Review of user parameter-free robust adaptive beamforming algorithms," *Digital Signal Processing*, vol. 19, no. 4, pp. 567–582, Jul. 2009.

[11] 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.

[12] 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.

[13] 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.

[14] 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, no. 5, pp. 1635–1644, Jul. 2007.

Paper C

Fast LCMV-based Methods for Fundamental Frequency Estimation

Jesper Rindom Jensen, George-Othon Glentis,
Mads Græsbøll Christensen, Andreas Jakobsson, and
Søren Holdt Jensen

The paper has been submitted to the
IEEE Trans. on Signal Processing, 2012.

In peer-review
The layout has been revised.

Abstract

Recently, optimal linearly constrained minimum variance (LCMV) filtering methods have been applied to fundamental frequency estimation. These estimators, however, suffer from a high computational complexity, since the cost functions are multimodal with narrow peaks, and they require matrix inversions and products for each point in the search grid. In this paper, we therefore propose a fast implementation of an LCMV-based fundamental frequency estimator, exploiting the estimator's inherently low displacement rank of the necessary products of Toeplitz-like matrices. This reduces the required computational complexity with several orders of magnitude. Moreover, we consider a recently proposed LCMV-based estimator in which the data covariance matrix is estimated using the iterative adaptive approach (IAA) for an increased spectral resolution. As the increased resolution comes at a notable computational cost, we also propose fast implementations of this estimator. One of these implementations is approximative and uses the preconditioned conjugates gradient method and a Quasi-Newton approach. Finally, we show how the considered pitch estimators can be efficiently updated when new observations become available. The time-recursive updating can reduce the computational complexity even further. The experimental results show that the performance of the proposed method is comparable or better than that of other competing methods in terms of spectral resolution. Furthermore, they show that the time-recursive implementations are able to track pitch fluctuations of synthetic as well as real-life signals.

1 Introduction

There exists a multitude of signal processing applications in which the fundamental frequency is an essential parameter, including, for instance, parametric coding of audio and speech, automatic music transcription, musical genre classification, tuning of musical instruments, separation and enhancement of audio and speech sources, and hearing aids. Due to the importance of knowing the fundamental frequency, numerous approaches and methods have been proposed for estimating this parameter, see, e.g., [1–7] and the references therein. An example of a recently proposed, high-resolution fundamental frequency estimator, is the linearly constrained minimum variance (LCMV) filtering method [6]. While this estimator is excellent for estimation of the fundamental frequency of closely spaced sources, it has a high computational complexity due to its cost function being multimodal with narrow peaks and requiring matrix inversion for each point in the search grid. The LCMV-based and many other fundamental frequency estimators rely on an estimate of the sample covariance matrix or its inverse, both commonly being formed by partitioning the available measurement into sub-vectors and forming the outer-product covariance matrix estimate. As is well-known, this approach will adversely affect the achievable spectral resolution, and there is therefore an interest in developing methods that achieve a higher spectral resolution. In this work, we consider the use of the iterative adaptive approach (IAA) for covariance matrix estimation in the LCMV method as proposed in [8]. The IAA was originally presented in [9] to provide sparse signal representation for passive sensing, channel estimation, and single-antenna radar applications, but having since found applications in areas as diverse as MIMO radar [10], missing data recovery [11], non-uniformly sampled spectral analysis [12], coherence and polyspectral estimation [13], spectroscopy [14], and blood velocity estimation using ultrasound [15]. The IAA estimate is a non-parametric data-dependent spectral estimate that does not require partitioning of the measurements. The estimate is instead formed iteratively and alternatingly by estimating the (amplitude) spectral estimate of the measurement as well as the covariance matrix formed from this amplitude spectrum. As shown in the above noted papers, the IAA-based estimation techniques are able to provide accurate estimates even when only a few samples are available. That is, by using the LCMV estimator in conjunction with the IAA for covariance matrix estimation, we achieve a substantially higher spectral resolution than what is normally achievable using the outer-product estimate. However, the improved performance comes at the cost of a considerable computational complexity. To alleviate this increase in complexity, we extend recent work on efficient IAA implementation [16, 17], exploiting the inherently low displacement rank of the necessary products of Toeplitz-like matrices, forming a computationally efficient implementation of the LCMV-based estimate. Moreover, we propose an even faster implementation that is approximative. This implementation is based on the preconditioned conjugates gradient method using a Quasi-Newton approach for the formulation of an appropriate preconditioning. We also show how the considered pitch estimators can be updated efficiently when new observations become available. By using such time-recursive implementations, the computationally complexity can be reduced much further as compared to batch processing, especially if time hopping is allowed. In the following section, we first briefly recall the LCMV-based fundamental frequency estimate and the IAA-based covariance matrix estimate. Then, in Sections 3 and IV, we introduce the proposed efficient exact and approximative implementations of the LCMV and the IAA-based LCMV methods. Section 5 discusses time-recursive updating of the estimates, followed, in Section 6, by extensive simulation examples illustrating the performance of the proposed implementations. Finally, Section VII concludes on the presented work.

2 Fundamental Frequency Estimation using Optimal Filtering

As audio and speech signals are quasi-periodic, one may well model such signals as (see, e.g., [7])

$$x(n) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n} + w(n), \tag{C.1}$$

for $n = 0, \ldots, N-1$, where L is the number of harmonics, $\alpha_l = A_l e^{j\phi_l}$, with $A_l > 0$ and ϕ_l denoting the real-valued amplitude and the phase of the lth harmonic, respectively. Furthermore, ω_0 denotes the sought fundamental frequency, and $w(n)$ is a complex-valued additive noise. For simplicity, we will here assume that the model order, L, is known, noting that this may otherwise be obtained using a model order estimator [18, 19], or by forming the model order and fundamental frequency estimation jointly, reminiscent to the ideas presented in [7]. The problem of interest is thus estimating ω_0 from $x(n)$ without making any strong assumptions on the statistics of the noise process.

2.1 Harmonic LCMV Method

Fundamental frequency estimation may, for instance, be conducted using the optimal filtering method introduced in [20], being based on an optimal LCMV filter. Consider an $(M-1)$th order FIR filter of which the output is given by

$$y(n) = \sum_{m=0}^{M-1} h(m)x(n-m) = \mathbf{h}^H \mathbf{x}_M(n), \tag{C.2}$$

for $n = M-1, \ldots, N-1$, where

$$\mathbf{h} = \begin{bmatrix} h(0) & \cdots & h(M-1) \end{bmatrix}^H \tag{C.3}$$

$$\mathbf{x}_M(n) = \begin{bmatrix} x(n) & \cdots & x(n-M+1) \end{bmatrix}^T, \tag{C.4}$$

with $(\cdot)^T$ and $(\cdot)^H$ denoting the transpose and conjugate transpose, respectively. The output power of the filter is

$$\mathrm{E}\{|y(n)|^2\} = \mathbf{h}^H \mathbf{R} \mathbf{h}, \tag{C.5}$$

where

$$\mathbf{R} = \mathrm{E}\{\mathbf{x}_M(n)\mathbf{x}_M^H(n)\}, \tag{C.6}$$

with $\mathrm{E}\{\cdot\}$ denoting the statistical expectation. The optimal filter response is found using the LCMV principle, such that the filter is designed to have a unit gain at the harmonic frequencies while having maximum noise suppression. This design procedure can also be formulated as

$$\min_{\mathbf{h}} \mathbf{h}^H \mathbf{R} \mathbf{h} \quad \text{subj. to } \mathbf{h}^H \mathbf{z}_M(l\omega_0) = 1 \ , \qquad \text{for } l = 1, \ldots, L, \tag{C.7}$$

where

$$\mathbf{z}_M(\omega_0) = \begin{bmatrix} 1 & e^{-j\omega_0} & \cdots & e^{-j(M-1)\omega_0} \end{bmatrix}^T . \tag{C.8}$$

The solution to the quadratic optimization problem with multiple equality constraints in (C.7) is well-known and given by [7]

$$\mathbf{h}_{\text{LCMV}} = \mathbf{R}^{-1} \mathbf{Z}_M(\omega_0) \left[\mathcal{G}_{\text{cov}}(\omega_0) \right]^{-1} \mathbf{1}, \tag{C.9}$$

with $\mathbf{1}$ denoting a vector of ones,

$$\mathcal{G}_{\text{cov}}(\omega_0) \triangleq \mathbf{Z}_M^H(\omega_0) \mathbf{R}^{-1} \mathbf{Z}_M(\omega_0), \tag{C.10}$$

and where

$$\mathbf{Z}_M(\omega_0) = \begin{bmatrix} \mathbf{z}_M(\omega_0) & \cdots & \mathbf{z}_M(L\omega_0) \end{bmatrix} . \tag{C.11}$$

An estimate of the fundamental frequency may thus be found by inserting (C.9) into (C.5) and maximize the output power, yielding

$$\hat{\omega}_0 = \arg \max_{\omega_0 \in \Omega_0} \mathbf{1}^T \left[\mathcal{G}_{\text{cov}}(\omega_0) \right]^{-1} \mathbf{1}, \tag{C.12}$$

where Ω_0 is a set of candidate fundamental frequencies. Here, we term the estimator in (C.12) the LCMV fundamental frequency estimate. The covariance matrix $\mathbf{R}$ is generally unknown, and is commonly replaced by the sample covariance matrix

$$\hat{\mathbf{R}} = \frac{1}{N - M + 1} \sum_{n=M-1}^{N-1} \mathbf{x}_M(n) \mathbf{x}_M^H(n). \tag{C.13}$$

To ensure that $\hat{\mathbf{R}}$ is invertible, the length of the sub-vectors, $\mathbf{x}_M(n)$, are restricted to $M < \frac{N}{2} + 1$, thereby limiting the spectral resolution to being on the order of $1/M$ [19]. A direct implementation of the estimator requires roughly

$$\mathcal{C}^{\text{cov}} \approx M^3 + M^2 \bar{N} + \bar{F} \left(ML^2 + LM^2 + L^3 \right) \tag{C.14}$$

operations, where $\bar{N} \triangleq N - M + 1$ and $\bar{F} \triangleq F/L$, with $F = |\Omega_0|$ being the size of the uniformly spaced grid of frequencies on the unit circle where the search for the optimum ω_0 is conducted. Typically, $F \gg N$, and due to the nature of the problem, the search is limited to candidate frequencies up to $\bar{F}$.

2.2 IAA-based Harmonic LCMV Method

We proceed to recall the IAA-based covariance matrix estimate, which is then used in conjunction with the LCMV method for fundamental frequency estimation. However, it should be stressed that this covariance matrix estimate could similarly be used in conjunction with other covariance based fundamental frequency estimators, thereby offering a similar improved spectral resolution. Following the usual IAA notation, let

$$\mathbf{x}_N = \begin{bmatrix} x(0) & x(1) & \cdots & x(N-1) \end{bmatrix}^T . \tag{C.15}$$

Then, the IAA estimate is formed by iteratively estimating the complex spectral amplitudes, $\alpha(\omega_k) \triangleq \alpha_k$, and the corresponding covariance matrix, $\tilde{\mathbf{R}}$, until practical convergence, as (see [9, 11] for further details)

$$\hat{\alpha}_k = \frac{\mathbf{z}_N^H(\omega_k)\tilde{\mathbf{R}}^{-1}\mathbf{x}_N}{\mathbf{z}_N^H(\omega_k)\tilde{\mathbf{R}}^{-1}\mathbf{z}_N(\omega_k)} \tag{C.16}$$

$$\tilde{\mathbf{R}} = \sum_{k=0}^{K-1} |\hat{\alpha}_k|^2 \mathbf{z}_N(\omega_k)\mathbf{z}_N^H(\omega_k) \tag{C.17}$$

for $k = 0, 1, \ldots, K-1$, with $\tilde{\mathbf{R}}$ being initialized to the identity matrix, $\mathbf{I}_N$. This implies that the complex amplitudes are initialized using the FFT of the sample vector. Typically, 10-15 iterations are sufficient for convergence in practice [9]. The expression in (C.16) can also be interpreted as a filtering operation, i.e.,

$$\hat{\alpha}_k = \mathbf{h}_{\text{IAA}}^H \mathbf{x}_N , \tag{C.18}$$

where the IAA filter, $\mathbf{h}_{\text{IAA}}$, is defined as

$$\mathbf{h}_{\text{IAA}} = \frac{\tilde{\mathbf{R}}^{-1}\mathbf{z}_N(\omega_k)}{\mathbf{z}_N^H(\omega_k)\tilde{\mathbf{R}}^{-1}\mathbf{z}_N(\omega_k)}, \tag{C.19}$$

from which it may be noted that the IAA filter resembles the optimal filter used in the traditional Capon method for spectrum estimation. Here, we instead propose a new filter, the IAA-based optimal LCMV (IAA-LCMV) filter, formed as

$$\mathbf{h}_{\text{IAA-LCMV}} = \tilde{\mathbf{R}}^{-1}\mathbf{Z}_N(\omega_0)\left[\mathcal{G}_{\text{IAA}}(\omega_0)\right]^{-1}\mathbf{1} \tag{C.20}$$

where

$$\mathcal{G}_{\text{IAA}}(\omega_0) \triangleq \mathbf{Z}_N^H(\omega_0)\tilde{\mathbf{R}}^{-1}\mathbf{Z}_N(\omega_0). \tag{C.21}$$

That is, we use the filter design in (C.9) together with the IAA covariance matrix estimate obtained after convergence has been achieved. Combining (C.18) and (C.20), one obtains an estimate of the output power of the IAA-LCMV filter as

$$\begin{aligned}\hat{P}_{\text{IAA-LCMV}} =& \mathbf{1}^T\left[\mathcal{G}_{\text{IAA}}(\omega_0)\right]^{-1}\mathbf{Z}_N^H(\omega_0)\mathbf{R}^{-1}\mathbf{X}_N \\ & \times \mathbf{R}^{-1}\mathbf{Z}_N(\omega_0)\left(\left[\mathcal{G}_{\text{IAA}}(\omega_0)\right]^{-1}\mathbf{1},\end{aligned} \tag{C.22}$$

with $\mathbf{X}_N = \mathbf{x}_N \mathbf{x}_N^H$. By taking the expected value of the output power, we get

$$\mathrm{E}\left\{\hat{P}_{\text{IAA-LCMV}}\right\} = \mathbf{1}^T \left(\boldsymbol{\mathcal{G}}_{IAA}(\omega_0)\right)^{-1} \mathbf{1}, \tag{C.23}$$

from which the expectation-based fundamental frequency estimate is obtained as

$$\hat{\omega}_0 = \arg \min_{\omega_0 \in \Omega_0} E\left\{\hat{P}_{\text{IAA-LCMV}}\right\}. \tag{C.24}$$

A direct implementation of (C.24) requires roughly

$$\mathcal{C}^{\text{IAA}} \approx m(N^3 + 3N^2K) + \bar{F}\left(NL^2 + LN^2 + L^3\right) \tag{C.25}$$

operations, where K denotes the size of the grid of frequencies utilized in the IAA implementation, with, usually, $K \leq F$, whereas m is the number of IAA iterations.

3 Efficient Implementation

An efficient implementation of (C.12) and (C.24) may alternatively be formed by exploiting the inherent displacement structure of the estimator, forming the implementation using Gohberg-Semencul (GS) factorizations of the involved inverse covariance matrices. Consider a Hermitian matrix $\mathbf{P} \in \mathcal{C}^{N \times N}$, and define the lower shifting matrix

$$\mathbf{D} = \begin{bmatrix} \mathbf{0}^T & 0 \\ \mathbf{I}_{N-1} & \mathbf{0} \end{bmatrix}. \tag{C.26}$$

Clearly, $(\mathbf{D})^N = \mathbf{0}$. Then, the displacement of $\mathbf{P}$ with respect to $\mathbf{D}$ and $\mathbf{D}^T$ is defined as

$$\nabla_{\mathbf{D},\mathbf{D}^T}\mathbf{P} \triangleq \mathbf{P} - \mathbf{D}\mathbf{P}\mathbf{D}^T. \tag{C.27}$$

Suppose that there exist integers ρ and $\sigma_i \in \{-1, 1\}$, for $i = 1, 2, \ldots, \rho$, such that (see also [21–23])

$$\nabla_{\mathbf{D},\mathbf{D}^T}\mathbf{P} = \sum_{i=1}^{\rho} \sigma_i \mathbf{t}_i \mathbf{t}_i^H = \mathbf{T}_\rho \boldsymbol{\Sigma}_\rho \mathbf{T}_\rho^H \tag{C.28}$$

where

$$\mathbf{T}_\rho = \begin{bmatrix} \mathbf{t}_1 & \cdots & \mathbf{t}_\rho \end{bmatrix} \tag{C.29}$$

$$\boldsymbol{\Sigma}_\rho = \text{diag}\left\{\begin{bmatrix} \sigma_1 & \cdots & \sigma_\rho \end{bmatrix}^T\right\}, \tag{C.30}$$

with diag$\{\mathbf{x}\}$ denoting the diagonal matrix formed with the vector $\mathbf{x}$ along its diagonal, and with $\mathbf{t}_i$ being the ith so-called generator vector. Then, the GS factorization of $\mathbf{P}$

may be expressed as

$$\mathbf{P} = \sum_{i=1}^{\rho} \sigma_i \mathcal{L}(\mathbf{D}, \mathbf{t}_i) \mathcal{L}^H(\mathbf{D}, \mathbf{t}_i) \ , \tag{C.31}$$

where $\mathcal{L}(\mathbf{D}, \mathbf{b})$ denotes a Krylov matrix of the form

$$\mathcal{L}(\mathbf{D}, \mathbf{b}) = \begin{bmatrix} \mathbf{b} & \mathbf{D}\mathbf{b} & \mathbf{D}^2\mathbf{b} & \cdots & \mathbf{D}^{M-1}\mathbf{b} \end{bmatrix} . \tag{C.32}$$

While this decomposition can be used to perform computationally demanding tasks such as matrix-vector multiplication in an efficient way, it does not provide an efficient way of computing the matrix itself when only its displacement representation is available. However, as

$$\mathbf{P} - \mathbf{D}\mathbf{P}\mathbf{D}^T = \sum_{i=1}^{\rho} \sigma_i \mathbf{t}_i \mathbf{t}_i^H , \tag{C.33}$$

multiplying both sides of (C.33) by $\mathbf{e}_{j+1}$ and noting that $\mathbf{D}^T\mathbf{e}_1 = \mathbf{0}$ and $\mathbf{D}^T\mathbf{e}_{j+1} = \mathbf{e}_j$, where $\mathbf{e}_j$ denotes a $N \times 1$ vector with a one at the jth entry and zeros elsewhere, implies that $\mathbf{P}$ may be recovered column-wise as

$$\mathbf{p}_j = \begin{cases} \displaystyle\sum_{i=1}^{\rho} \sigma_i \mathbf{t}_i \mathbf{t}_i^H \mathbf{e}_j \ , & j = 1 \\ \mathbf{D}\mathbf{p}_{j-1} + \displaystyle\sum_{i=1}^{\rho} \sigma_i \mathbf{t}_i \mathbf{t}_i^H \mathbf{e}_j \ , & j > 1 \end{cases} \tag{C.34}$$

for $j = 1, 2, \ldots, N-1$, with $\mathbf{p}_j$ denoting the jth column of $\mathbf{P}$. Estimating $\mathbf{P}$ in this way will require roughly ρN^2 operations. The coefficients of the trigonometric polynomial

$$\varphi(\omega) \triangleq \mathbf{z}^H(\omega)\mathbf{P}\mathbf{z}(\omega) = \sum_{\kappa=-N+1}^{N-1} c_\kappa e^{-j\kappa\omega} \tag{C.35}$$

can then be formed at a cost of $\mathcal{O}(\rho N \log_2 N)$ using the method detailed in [24]. However, to form the coefficients of the trigonometric polynomials

$$\psi_{l_1,l_2}(\omega) \triangleq \mathbf{z}^H(l_1\omega)\mathbf{P}\mathbf{z}(l_2\omega), \tag{C.36}$$

for l_1 and $l_2 \in \mathcal{Z}$, one needs to consider the augmented frequency vectors

$$\mathbf{z}_k(\omega) = \mathbf{S}_{l_k} \begin{bmatrix} \mathbf{z}(l_k\omega) \\ \times \end{bmatrix} , \tag{C.37}$$

for $k = 1$ or 2, where $\mathbf{S}_{l_k}$ is the selection matrix with zeros and ones indicating the presence or absence of a harmonic component, $\mathbf{S}_{l_k}\mathbf{S}_{l_k}^T = \mathbf{I}_{l_k N}$, and $\times$ denotes terms of no relevance. Using (C.37), (C.36) may be written as

$$\psi_{l_1,l_2}(\omega) = \mathbf{z}_1^H(\omega)\bar{\mathbf{P}}\mathbf{z}_2(\omega) = \sum_{\kappa=-l_1(N-1)}^{l_2(N-1)} \bar{c}_\kappa e^{-j\kappa\omega} \tag{C.38}$$

where $\bar{\mathbf{P}} \triangleq \mathbf{S}_{l_1}^T \mathbf{P}\mathbf{S}_{l_2}$ is an expanded rectangular matrix of size $(l_1(N-1)+1) \times (l_2(N-1)+1)$. Thus, the coefficients $\bar{c}_\kappa$ can be computed by summing all elements upon the diagonal of $\bar{\mathbf{P}}$. In practice, there is no need to form $\bar{\mathbf{P}}$, as one can easily show that these may be computed recursively as

$$\bar{\mathbf{C}}_{i+1} = \bar{\mathbf{C}}_i + \begin{bmatrix} \mathbf{0}_{l_2(N-1-i)} \\ \mathbf{S}_{l_1}^T \mathbf{p}_{i+1} \\ \mathbf{0}_{l_2 i} \end{bmatrix}, \tag{C.39}$$

for $i = 0, 1, \ldots, N-1$, where

$$\bar{\mathbf{C}} \triangleq \begin{bmatrix} \bar{c}_{-l_1(N-1)} & \cdots & \bar{c}_{l_2(N-1)} \end{bmatrix}^T, \tag{C.40}$$

at a cost of no more than $\mathcal{O}(N^2)$ operations. With (C.34) and (C.39) at our disposal, we proceed further with the proposal of fast implementation methods for (C.12) and (C.24).

3.1 Fast Harmonic LCMV Method

Restricting the set of candidate fundamental frequencies, Ω_0, to the frequencies uniformly spanned on the unit cycle, the maximization of (C.12) may be performed indirectly by exhaustive searching. This results in the evaluation of the trigonometric matrices in (C.10), and the computation of their inverses over the set of uniformly spaced frequencies of interest, which enables the use of the Fast Fourier Transform (FFT) for computational speed up. First, in order to form the required inversion of $\hat{\mathbf{R}}$ given by (C.13), we exploit the generalized Levinson algorithm presented in [24] to form $\hat{\mathbf{R}}^{-1}$ from its displacement representation. As discussed in [24], the computation of this task, requires about

$$\mathcal{C}^{\text{FCOV}}(M,N) \approx 4.5M^2 + 1.5N\log_2 N \tag{C.41}$$

operations, while $2M^2$ additional operations are needed for the computation of $\hat{\mathbf{R}}^{-1}$ from its displacement representation using (C.34). Subsequently, (C.10) is evaluated component-wise since the (l_1, l_2)th component of $\mathcal{G}(\omega_0)$, given by

$$[\mathcal{G}(\omega_0)]_{(l_1,l_2)} = \mathbf{z}_M^H(l_1\omega_0)\hat{\mathbf{R}}^{-1}\mathbf{z}_M(l_2\omega_0), \tag{C.42}$$

can be written as a polynomial of which the coefficients can be formed efficiently using (C.39). The cost of this is $\mathcal{O}(0.5L^2M^2)$ operations (non-trivial additions), although it should be noted that due to the Hermitian symmetry, only half of the polynomials actually have to be computed. Evaluating these on a uniformly spaced grid of frequencies using the FFT can be done at a cost of $\mathcal{O}(0.25L^2F\log_2 F)$, or $\mathcal{O}(0.25L^2F\log_2 F/L)$ if using FFT algorithms comprising output pruning. Finally, one may compute (C.12) at a cost of $\mathcal{O}(L^2F)$, implying that the overall computational cost of the proposed approach is approximately

$$\begin{aligned}\mathcal{C}^{\text{FLCMV}} \approx &\mathcal{C}^{\text{FCOV}}(M,N) + 2N^2 + \\ &F\left(0.25\log_2 F + 1\right)L^2\,. \end{aligned} \tag{C.43}$$

3.2 Fast IAA-based LCMV Method

Estimation of the fundamental frequency using the IAA-LMCV method is performed by maximizing (C.24) Restricting the search on an equally spaced set of frequencies on the unit circle, this task can be efficiently tackled by means of the FFT as in the LCMV approach discussed before. First, though, the covariance matrix $\tilde{\mathbf{R}}$ is estimated using IAA as described by (C.16) and (C.17), which can efficiently implemented without the need of direct estimation of the covariance matrix and its inverse. This can be accomplished using the celebrated Levinson-Durbin (LD) algorithm and some fast techniques for the evaluation of trigonometric polynomials related to structured matrices as detailed in [16, 17]. In this way, given the available data set $\mathbf{x}_N$, the displacement representation of the Toeplitz matrix (C.17), as well as the displacement representation of its inverse $\tilde{\mathbf{R}}^{-1}$, are iteratively estimated at a cost of

$$\begin{aligned}\mathcal{C}^{\text{FIAA}}(N,m) \approx m[&N^2 + 12N\log_2(2N) \\ &+ 1.5K\log_2 K]\end{aligned} \tag{C.44}$$

operations. With the displacement representation of $\tilde{\mathbf{R}}^{-1}$ at hand, the inverse itself is computed using (C.34) at an additional cost of N^2 operations. Subsequently, (C.21) is component wise evaluated by computing the coefficients of the associated trigonometric polynomials using (C.39) and the FFT, at a cost of $\left(0.5L^2N^2 + 0.25L^2F\log_2 F/L\right)$ operations. Finally, one may compute (C.24) at a cost of $\mathcal{O}(L^2F)$, implying that the overall computational cost of the proposed approach is approximately

$$\begin{aligned}\mathcal{C}^{\text{FIAA-LCMV}} \approx &\mathcal{C}^{\text{FIAA}}(N,m) + N^2 + \\ &F\left[0.25\log_2 F + 1\right]L^2\,. \end{aligned} \tag{C.45}$$

4 The Fast Approximative IAA-based LCMV Method

Substantial computational savings can be achieved by using the recently proposed approximative IAA algorithm [25] for the estimation of the covariance matrix and its inverse required in (C.24). In [25], a superfast implementation of the IAA algorithm, using the preconditioned conjugates gradient method and a Quasi Newton (QN) approach, was proposed for the formulation of an appropriate preconditioning. Building on these results, we here present a novel approximative fundamental frequency estimation method. The proposed approach is motivated by the QN algorithms formulated in [25, 26], where the inverse of Toeplitz-like matrices is approximated by extrapolating the inverse of a lower size matrix. The lower size matrix is treated as if it has been associated with an autoregressive (AR) model of lower order $q \leq N$. Thus, instead of computing $\tilde{\mathbf{R}}^{-1}$, a lower order extrapolated estimate is adopted. This results in an approximate IAA algorithm, where $\alpha(\omega_k)$ and $\mathbf{Q}$ are estimated iteratively as

$$\hat{\alpha}_k = \frac{\mathbf{z}_N^H(\omega_k)\mathbf{Q}^{-1}\mathbf{x}_N}{\mathbf{z}_N^H(\omega_k)\mathbf{Q}^{\mathbf{-1}}\mathbf{z}_N(\omega_k)} \tag{C.46}$$

$$\bar{\mathbf{R}} = \sum_{k=0}^{K-1} |\hat{\alpha}_k|^2 \, \mathbf{z}_q(\omega_k)\mathbf{z}_q^H(\omega_k) \tag{C.47}$$

for $k = 0, 1, \dots, K-1$, until practical convergence. The qth order autocorrelation matrix $\bar{\mathbf{R}}$ is initialized to the identity matrix, $\mathbf{I}_q$, and

$$\mathbf{Q}^{-1} = \begin{bmatrix} \mathbf{0} & \mathbf{0}^T \\ \mathbf{0} & \bar{\mathbf{R}}^{-1} \end{bmatrix} + \mathbf{A}_{N,N-q+1}\mathbf{A}_{N,N-q+1}^H \tag{C.48}$$

with

$$\mathbf{A}_{N,N-q+1} \triangleq \left[\bar{\mathbf{a}}\,\mathbf{D}\bar{\mathbf{a}}\,\dots\,\mathbf{D}^{N-q}\bar{\mathbf{a}}\right], \tag{C.49}$$

$$\bar{\mathbf{a}} = [\hat{\mathbf{a}}\,\mathbf{0}_{N-q}^T]^T, \tag{C.50}$$

where $\hat{\mathbf{a}}$ is the displacement generator associated with the power normalized forward predictor of $\bar{\mathbf{R}}$ as detailed in [25]. Then, the resulting approximative QN-IAA-based harmonic LCMV (QN-IAA-LCMV) method is formed by considering the cost function related to the estimate of $\mathbf{Q}$ as

$$\mathrm{E}\{\hat{P}_{\text{QN-IAA-LCMV}}\} = \mathbf{1}^T \left[\mathcal{G}_{\text{QN-IAA}}(\omega_0)\right]^{-1} \mathbf{1} \tag{C.51}$$

where

$$\mathcal{G}_{\text{QN-IAA}}(\omega_0) \triangleq \mathbf{Z}_N^H(\omega_0)\mathbf{Q}^{-1}\mathbf{Z}_N(\omega_0). \tag{C.52}$$

The fundamental frequency is then estimated by maximizing the output power using

$$\hat{\omega}_0 = \arg\min_{\omega_0 \in \Omega_0} E\left\{\hat{P}_{\text{QN-IAA-LCMV}}\right\} \tag{C.53}$$

Choosing $q \ll N$, a significant computation reduction can be achieved at the expense of a possible degradation in the quality of the spectrum estimate. The displacement representation of the approximate inverse covariance matrix $\mathbf{Q}^{-1}$ is estimated from the available data $\mathbf{x}_N$ using the QN-IAA algorithm detailed in [25], where the LD algorithm is employed for the computation of the displacement representation of $\bar{\mathbf{R}}^{-1}$. This representation is subsequently utilized in (C.48), at a cost of

$$\begin{aligned}\mathcal{C}^{\text{QN-IAA}}(m, N, q) \approx & m\left[q^2 + 12q\log_2(2q) + \right. \\ & \left. N\log_2(N) + 1.5K\log_2(K)\right]\end{aligned} \tag{C.54}$$

operations. Evaluating (C.52) component-wise, and using (C.48), one gets

$$\begin{aligned}\psi_{l_1,l_2}(\omega_0) &\triangleq \mathbf{z}_N^H(l_1\omega_0)\mathbf{Q}\mathbf{z}_N(l_2\omega_0) \\ &= \psi_{l_1,l_2}^{(1)}(\omega_0)e^{-j\omega_0(l_1-l_2)(N-q)} + \psi_{l_1,l_2}^{(2)}(\omega_0)\end{aligned}$$

where

$$\begin{aligned}\psi_{l_1,l_2}^{(1)}(\omega_0) &\triangleq \mathbf{z}_q^H(l_1\omega_0)\bar{\mathbf{R}}^{-1}\mathbf{z}_q(l_2\omega_0) \\ \psi_{l_1,l_2}^{(2)}(\omega_0) &\triangleq \mathbf{z}_N^H(l_1\omega_0)\mathbf{A}_{N,N-q}\mathbf{A}_{N,N-q}^H\mathbf{z}_N(l_2\omega_0).\end{aligned}$$

It should be noted that using (C.49), and recalling that $\mathbf{D}$ is a lower shifting matrix, $\psi_{l_1,l_2}^{(2)}(\omega_0)$ may be rewritten as

$$\psi_{l_1,l_2}^{(2)}(\omega_0) = N - q \tag{C.55}$$

when $l_1 = l_2$ and $\omega_0 = 0$, and

$$\psi_{l_1,l_2}^{(2)}(\omega_0) = \frac{\phi_{l_1}(\omega_0)\phi_{l_2}^*(\omega_0)(1 - e^{-j\omega_0(l_1-l_2)(N-q)})}{1 - e^{-j\omega_0(l_1-l_2)}}, \tag{C.56}$$

otherwise, where

$$\phi_l(\omega_0) = \mathbf{z}_q^H(l\omega_0)\bar{\mathbf{a}}. \tag{C.57}$$

Finally, since the search for the optimum fundamental frequency is restricted on a set of equally spaced frequencies on the unit circle, for frequencies ω_k up to $k = F/L$, (C.57) is efficiently evaluated as

$$\phi_l(\omega_k) = \phi_1(\omega_{lk}). \tag{C.58}$$

Thus, the overall computational complexity of the proposed QN-IAA-LCMV method is given by

$$\begin{aligned}\mathcal{C}^{\text{QN-IAA-LCMV}} \approx & \mathcal{C}^{\text{QN-IAA}}(m, N, q) + \\ & q^2 + F\left(0.25\log_2 F + 1\right)L^2 .\end{aligned} \tag{C.59}$$

5 Time-Recursive Implementations

We proceed to examine how the discussed methods may be efficiently updated as additional measurements becomes available.

5.1 Fast Harmonic LCMV

To allow for such an updating, the required covariance matrices should be replaced by suitable time-recursive estimates. To do so, an exponentially forgetting window approximation may be formed in place of (C.6) as

$$\begin{aligned}\hat{\mathbf{R}}(n) &= \sum_{m=0}^{n} \lambda^{n-m} \mathbf{x}_M(m)\mathbf{x}_M^H(m) && \text{(C.60)}\\ &= \lambda\hat{\mathbf{R}}(n-1) + \mathbf{x}_M(n)\mathbf{x}_M^H(n) && \text{(C.61)}\end{aligned}$$

where $\lambda \in (0,1)$ is the forgetting factor controlling the memory fading of the recursive estimator, with $\hat{\mathbf{R}}(-1)$ initialized by the scaled identity matrix $\sigma\mathbf{I}$, for $\sigma > 0$ (see also [27]). Exponentially forgetting updating is here selected in favor of a rectangular sliding window updating as the former allows for a computationally simpler algorithm as well as that the associated spectral variables are then updated in a more stable manner [28]. As shown in [28], $\hat{\mathbf{R}}^{-1}(n)$ obeys a particularly interesting identity of the form

$$\begin{aligned}\begin{bmatrix}\hat{\mathbf{R}}^{-1}(n) & \mathbf{0}\\ \mathbf{0}^T & 0\end{bmatrix} &= \begin{bmatrix}0 & \mathbf{0}^T\\ \mathbf{0} & \lambda\hat{\mathbf{R}}^{-1}(n)\end{bmatrix} + \mathbf{t}_1(n)\mathbf{t}_1^H(n) - \\ &\quad \mathbf{t}_2(n)\mathbf{t}_2^H(n) + \mathbf{t}_3(n)\mathbf{t}_3^H(n) && \text{(C.62)}\end{aligned}$$

where the vectors $\mathbf{t}_1(n)$, $\mathbf{t}_2(n)$, and $\mathbf{t}_3(n)$ are defined in terms of the power normalized forward and backward predictors as well as the Kalman gain vector (often denoted $\mathbf{a}(n)$, $\mathbf{b}(n)$, and $\mathbf{w}(n)$ in the adaptive signal processing nomenclature) associated with the sample covariance matrix $\hat{\mathbf{R}}(n)$ at time instant n. Using (C.62) in conjunction with (C.10) results in an efficient way for the component-wise estimation of the matrix $\boldsymbol{\mathcal{G}}^{\text{cov}}(n,\omega)$ at time instant n required in (C.12). The (l_1, l_2)th component of this matrix can also be written as

$$[\boldsymbol{\mathcal{G}}^{\text{cov}}(n,\omega_0)]_{(l_1,l_2)} = \psi_{l_1,l_2}(n,\omega_0) \qquad \text{(C.63)}$$

where $\psi_{l_1,l_2}(n,\omega_0) \triangleq \mathbf{z}_M^H(l_1\omega_0)\hat{\mathbf{R}}^{-1}(n)\mathbf{z}_M(l_2\omega_0)$, which, using (C.62), takes the form

$$\begin{aligned}\psi_{l_1,l_2}(n,\omega_0) =& \frac{1}{1-\lambda e^{-j(l_1-l_2)\omega_0}}\\ &\times \big[\varphi_{1,l_1}(n,\omega_0)\varphi_{1,l_2}^*(n,\omega_0)\\ &- \varphi_{2,l_1}(n,\omega_0)\varphi_{2,l_2}^*(n,\omega_0)\\ &+ \varphi_{3,l_1}(n,\omega_0)\varphi_{3,l_2}^*(n,\omega_0)\big] && \text{(C.64)}\end{aligned}$$

with

$$\varphi_{i,l}(n,\omega_0) \triangleq \mathbf{z}_M^H(l\omega_0)\mathbf{t}_i(n). \tag{C.65}$$

As the resulting search is restricted on a set of equally spaced frequencies on the unit circle for frequencies ω_k up to $k = F/L$, one may write

$$\varphi_{i,l}^n(\omega_k) = \varphi_{i,l}^n(\omega_{lk}). \tag{C.66}$$

The time-varying generator vectors of $\hat{\mathbf{R}}_M^{-1}(n)$, namely $\mathbf{t}_i(n)$, for $i = 1, 2, 3$, are computed using a standard Recursive Least Squares (RLS) algorithm at a cost of approximately $2M^2$ operations. Alternatively, a stabilized fast RLS (SFRLS) algorithm can be employed at a reduced cost of $7M$. Summarizing, the proposed time-recursive, fast, harmonic LCMV method consists of the following steps:

1. computation of the time varying generator vectors $\mathbf{t}_i(n)$, for $i = 1, 2, 3$, using the standard RLS or fast RLS (FRLS) algorithms, or using any other well behaved method,
2. element-wise computation of (C.64) using three FFTs as implied by (C.64) and (C.65), and
3. the search for the optimal fundamental frequency using (C.12).

It is worth noting that step 1) only needs to be updated at each time instant in a time-recursive way. The computations involved in the remaining steps 2) and 3), albeit their time-varying formalism, are not truly time-recursive in nature, since these depend on variables at the current time instant n only. This feature enables the development of time-recursive fundamental frequency algorithms with time hopping, in cases when the estimation of the fundamental frequency is not required at each time instant n, but only at every K_{hop} time units instead. The computational complexity per processed sample of the proposed time recursive harmonic LCMV method is therefore approximately

$$\mathcal{C}_{\text{TR}}^{\text{FLCMV}} \approx 2M^2 + \left(1.5F\log_2 F + L^2F\right)/K_{\text{hop}} \tag{C.67}$$

operations, when the RLS is used at step 1), or

$$\mathcal{C}_{\text{TR-F}}^{\text{FLCMV}} \approx 7M + \left(1.5F\log_2 F + L^2F\right)/K_{\text{hop}} \tag{C.68}$$

operations, when the FRLS is employed instead, with K_{hop} being an integer designating the possible time hopping, taking the value $K_{\text{hop}} = 1$ otherwise.

5.2 Fast IAA-based LCMV

Regrettably, the time frequency interleaving imposed by the iterative scheme in (C.16) and (C.17) of the IAA-based approach does not allow for a pure time-recursive estimation of the IAA-based covariance matrix and its subsequent use for a time-recursive computation of (C.21), required by the IAA-based cost function in (C.24). However, the development of a time-recursive scheme for the IAA-based fundamental frequency estimation is still feasible. As recently proposed in [29], the estimate of the covariance matrix at time instant n should be approximately equal to the estimate of the covariance matrix at time instant $n-1$ upon convergence. Thus, an approximative time-recursive update of the covariance matrix in (C.17) may be constructed as

$$\hat{\alpha}_k(n) = \frac{\mathbf{z}_N^H(\omega_k)\tilde{\mathbf{R}}^{-1}(n-1)\mathbf{x}_N(n)}{\mathbf{z}_N^H(\omega_k)\tilde{\mathbf{R}}^{-1}(n-1)\mathbf{z}_N(\omega_k)} \tag{C.69}$$

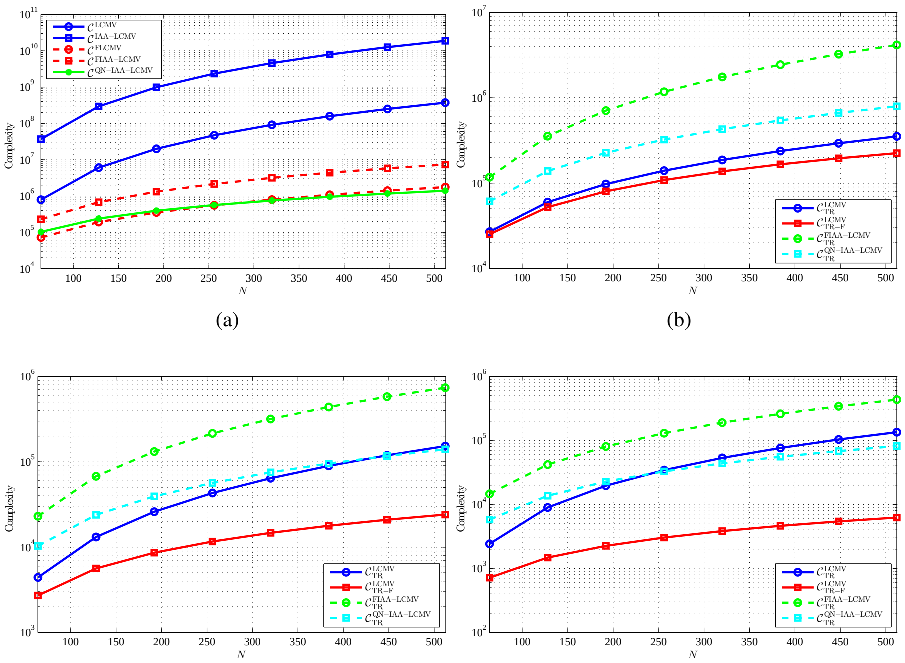
$$\tilde{\mathbf{R}}(n) = \sum_{k=0}^{K-1} |\hat{\alpha}_k(n)|^2 \, \mathbf{z}_N(\omega_k)\mathbf{z}_N^H(\omega_k) \tag{C.70}$$

where $\mathbf{x}_N(n) = [x(n-N+1)\ x(n-N+2)\ \ldots\ x(n)]^T$ is the data vector at time instant n. Although $\mathcal{G}^{\text{IAA}}(n,\omega_0)$, resulting from (C.21) and (C.70), is time-dependent, the required computations are not time-recursive in nature. This enables time hopping in the IAA-based fundamental frequency estimation case as well if desired. The computational complexity of the proposed time-recursive, IAA-based, harmonic LCMV scheme is therefore approximately given by

$$\begin{aligned}\mathcal{C}_{\text{TR}}^{\text{FIAA-LCMV}} \approx & \mathcal{C}^{\text{FIAA}}(N,1) + \\ & \left[N^2 + F\left(0.25\log_2 F + 1\right)L^2\right]/K\end{aligned} \tag{C.71}$$

The time-recursive, QN-, and IAA-based, harmonic LCMV method is organized in a similar way.

In Fig. C.1, we have depicted the computational complexities as a function of the number of samples, N, for the different implementations described in the previous sections. First, we considered the computational complexity for batch processing as shown in Fig. C.1a. From this figure, we can see that the fast implementations indeed have lower complexities than the direct implementations of the LCMV and IAA-LCMV methods. Furthermore, we observe that the implementations of the IAA-LCMV method generally have higher complexities than the corresponding implementations of the LCMV method. Finally, we note that the QN-based approximative implementation of the IAA-LCMV methods has computational complexity comparable to that of the fast implementation of the LCMV method. Then we considered the computational complexities for the proposed time-recursive fundamental frequency estimators in Figs. C.1b-C.1d. These complexities have the same trend, i.e., the implementations of the IAA-LCMV method have the highest computational complexity, but by using the QN-based approximative implementation, the complexity gets closer to that of the implementations of the LCMV method. Finally, we observe that the computational complexity of all implementations can be decreased with several orders of magnitude if time hopping is allowed.



(a)

(b)

(c)

(d)

Fig. C.1: Computational complexity of the harmonic LCMV fundamental frequency estimation algorithms using the data covariance approach, where $M = N/2 + 1$, the IAA approach, where $m = 10$ and $K = 4N$, and their fast implementation. In all cases, $F = 10N$ and $L = 5$. The complexities are shown for (a) batch processing, and for time-recursive processing with (b) $K_{\text{hop}} = 1$, (c) $K_{\text{hop}} = 10$, (d) $K_{\text{hop}} = 50$, respectively. Note the difference in scale.

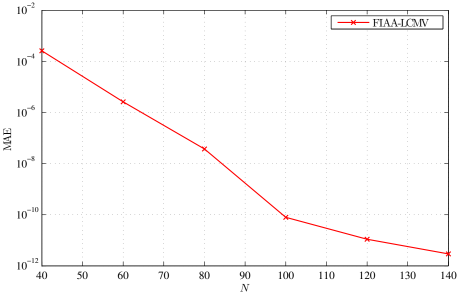
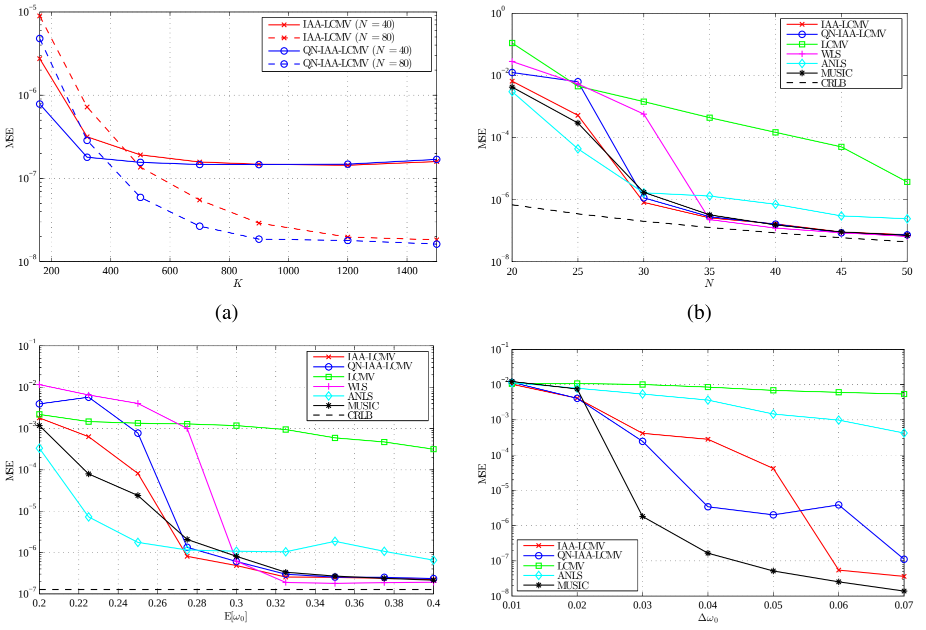


Fig. C.2: Mean absolute errors between the cost-functions obtained using the direct IAA-LCMV implementation and its fast implementation (FIAA-LCMV).



(a)

(b)

(c)

(d)

Fig. C.3: Mean squared errors of different fundamental frequency estimators as a function of (a) the number of frequency grid points used for the IAA-based covariance matrix estimate, (b) the number of available samples, (c) the expected fundamental frequency, and (d) the spacing between fundamental frequencies in a two source scenario. Moreover, the Cramér-Rao lower bound (CRLB) is depicted in (b) and (c).

6 Experimental Results

The experimental results obtained during the evaluation of the proposed methods are divided into two parts. First, we evaluate the statistical performance of the pitch estimators proposed for batch processing. Then, in the latter part, we evaluate the tracking performance of the proposed time-recursive pitch estimators.

6.1 Statistical Evaluation

We proceed to evaluate the accuracy of the efficient implementation of the proposed estimator. For this investigation, we used a harmonic signal with $L = 5$ in white

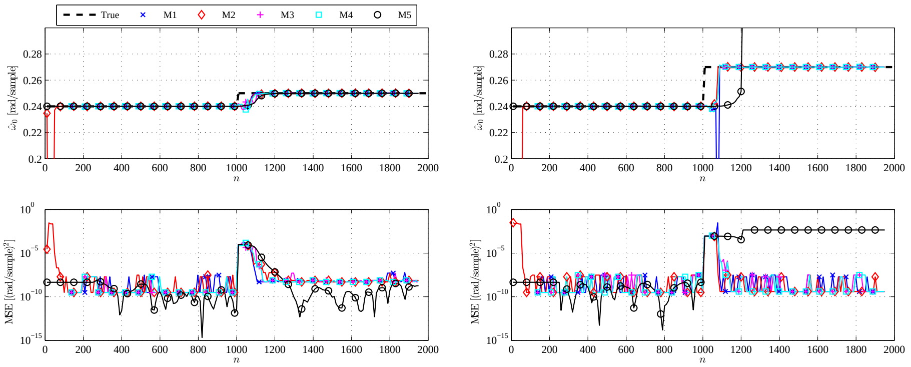
Gaussian noise at an SNR of 20 dB, with the SNR being defined as

$$\text{SNR} = 10 \log_{10} \left(\frac{1}{\sigma_w^2} \sum_{l=1}^{L} |\alpha_l|^2 \right). \tag{C.72}$$

The number of grid points used for the IAA-based covariance matrix estimate was $K = 1000$, the number of candidate fundamental frequencies was $|\Omega_0| = 2^{13}$, and the fundamental frequency was sampled from $\mathcal{U}(0.2, 0.3)$. Using this setup, we measured the mean absolute error (MAE) over all frequency points and Monte-Carlo simulations for different Ns. The results are provided in Fig. C.2. We note that the error between the direct and fast IAA-LCMV implementations are close to numerical precision for $N > 100$. The MAE is larger for lower Ns, since the matrix product $\mathbf{Z}_N^H(\omega_0)\tilde{\mathbf{R}}^{-1}\mathbf{Z}_N(\omega_0)$ becomes nearly ill-conditioned at low frequencies.

Then, we evaluate the performance of the proposed method, investigating the influence of K, N, the expected fundamental frequency, and the spacing between fundamental frequencies (the last in a two source scenario). For these experiments, the number of candidate fundamental frequencies was $|\Omega_0| = 2^{16}$. Initially, we consider a noisy harmonic signal as in the previous investigation. Fig. C.3a shows the measured mean squared error (MSE) of the IAA-LCMV and QN-IAA-LCMV estimators as a function of K, with the fundamental frequency being samples from $\mathcal{U}(0.3, 0.4)$. The results show the performance of the estimators for two different sample lengths, i.e., $N = 40$ and $N = 80$. As is clear from the figure, one needs more frequency points when N is increased to achieve the maximum possible performance for both the IAA-LCMV and the QN-IAA-LCMV methods. For $N = 40$, $K \approx 400$ seems to be sufficient, whereas at least $K \approx 1200$ frequency points are needed for $N = 80$. In this and all the following experiments, the order of the autocorrelation matrix was lowered to $q = \lfloor N/2 \rfloor$ in the QN-IAA-LCMV implementation.

Fig. C.3b shows the MSE as function of N, for $K = 1000$ frequency grid points, showing the performance of the IAA-based estimators as compared with a WLS method [1, 7], a LCMV method [7], an approximate NLS (ANLS) method [7], and a MUSIC method [7]. One may note from the figure that the IAA-based estimator show better performance as compared to the other methods for data lengths in the interval, say $30 < N < 35$. For larger data lengths, the IAA estimators outperform the ANLS and LCMV methods, while their performance is similar to that of the WLS and MUSIC method. Examining the influence of the fundamental frequency, Fig. C.3c shows the MSE as a function of the expected fundamental frequencies, $\text{E}[\omega_0]$, where, in each simulation, the fundamental frequency was sampled from $\text{E}[\omega_0] + \mathcal{U}(-0.001, 0.001)$, using $N = 35$ and $K = 1000$. As is clear from the results, the IAA estimators outperforms the LCMV and ANLS methods for $\text{E}[\omega_0] > 0.28$, while their performances are comparable to those of the MUSIC and WLS methods for $\text{E}[\omega_0] > 0.3$. Finally, we compared the discussed methods in a scenario with two harmonic sources, examining two sources with $L = 3$ unit amplitude harmonics. The ratio between each of the sources and a



(a) (b)

Fig. C.4: Plot of (top) the true pitch and pitch estimates obtained using M1-M5, and (bottom) the MSE associated with the pitch estimates.

white Gaussian noise source was 40 dB. In each simulation, the fundamental frequency ω_0^1 of first source was sampled from $\mathcal{U}(0.299, 0.301)$ and the fundamental frequency of the second source was $\omega_0^2 = \omega_0^1 + \Delta\omega_0$, where $\Delta\omega_0$ is the spacing, using $N = 60$, and $K = 1,000$. As seen in Fig. C.3d, the performances of the IAA methods are generally better than those of the LCMV and ANLS methods, while the MUSIC method shows the best performance. All the above presented results have been obtained using 500 Monte-Carlo simulations in which the phases of the harmonics and the noise signal were randomized. In summary and maybe most importantly, all the results obtained from the statistical evaluation show that the IAA can be used to improve the spectral resolution of the LCMV method that uses the sample covariance matrix estimate as we claimed in the introduction.

6.2 Qualitative Evaluation

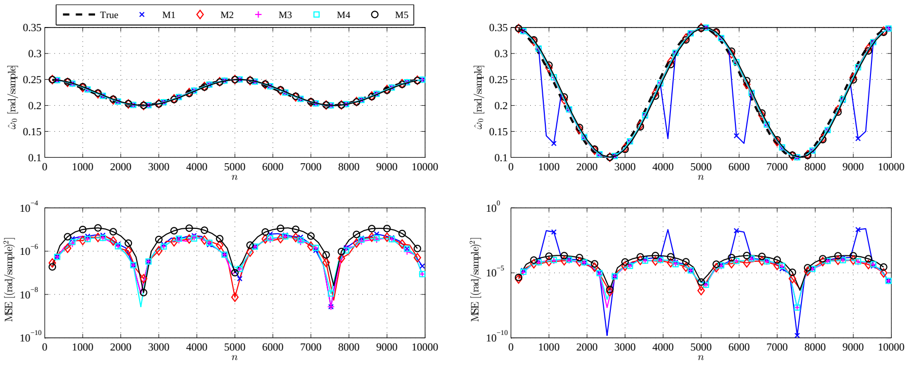
We proceed to evaluate the performance of the time-recursive implementations. These implementations are evaluated qualitatively on synthetic and real-life signals. In the evaluation, we consider

M1 the LCMV method implemented by applying (C.12) on rectangular sliding windowed data,

M2 the LCMV method implemented by (C.62)-(C.66),

M3 the IAA-LCMV method implemented by (C.69)-(C.70),

M4 the IAA-LCMV method implemented by (C.69)-(C.70) with an exponential forgetting factor on the IAA amplitude spectrum estimate,



(a)

(b)

Fig. C.5: Plots of (top) the true pitch and pitch estimates obtained using M1-M5, and (bottom) the MSE associated with the pitch estimates.

M5 the NLS pitch tracker proposed in [30] without Kalman filtering.

The NLS pitch tracker (M5) is included in the evaluation for benchmarking purposes. In all methods, observed data vectors of length $N = 128$ were considered. In M1-M2, a filter length of $M = 50$ were used, and the forgetting factor λ in the RLS algorithm of M2 was set to 0.98. The IAA-based methods (M3-M4) were set up with $K = 2000$ and 10 iterations for initialization. The forgetting factor in M4 was also set to $\lambda = 0.98$. In all methods, only pitch candidates in the interval $2\pi[0.004, 0.1]$ were considered, with $|\Omega_0| = 40000$ for M1-M4 and $|\Omega_0| = 2^{14}$ for M5. Using the above setup, we first applied the methods under evaluation on a synthetic signal constituted by a harmonic signal with $L = 5$ with an abrupt pitch change and white Gaussian noise; the SNR was 20 dB. A total of 2000 samples of the signal was observed. For the first 1000 samples, the true pitch was $\omega_0 = 0.24$, after which it changed to $\omega_0 = 0.24 + \delta$. In Fig. C.4a and C.4b, we show simulation results for $\delta = 0.01$ and $\delta = 0.03$, respectively. For $\delta = 0.01$, all methods show similar tracking performance. The NLS tracker (M5) obtains the lowest MSEs, which is explained by the fact that it obtains the pitch estimate using a gradient search rather than using a grid search as in the methods M1-M4. For a larger change in pitch ($\delta = 0.03$), the NLS tracker (M5) fails to track the pitch compared to the proposed methods. This can also be explained by the gradient search. We also evaluated the methods on a synthetic signal with smooth pitch changes. Again, the number of harmonics was $L = 5$, the noise was white Gaussian, and the SNR was 20 dB. Using frequency modulation, we obtained a harmonic signal with a pitch of

$$\omega_{0,\text{FM}}(n) = \omega_0 + \delta \cos\left(2\pi f_{\text{FM}} n / N_{\text{total}}\right) \tag{C.73}$$

at time instance n, where $\omega_0 = 0.225$, $f_{\text{FM}} = 2$ sample^{-1} is the modulation frequency, and $N_{\text{total}} = 10000$ is the total number observed samples. Simulation results for $\delta =$

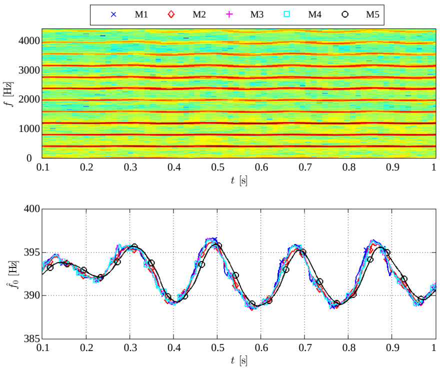


Fig. C.6: Plots of (top) the spectrogram of a real-life violin signal with vibrato, and (bottom) pitch estimates obtained using M1-M5 when applied on the violin signal.

0.025 and $\delta = 0.125$ are depicted in C.5a and C.5b, respectively. For $\delta = 0.025$, the performance in terms of MSE is similar for the methods M1-M4, whereas that MSE is generally larger for M5. The conclusions are the same for $\delta = 0.125$, except that the LCMV method with a sliding rectangular window (M1) has problems with pitch halving. Finally, we applied the methods on a real-life violin signal with vibrato. The spectrogram of the signal and the estimation results are shown in Fig. C.6. As it appears from these results, all methods were able to track the pitch fluctuations of this real-life signal.

7 Conclusions

In this paper, we consider fast implementations of two recently proposed pitch estimators. The estimators considered are both based on optimal filtering using the linearly constrained minimum variance (LCMV) principle, but uses different estimates estimates of the data covariance matrix; in one of the estimators, the sample covariance matrix estimate is used, whereas an iterative adaptive approach (IAA) estimate is used in the other. We propose fast implementations for both of the pitch estimators, exploiting the low displacement rank of the necessary products of Toeplitz-like matrices. As shown, this reduces the computational complexity by several orders of magnitude. We also propose an approximative fast implementation, using covariance matrices of lower size and extrapolation. This implementation has an even lower computational complexity. Finally, we propose time-recursive implementations of both estimators. These provide yet another mean for reducing the complexity. Our quantitative evaluation show that the IAA-based estimator considered outperforms other state-of-the-art pitch estimators in terms of means squared error in many scenarios, and that the difference between the proposed fast implementation and the direct implementation is negligible. Moreover, the qualitative evaluations show that the proposed time-recursive implementations can be used for tracking abrupt as well as smooth pitch changes of both synthetic and real-life signals.

References

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

[2] K. W. Chan and H. C. So, “Accurate frequency estimation for real harmonic sinusoids,” *IEEE Signal Processing Letters*, vol. 11, no. 7, pp. 609–612, Jul. 2004.

[3] 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.

[4] V. Emiya, B. David, and R. Badeau, “A parametric method for pitch estimation of piano tones,” in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 2007, pp. 249–252.

[5] S. Godsill and M. Davy, “Bayesian harmonic models for musical pitch estimation and analysis,” in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, May 2002, pp. 1769–1772.

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

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

[8] J. R. Jensen, M. G. Christensen, and S. H. Jensen, “A single snapshot optimal filtering method for fundamental frequency estimation,” in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2011, pp. 4272–4275.

[9] T. Yardibi, J. Li, P. Stoica, M. Xue, and A. B. Baggeroer, “Source localization and sensing: A nonparametric iterative approach based on weighted least squares,” *IEEE Trans. on Aerospace and Electronic Systems*, vol. 46, no. 1, pp. 425–443, Jan. 2010.

[10] W. Roberts, P. Stoica, J. Li, T. Yardibi, and F. A. Sadjadi, “Iterative adaptive approaches to MIMO radar imaging,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 1, pp. 5–20, Feb. 2010.

[11] P. Stoica, J. Li, and J. Ling, “Missing data recovery via a nonparametric iterative adaptive approach,” *IEEE Signal Processing Letters*, vol. 16, no. 4, pp. 241–244, Apr. 2009.

[12] P. Stoica, J. Li, and H. He, “Spectral analysis of nonuniformly sampled data: A new approach versus the periodogram,” *IEEE Trans. on Signal Processing*, vol. 57, no. 3, pp. 843–858, Mar. 2009.

[13] N. R. Butt and A. Jakobsson, “Coherence spectrum estimation from nonuniformly sampled sequences,” *IEEE Signal Processing Letters*, vol. 17, no. 4, pp. 339–342, Apr. 2010.

[14] E. Gudmundson, P. Stoica, J. Li, A. Jakobsson, M. D. Rowe, J. A. S. Smith, and J. Ling, “Spectral estimation of irregularly sampled exponentially decaying signals with applications to rf spectroscopy,” *Journal of Magnetic Resonance*, vol. 203, no. 1, pp. 167–176, Mar. 2010.

[15] E. Gudmundson, A. Jakobsson, J. A. Jensen, and P. Stoica, “Blood velocity estimation using ultrasound and spectral iterative adaptive approaches,” *Signal Processing*, vol. 91, no. 5, pp. 1275–1283, May 2011.

[16] G.-O. Glentis and A. Jakobsson, “Efficient implementation of iterative adaptive approach spectral estimation techniques,” *IEEE Trans. Signal Processing*, vol. 59, no. 9, pp. 4154–4167, Sep. 2011.

[17] M. Xue, L. Xu, and J. Li, “IAA spectral estimation: Fast implementation using the Gohberg-Semencul factorization,” *IEEE Trans. Signal Processing*, vol. 59, no. 7, pp. 3251–3261, Jul. 2011.

[18] P. Stoica and Y. Selén, “Model-order selection — a review of information criterion rules,” *IEEE Signal Processing Magazine*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

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

[20] M. G. Christensen, J. H. Jensen, A. Jakobsson, and S. H. Jensen, “On optimal filter designs for fundamental frequency estimation,” *IEEE Signal Processing Letters*, vol. 15, pp. 745–748, 2008.

[21] T. Kailath and A. H. Sayed, “Displacement structure: Theory and applications,” *SIAM Review*, vol. 37, no. 3, pp. 297–386, Sep. 1995.

[22] I. Gohberg and V. Olshevksy, “Complexity of multiplication with vectors for structured matrices,” *Linear Algebra Appl.*, vol. 202, pp. 163–192, Apr. 1994.

[23] D. Wood, “Product rules for the displacement of near-Toeplitz matrices,” *Linear Algebra Appl.*, vol. 188/189, pp. 641–663, Jul.-Aug. 1993.

[24] G.-O. Glentis, “A fast algorithm for APES and Capon spectral estimation,” *IEEE Trans. Signal Processing*, vol. 56, no. 9, pp. 4207–4220, Sep. 2008.

[25] G.-O. Glentis and A. Jakobsson, “Superfast approximative implementation of the IAA spectral estimate,” *IEEE Trans. Signal Processing*, vol. 60, no. 1, pp. 472–478, Jan. 2012.

[26] G. V. Moustakides and S. Theodoridis, “Fast Newton transversal filters - a new class of adaptive estimation algorithms,” *IEEE Trans. Signal Processing*, vol. 39, no. 10, pp. 2184–2193, Oct. 1991.

[27] G.-O. Glentis, K. Berberidis, and S. Theodoridis, “Efficient least squares adaptive algorithms for FIR transversal filtering,” *IEEE Signal Processing Magazine*, vol. 16, no. 4, pp. 13–41, Jul. 1999.

[28] G.-O. Glentis, “Efficient algorithms for adaptive Capon and APES spectral estimation,” *IEEE Trans. Signal Processing*, vol. 58, no. 1, pp. 84–96, Jan. 2010.

[29] G.-O. Glentis and A. Jakobsson, “Time-recursive IAA spectral estimation,” *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 111–114, Feb. 2011.

[30] M. G. Christensen, “A method for low-delay pitch tracking and smoothing,” in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 2012.

Paper D

Joint DOA and Fundamental Frequency Estimation Methods Based on 2-D Filtering

Jesper Rindom Jensen, Mads Græsbøll Christensen and
Søren Holdt Jensen

The paper has been published in the
Proc. European Signal Processing Conference, 2010.

Abstract

It is well-known that filtering methods can be used for processing of signals in both time and space. This comprises, for example, fundamental frequency estimation and direction-of-arrival (DOA) estimation. In this paper, we propose two novel 2-D filtering methods for joint estimation of the fundamental frequency and the DOA of spatio-temporarily sampled periodic signals. The first and simplest method is based on the 2-D periodogram, whereas the second method is a generalization of the 2-D Capon method. In the experimental part, both qualitative and quantitative measurements show that the proposed methods are well-suited for solving the joint estimation problem. Furthermore, it is shown that the methods are able to resolve signals separated sufficiently in only one dimension. In the case of closely spaced sources, however, the 2-D Capon-based method shows the best performance.

1 Introduction

In the last couple of decades, filtering methods have been used for processing of both spatial and temporal signals [1, 2]. Processing of spatial signals is also known as array signal processing and within this field, filtering methods are better known as beamformers [1]. One application of this, is processing of audio and speech signals recorded using a multi-microphone setup as considered in [3]. A common task is to estimate the direction-of-arrival (DOA) of such signals. However, often it is also desired to estimate the fundamental frequency of the self-same signals. Lately, it has been shown [4] that filtering methods are useful in this context as well. In the rest of the paper will refer to the fundamental frequency as the pitch. It is essential to estimate both the DOA and the pitch, since these features are useful for both separation and enhancement [5]. Furthermore, the pitch is also relevant regarding compression [6].

In many applications, such as hands-free communication, teleconferencing, surveillance systems and hearing-aids, it is necessary to know both the DOA and the pitch of speech and audio signals. This is needed for, e.g., tracking, separation and enhancement purposes. Therefore, joint pitch and DOA estimation is a relevant problem. We can formulate the joint estimation problem as follows: consider a periodic source $s(n_t)$ impinging on an array containing N_s sensors. On the n_sth sensor, the periodic source is corrupted by the noise source $w_{n_s}(n_t)$. The signal sampled by the n_sth sensor, for $n_t = 0, \dots, N_t - 1$ and $n_s = 0, \dots, N_s - 1$, can then be written as

$$x_{n_s}(n_t) = s(n_t - \tau_{n_s}) + w_{n_s}(n_t) \; , \tag{D.1}$$

where τ_{n_s} is the time delay of the signal on sensor n_s compared to a reference point. Note that the DOA can be estimated by realizing that there is a relationship between the DOA and the time delay. The relation depend on the array structure. In this paper, we assume a uniform linear array (ULA) and that the signals of interest are located in the so-called far-field. This implies a simple relationship between the DOA and the time delay. The problem considered in this paper, is join estimation of the DOA θ and the pitch ω_t of the periodic source $s(n_t)$ which can also be modeled as

$$s(n_t) = \sum_{l=1}^{L} \alpha_l e^{j\omega_t l n_t} \;, \tag{D.2}$$

where L is the model order and $\alpha_l = \Upsilon_l e^{j\phi_l}$ is the complex amplitude of the lth sinusoid with $\Upsilon_l > 0$ and ϕ_l being the amplitude and the phase, respectively. In this paper, we consider the model order as being known. The model in (D.2) allows us to consider the signal of interest as several narrowband sources. The narrowband assumption is a key assumption in many array processing methods and we also employ in this paper.

Recently, the problem of joint pitch and DOA estimation has attracted considerable attention. Some of the first approaches to solve the problem only considered how the frequencies of single 2-D sinusoids can be estimated. A few examples of such methods are [7], where a state-space realization technique is used, [8] which is based on the 2-D Capon method, [9] which is based on the ESPRIT method, and [10] where a signal-dependent multistage wiener filter (MWF) is used. The DOA and the pitch should, when possible, be estimated jointly for several reasons. For example, by estimating the parameters jointly we can process signals separated sufficiently in only one dimension as opposed to 1-D methods. Recently, a few methods have been proposed for joint DOA and pitch estimation; in [11] a ML-based method is proposed; in [11, 12] subspace-based methods are proposed; in [13] a correlation-based method is proposed; and in [14] a spatio-temporal filtering method based on the LCMV beamformer is proposed. In this paper, we present two novel methods for DOA and pitch estimation. Both methods are 2-D filtering methods based on a filter-bank interpretation of the periodogram and a generalization of the 2-D Capon method, respectively. As opposed to the method in [14], we do not require any prior knowledge on the spatial or temporal characteristics, other than a harmonic structure, since we propose to estimate the DOA and pitch jointly. In cases with colored noise, the 2-D Capon-based method is preferred because of its excellent performance in a multi-source scenario. However, the 2-D periodogram-based method may in some cases be preferred because of its lower computational complexity.

The rest of the paper is organized as follows: in Section 2 we introduce the joint estimation problem and present the proposed methods. Section 3 contains the experimental part of the paper and, finally, Section 4 concludes our work.

2 Proposed Methods

In this section, we briefly review the concept of 2-D filtering methods for spectral estimation and we present the proposed methods. In 2-D filtering methods for spectral estimation, it is desired to design a filter which passes a signal component with a given frequency pair undistorted. At the same time, the filter should attenuate signal components at all other frequency pairs. The two different filter design procedures used in the two proposed methods are described following.

Assume that we have a matrix $\mathbf{X}_D$ of dimension $N_s \times N_t$ containing our spatio-temporarily sampled data. Note that N_s and N_t are the numbers of spatial and temporal samples, respectively. The input data is then to be filtered by a $M_s \times M_t$ order 2-D finite impulse response (FIR) filter

$$\mathbf{H}_{\omega_t,\omega_s} = \begin{bmatrix} H_{\omega_t,\omega_s}(0,0) & \cdots & H_{\omega_t,\omega_s}(0,M_t') \\ \vdots & \ddots & \vdots \\ H_{\omega_t,\omega_s}(M_s',0) & \cdots & H_{\omega_t,\omega_s}(M_s',M_t') \end{bmatrix}, \tag{D.3}$$

where $M_s' = M_s - 1$, $M_t' = M_t - 1$ and the filter is designed for the temporal and spatial frequencies ω_t and ω_s. The filter is then applied on sub-blocks $\mathbf{X}_{n_s}(n_t)$ of the data matrix. One sub-block is defined as

$$\mathbf{X}_{n_s}(n_t) = \begin{bmatrix} x_{n_s}(n_t) & \cdots & x_{n_s}(n_t - M_t') \\ \vdots & \ddots & \vdots \\ x_{n_s+M_s'}(n_t) & \cdots & x_{n_s+M_s'}(n_t - M_t') \end{bmatrix}. \tag{D.4}$$

Due to the ULA and far-field assumptions, the spatial frequency is given by $\omega_s = \omega_t f_s \frac{d \sin\theta}{c}$, where f_s is the sampling frequency, d is the inter-sensor spacing, θ is the DOA in radians, and c is the wave propagation velocity. Following, we stack the filter response and the sub-blocks in (D.3) and (D.4), i.e.,

$$\mathbf{h}_{\omega_t,\omega_s} = \text{vec}\{\mathbf{H}_{\omega_t,\omega_s}\} \tag{D.5}$$

$$\mathbf{x}_{n_s}(n_t) = \text{vec}\{\mathbf{X}_{n_s}(n_t)\}, \tag{D.6}$$

with vec$\{\cdot\}$ denoting the column-wise stacking operator. Since we have mapped the filtering operation from 2-D to 1-D, it can be seen that the filter design somehow resembles that of 1-D filtering methods. As the first step in the design procedure we need to find an expression for the filter output power

$$\text{E}\{|y_{n_s}(n_t)|^2\} = \text{E}\{\mathbf{h}_{\omega_t,\omega_s}^H \mathbf{x}_{n_s}(n_t)\mathbf{x}_{n_s}^H(n_t)\mathbf{h}_{\omega_t,\omega_s}\} \tag{D.7}$$

$$= \mathbf{h}_{\omega_t,\omega_s}^H \mathbf{R}\mathbf{h}_{\omega_t,\omega_s}, \tag{D.8}$$

where $\mathbf{R}$ is the covariance matrix

$$\mathbf{R} = \text{E}\{\mathbf{x}_{n_s}(n_t)\mathbf{x}_{n_s}^H(n_t)\}. \tag{D.9}$$

Note that E$\{\cdot\}$ and $(\cdot)^H$ denotes the expectation operator and the complex transpose, respectively. Often we do not have access to the true covariance matrix, which we therefore will replace with the sample covariance matrix

$$\hat{\mathbf{R}} = \frac{1}{(N_s - M_s')(N_t - M_t')} \sum_{p=0}^{N_s-M_s} \sum_{q=0}^{N_t-M_t} \mathbf{x}_p(n_t - q)\mathbf{x}_p^H(n_t - q) \, . \tag{D.10}$$

The next task is to design the filter such that the output power is minimized subject to a distortionless constraint at desired frequencies. The filter design procedure is what differs between the two proposed method.

2.1 2-D Periodogram-based Method

In the 2-D periodogram-based method, we use a filter-bank structure constituted by stacked 2-D filter responses. We assume that there is no cross-talk between the filters in the filter-bank, which allows us to write the total filter-bank output power as

$$\sum_{l=1}^{L} \mathrm{E}\{|y_{n_s,l}(n_t)|^2\} = \sum_{l=1}^{L} \mathbf{h}_{l\omega_t,l\omega_s}^H \mathbf{R} \mathbf{h}_{l\omega_t,l\omega_s} \, . \tag{D.11}$$

Note that $\mathbf{h}_{l\omega_t,l\omega_s}$ is the impulse response of the lth filter in the filter-bank. We then define the filter-bank matrix as

$$\mathbf{H}_{\text{fb},(\omega_t,\omega_s)} = \begin{bmatrix}\mathbf{h}_{\omega_t,\omega_s} & \cdots & \mathbf{h}_{L\omega_t,L\omega_s}\end{bmatrix} \, . \tag{D.12}$$

Having introduced the filter-bank matrix, we can rewrite the total filter-bank output power in D.11 as

$$\sum_{l=1}^{L} \mathrm{E}\{|y_{n_s,l}(n_t)|^2\} = \mathrm{Tr}\left\{\mathbf{H}_{\text{fb},(\omega_t,\omega_s)}^H \mathbf{R} \mathbf{H}_{\text{fb},(\omega_t,\omega_s)}\right\} \, , \tag{D.13}$$

with $\mathrm{Tr}\{\cdot\}$ denoting the trace operator. In the 2-D periodogram-based method, it is assumed that the input signal is white Gaussian noise, hence, the method is independent on the signal statistics. Using this assumption, it can be shown that the individual filters in the filter-bank are constituted by Fourier vectors which will ensure a unit gain at the desired frequencies. However, the attenuation of other frequency components will not be optimal since the signal statistics are not used in the design procedure. The lth filter response is then given by

$$\mathbf{h}_{l\omega_t,l\omega_s} = \mathbf{a}_{l\omega_t,l\omega_s} \, , \tag{D.14}$$

where

$$\mathbf{a}_{\omega_t,\omega_s} = \mathbf{a}_{\omega_t} \otimes \mathbf{a}_{\omega_s} \tag{D.15}$$

$$\mathbf{a}_{\omega_k} = \begin{bmatrix}1 & e^{-j\omega_k} & \cdots & e^{-jM_k'\omega_k}\end{bmatrix}^T \, . \tag{D.16}$$

By inserting (D.14) into (D.13) we get that

$$\sum_{l=1}^{L} \mathrm{E}\{|y_{n_s,l}(n_t)|^2\} = \mathrm{Tr}\{\mathbf{A}_{\omega_t,\omega_s}^H \mathbf{R} \mathbf{A}_{\omega_t,\omega_s}\} \tag{D.17}$$

$$= J_{\mathrm{P},(\omega_t,\omega_s)} \ , \tag{D.18}$$

with $\mathbf{A}_{\omega_t,\omega_s} = \begin{bmatrix}\mathbf{a}_{\omega_t,\omega_s} & \cdots & \mathbf{a}_{L\omega_t,L\omega_s}\end{bmatrix}$. We can then obtain joint estimates of the pitch and the DOA by maximizing the total filter-bank output power over sets of candidate DOAs Θ and pitch frequencies Ω, i.e.,

$$(\hat{\theta}, \hat{\omega}) = \arg \max_{(\theta,\omega)\in\Theta\times\Omega} J_{\mathrm{P},(\omega_t,\omega_s)} \ . \tag{D.19}$$

2.2 2-D Capon-based Method

The proposed 2-D Capon-based method is a generalization of the 2-D Capon method [15]. The generalization is obtained by introducing multiple harmonic constraints in the filter design. That is, the proposed 2-D Capon-based method relies on a single 2-D filter. We design the 2-D filter by minimizing the output power subject to distortionless constraints on the harmonics, i.e.,

$$\min_{\mathbf{h}} \mathbf{h}_{\omega_t,\omega_s}^H \mathbf{R} \mathbf{h}_{\omega_t,\omega_s} \ \text{ s.t. } \ \mathbf{h}_{\omega_t,\omega_s}^H \mathbf{a}_{l\omega_t,\omega_s} = 1 \ , \tag{D.20}$$

$$\text{for } \ l = 1, \ldots, L \ . \tag{D.21}$$

The optimization problem is easily solved by using the Lagrange multiplier method which leads to the following result

$$\mathbf{h}_{\omega_t,\omega_s} = \mathbf{R}^{-1} \mathbf{A}_{\omega_t,\omega_s} (\mathbf{A}_{\omega_t,\omega_s}^H \mathbf{R}^{-1} \mathbf{A}_{\omega_t,\omega_s})^{-1} \mathbf{1} \ , \tag{D.22}$$

with $\mathbf{1}$ being a $L \times 1$ vector containing ones. Inserting the optimal filter expression into the filter output power expression leads to

$$\mathrm{E}\{|y_{n_s}(n_t)|^2\} = \mathbf{1}^H (\mathbf{A}_{\omega_t,\omega_s}^H \mathbf{R}^{-1} \mathbf{A}_{\omega_t,\omega_s})^{-1} \mathbf{1} \tag{D.23}$$

$$= J_{\mathrm{C},(\omega_t,\omega_s)} \ . \tag{D.24}$$

Finally, we can jointly estimate the DOA and the pitch by maximizing the filter output power for a sets of candidate DOAs Θ and pitch frequencies Ω

$$(\hat{\theta}, \hat{\omega}) = \arg \max_{(\theta,\omega)\in\Theta\times\Omega} J_{\mathrm{C},(\omega_t,\omega_s)} \ . \tag{D.25}$$

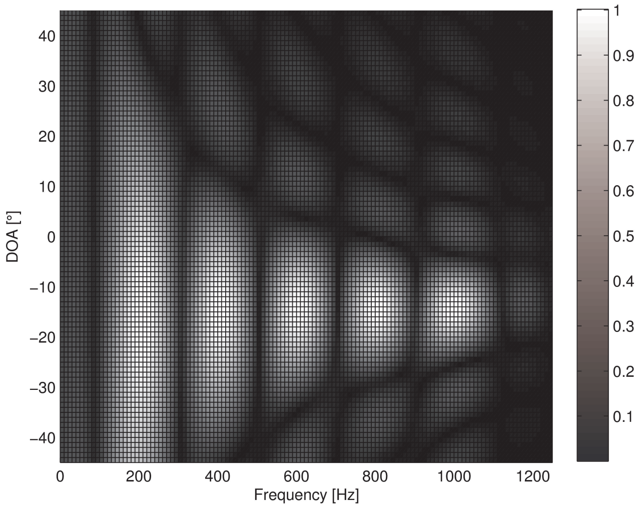


Fig. D.1: Contour plot of a 2-D periodogram-based filter-bank response designed for a signal constituted by five harmonics having a pitch of 200 Hz and a DOA of -15°.

2.3 Refined Estimates

In cases where there are high resolution requirements, the proposed methods may imply a huge computational burden since they rely on a grid search. However, to circumvent this issue we can instead use a smaller grid to obtain an initial estimate of the parameters which can then be refined by making a gradient search. Note that in the following derivations we leave out the dependencies on ω_s and ω_t for a simpler notation. The first order derivatives of the filter output powers are readily obtained as

$$\begin{bmatrix} \frac{\partial J_{\mathrm{P}}}{\partial \theta} \\ \frac{\partial J_{\mathrm{P}}}{\partial \omega_t} \end{bmatrix} = \frac{2}{M^2} \mathrm{Re} \left\{ \begin{bmatrix} \mathrm{Tr}\{\mathbf{A}^H \mathbf{R} \mathbf{B}_\theta\} \\ \mathrm{Tr}\{\mathbf{A}^H \mathbf{R} \mathbf{B}_{\omega_t}\} \end{bmatrix} \right\} \tag{D.26}$$

$$\begin{bmatrix} \frac{\partial J_{\mathrm{C}}}{\partial \theta} \\ \frac{\partial J_{\mathrm{C}}}{\partial \omega_t} \end{bmatrix} = -2\mathrm{Re} \left\{ \begin{bmatrix} \mathbf{1}^H \mathbf{Q} \mathbf{A}^H \mathbf{R}^{-1} \mathbf{B}_\theta \mathbf{Q} \mathbf{1} \\ \mathbf{1}^H \mathbf{Q} \mathbf{A}^H \mathbf{R}^{-1} \mathbf{B}_{\omega_t} \mathbf{Q} \mathbf{1} \end{bmatrix} \right\} , \tag{D.27}$$

where $\mathbf{Q} = (\mathbf{A}^H \mathbf{R}^{-1} \mathbf{A})^{-1}$, and Re$\{\cdot\}$ denotes taking the real part. The entries in the $\mathbf{B}$ matrices are given by

$$[\mathbf{B}_\theta]il = -j\omega_t f_s l \frac{d\cos\theta}{c} k_{s,i} e^{-j\omega_t l \left(f_s \frac{d\sin\theta}{c} k_{s,i} + k_{t,i}\right)} \tag{D.28}$$

$$[\mathbf{B}_{\omega_t}]_{il} = -jl \left(f_s \frac{d\sin\theta}{c} k_{s,i} + k_{t,i} \right) e^{-j\omega_t l \left(f_s \frac{d\sin\theta}{c} k_{s,i} + k_{t,i}\right)} . \tag{D.29}$$

The intermediate k variables are defined as

$$k_{s,i} = (i-1) \bmod M_s \tag{D.30}$$

$$k_{t,i} = \left\lfloor \frac{i-1}{M_s} \right\rfloor , \tag{D.31}$$

with $(\cdot \bmod \cdot)$ and $\lfloor \cdot \rfloor$ denoting the modulus and the flooring operators, respectively. We can then obtain refined parameter estimates by using an iterative procedure

$$\begin{bmatrix} \hat{\theta}^{(i+1)} \\ \hat{\omega}_t^{(i+1)} \end{bmatrix} = \begin{bmatrix} \hat{\theta}^{(i)} \\ \hat{\omega}_t^{(i)} \end{bmatrix} + \delta \nabla J , \tag{D.32}$$

where i is the iteration number, δ is a small positive constant found through line search and $\nabla J = \begin{bmatrix} \frac{\partial J}{\partial \theta} & \frac{\partial J}{\partial \omega_t} \end{bmatrix}^T$.

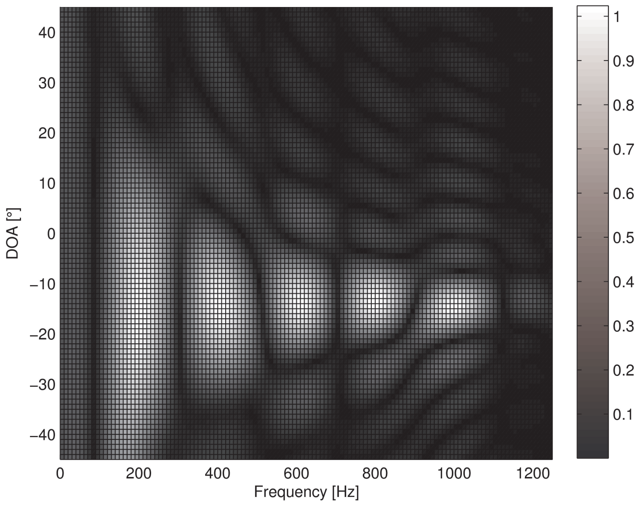


Fig. D.2: Contour plot of a 2-D Capon-based filter response designed for a signal constituted by five harmonics having a pitch of 200 Hz and a DOA of -15°. The signal was corrupted by white Gaussian noise with an SNR of -40 dB.

3 Experimental Results

We will now consider the evaluation of the proposed methods. In all of the experiments described in the rest of this section, an ULA was assumed having an inter-sensor spacing of $d = \frac{c}{f_s}$ where $c = 343.2$ m/s is the speed of sound in air at 20° C. Furthermore,

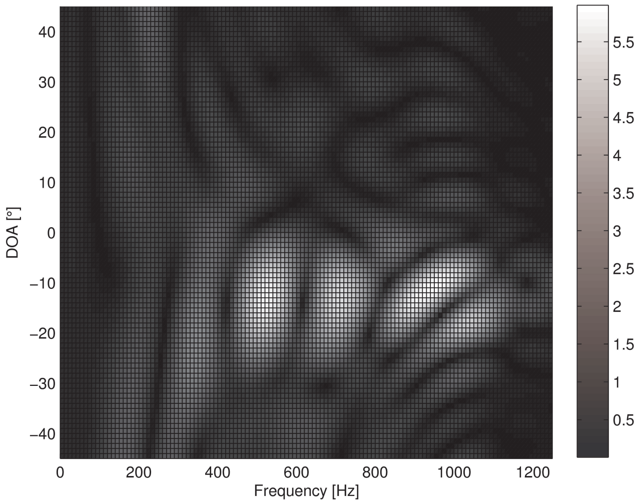


Fig. D.3: Contour plot of a 2-D Capon-based filter response designed for a signal constituted by five harmonics having a pitch of 200 Hz and a DOA of -15°. The signal was corrupted by white Gaussian noise with an SNR of 10 dB.

in all experiments the sampling frequency was $f_s = 2.5$ kHz. First, we evaluate the functionality of the filters in terms of investigating the filter response. In Fig. D.1, a filter response of a 2-D periodogram-based filter-bank is shown. The filter orders were in this case set to $M_t = M_s = 20$. The filter-bank was designed for a signal constituted by five harmonics with a pitch of 200 Hz and a DOA of -15°. As mentioned, this filter is independent of the signal statistics and will therefore have a rather similar attenuation of signal components not having both one of the target harmonic frequencies and the target DOA. At the target harmonic frequencies and DOA, the filter will have a unit gain. These observations can also be made from the experimental result which verifies the filter design. Likewise, we also conducted experiments on the 2-D Capon-based filter response. In these experiments the sample lengths were set to $N_t = N_s = 80$ and the filter orders were the same as in the previous experiment. The filter was designed for a signal having five harmonics with a pitch of 200 Hz and a DOA of -15°. Also, the signal was corrupted by white Gaussian noise and the experiment was conducted for SNRs of -40 dB and 10 dB. The results are depicted in Fig. D.2 and D.3, respectively. For the low SNR of -40 dB the filter response somehow resembles that of the 2-D periodogram-based filter which can also be verified mathematically. At high SNRs the 2-D Capon-based method seems to suffer from leakage, since it has a relatively high gain at off pitch frequencies and DOAs. This is, however, characteristic for the

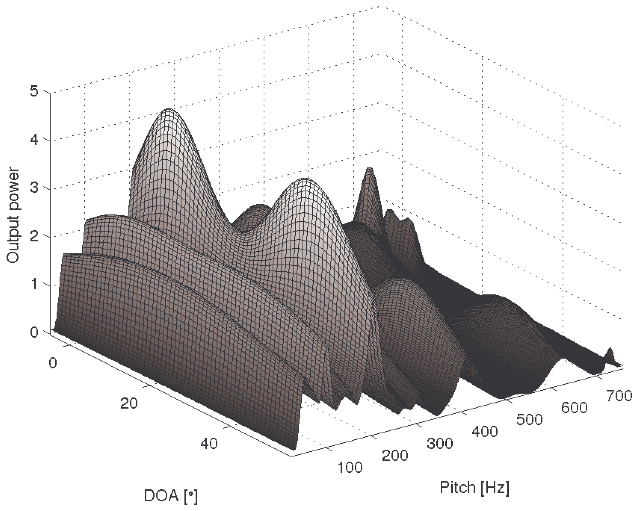


Fig. D.4: Output power of the 2-D periodogram-based filter-bank of orders $M_t = 30$ and $M_s = 10$ applied on a mixture of two signals with DOAs of 4° and 40°, respectively, and both with a pitch of 213 Hz. The SNR with respect to each signal was 10 dB.

minimum variance distortionless response (MVDR) principle.

Following, we evaluate the proposed methods ability to jointly estimate the pitch and DOA in a multi-source scenario. First, we have calculated the output power of both the 2-D periodogram-based filter and the 2-D Capon-based filter for several candidate pitch frequencies and DOAs. In these experiments, the number of sensors was $N_s = 30$, the sample length was $N_t = 100$ and the filter orders were $M_t = 30$ and $M_s = 10$. The filters were applied on a mixture of two signals both constituted by four sinusoids. The two signals had DOAs of 4° and 40°, respectively, and they both had a pitch of 213 Hz. The results from the experiments are depicted in Fig. D.4 and D.5, respectively. The first observation from these experiments are, that the pitch frequencies and DOAs of the two sources can be estimated correctly using both methods by taking the arguments of the two largest peaks of the filter output powers. Also, we observe that the peaks in the filter output power are much narrower for the 2-D Capon-based method which indicates that the 2-D Capon-based method will be superior when it comes to resolving closely-spaced sources.

We will now evaluate further on the proposed methods ability to resolve closely spaced sources. For this purpose we have conducted some Monte-Carlo simulations on the estimation error as a function of the source spacing in a two-source scenario. Due to a high computational complexity, we assumed that the pitch and DOA estimates

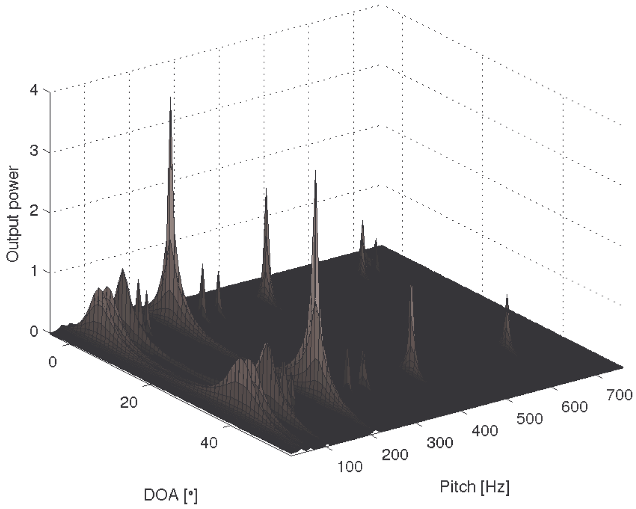


Fig. D.5: Output power of the 2-D Capon-based filter of orders $M_t = 30$ and $M_s = 10$ applied on a mixture of two signals with DOAs of 4° and 40°, respectively, and both with a pitch of 213 Hz. The SNR with respect to each signal was 10 dB.

were close to the true pitch and DOA which allowed us to just doing a gradient search from the true pitch and DOA instead of doing a fine grid search. In the first series of Monte-Carlo simulations we measured the MSE of the DOA estimates as a function of the source spacing in degrees. The two sources were each constituted by four harmonics having a pitch of 236 Hz. The DOA was fixed to -20° for one of the sources while the other was varied. White Gaussian noise was added to the signal such that the SNR with respect to each source was 10 dB. The sample lengths were $N_t = N_s = 30$ while the filter orders were $M_t = M_s = 10$. We conducted 500 Monte-Carlo simulations for each of the different source spacings and the results are depicted in Fig. D.6. From the results, it is clearly seen that the 2-D Capon-based method are superior in multi-source scenarios. The 2-D periodogram-based method shows some thresholding behavior around a source spacing of 35° while the 2-D Capon-based method does not show any thresholding behavior before a source spacing of only 10°. Note that the MSE for the 2-D periodogram-based method is not necessarily decreasing when the source spacing is increased. This is due to the fact that this method is signal independent and the MSE will therefore depend heavily on if the filter response occasionally has a dip at the pitch and DOA of the interfering source or not. We also conducted a series of Monte-Carlo simulations where the MSE was measured as a function of the source spacing in Hz. In these simulations, the two sources were constituted by one

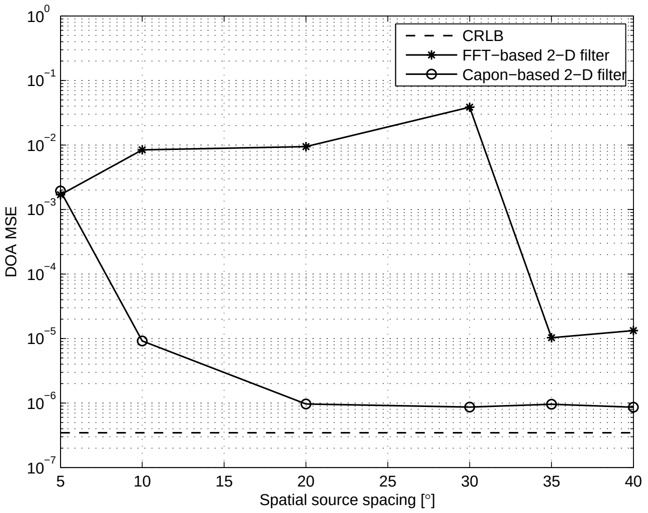


Fig. D.6: MSE in a two-source scenario as a function of the source spacing in degrees, and the CRLB for the single-source scenario.

2-D sinusoid. Both signals were having a DOA of 7°. The frequency of one of the sources was fixed to 200 Hz while the frequency of the other source was varied. The noise conditions, sample lengths and filter orders were the same as in the other series of Monte-Carlo simulations. Again, we conducted 500 Monte-Carlo simulations for each source spacing and the results are depicted in Fig. D.7. We see from the results that the 2-D Capon-based method is better for resolving closely-spaced sources. The dips in MSE at certain spacings for the 2-D periodogram-based method can be explained in the same way as in the previous series of Monte-Carlo simulations. Note that the 2-D periodogram-based method shows thresholding behavior below a spacing of 250 Hz whereas the 2-D Capon-based method does not show any thresholding behavior until below a spacing of 100 Hz.

4 Conclusion

In this paper, we proposed two new 2-D filtering methods for joint estimation of the pitch and the DOA of periodic signals recorded in space and time by a ULA. Since the proposed methods are based on a harmonic model, they are relevant for all signals being periodic of nature such as audio and speech signals. The first proposed method is based on the 2-D periodogram, and it can thereby be implemented easily using the

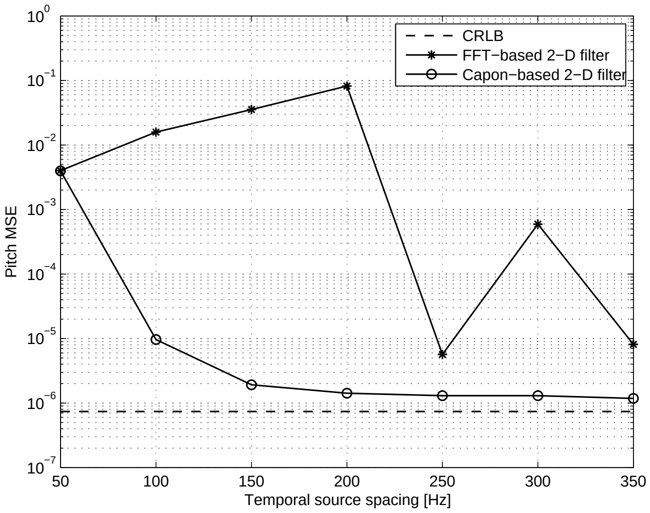


Fig. D.7: MSE in a two-source scenario as a function of the source spacing in Hz, and the CRLB for the single-source scenario.

2-D fast Fourier transform (FFT). By doing this, the computational complexity can be reduced significantly. The second proposed method is the 2-D Capon-based method, which is a generalization of the 2-D Capon method. Our experiments showed that both proposed methods can be used for jointly estimating the pitch frequencies and DOAs of multiple sources that are only separated sufficiently in one dimension, i.e., either space or time. We evaluated the proposed methods with respect to the sufficiency condition in regard to separation and these experiments showed that the 2-D Capon-based method outperforms the 2-D periodogram-based method in multi-source scenarios under adverse conditions.

References

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

[2] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Education, Inc., 2005.

[3] M. Brandstein and D. Ward, Eds., *Microphone Arrays - Signal Processing Techniques and Applications*. Springer-Verlag, 2001.

[4] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[5] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis - Principles, Algorithm, and Applications*. John Wiley & Sons, Inc., 2006.

[6] B. Edler and H. Purnhagen, "Parametric audio coding," in *Proc. Conf. Signal Process.*, vol. 1, 2000, pp. 21–24.

[7] M. Viberg and P. Stoica, "A computationally efficient method for joint direction finding and frequency estimation in colored noise," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, vol. 2, Nov. 1998, pp. 1547–1551.

[8] A. Jakobsson, S. L. Jr. Marple, and P. Stoica, "Computationally efficient two-dimensional Capon spectrum analysis," *IEEE Trans. Signal Process.*, vol. 48, no. 9, pp. 2651–2661, Sep. 2000.

[9] A. N. Lemma, A.-J. van der Veen, and E. F. Deprettere, "Analysis of joint angle-frequency estimation using ESPRIT," *IEEE Trans. Signal Process.*, vol. 51, no. 5, pp. 1264–1283, May 2003.

[10] T. Shu and X. Liu, "Robust and computationally efficient signal-dependent method for joint DOA and frequency estimation," *EURASIP J. on Advances in Signal Processing*, vol. 2008, no. 1, pp. 1–16, Apr. 2008.

[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] 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.

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

[14] 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.

[15] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

Paper E

Nonlinear Least Squares Methods for Joint DOA and Pitch Estimation

Jesper Rindom Jensen, Mads Græsbøll Christensen,
and Søren Holdt Jensen

The paper has been submitted to the
IEEE Trans. on Audio, Speech and Language Processing, 2012.

In peer-review

The layout has been revised.

Abstract

In this paper, we consider the problem of joint direction-of-arrival (DOA) and fundamental frequency estimation. Joint estimation enables robust estimation of these parameters in multi-source scenarios where separate estimators may fail. First, we derive the exact and asymptotic Cramér-Rao bounds for the joint estimation problem. Then, we propose nonlinear least squares (NLS) and an approximate NLS (aNLS) estimators for joint DOA and fundamental frequency estimation. The proposed estimators are maximum likelihood estimators when: 1) the noise is white Gaussian, 2) the environment is anechoic, and 3) the source of interest is in the far-field. Otherwise, the methods still yield approximately maximum likelihood estimates. Simulations on synthetic data show that the proposed methods have similar or better performance than state-of-the-art methods for DOA and fundamental frequency estimation. Moreover, simulations on real-life data indicate that the NLS and aNLS methods are applicable even when reverberation is present and the noise is not white Gaussian.

1 Introduction

Both direction-of-arrival (DOA) estimation and fundamental frequency estimation are very important signal processing topics, and, individually, these two estimation problems are widely studied research topics. DOA estimation, for example, has been treated in many text books and research papers (see, e.g., [1–6]) and has a multitude of applications in areas such as geophysics, radio astronomy, biomedical engineering, radar and microphone arrays. Fundamental frequency estimation (we will also refer to this as pitch estimation), on the other hand, has applications such as compression, separation and enhancement of, e.g., audio and voiced speech [7, 8], automatic music transcription and music classification [9]. For an overview of existing pitch estimation techniques, see, e.g., [9–13]. That is, both DOA and pitch estimation are relevant for processing of audio and speech signals. A few examples of applications which can benefit from the knowledge of both the DOA and the pitch are hands-free communication, teleconferencing, surveillance applications and hearing aids.

It is therefore natural to consider joint spatio-temporal processing of audio and speech signals which is the topic of this paper. More specifically, we consider joint DOA and pitch estimation. Besides the convenience of being able to estimate the DOA and the pitch simultaneously, joint spatio-temporal processing potentially has two significant advantages. For instance, if a source parameter is equal for both sources in a two-source scenario, the sources are not resolvable if we only estimate this parameter separately. If joint parameter estimation of several parameters is performed and just some of the parameters are distinct, then the sources are possibly still resolvable. Another important advantage of joint estimation relates to the estimation accuracy. For example, DOA and pitch estimation of periodic sources such as audio and voiced speech can be conducted by first estimating the DOA, then by extracting the signal impinging from that DOA, and finally by estimating the pitch from the extracted signal. However, the extraction can be seen as a linear data transformation which potentially increases the Cramér-Rao bound (CRB) for the pitch estimate, meaning that the resulting estimates may be suboptimal. Other important issues regarding processing of multi-channel signals are, e.g., reverberation and array calibration errors. We refer the interested reader to [14, 15] for an overview of methods dealing with these problems as these topics are out of the scope of this paper.

Motivated by the above observations, and due to an increasing computational capability, the computationally demanding problem of joint DOA and pitch estimation has attracted considerable attention in the recent years. As a result, some methods have been proposed for solving the joint estimation problem. Basically, these methods can be divided into two groups. The first group jointly estimates the frequency and the DOA of a single sinusoid defined in two dimensions (e.g., time and space). A few examples of such methods are [16], where a state-space realization technique is used, [17–19] which is based on the 2-D minimum variance distortionless response (MVDR) method, [20] which is based on the ESPRIT method, and [21] where a signal-dependent multistage Wiener filter (MWF) [22] is used. This group of methods is not commonly used in speech and audio processing. In most of the literature (see, e.g., [23–27]), DOA estimation of audio and speech recorded using a microphone array, has been treated as a broadband problem. In this paper, however, we shall assume a harmonic model which describes audio and voiced speech well; this will incontrovertibly allow us to treat the joint DOA and pitch estimation problem as L narrowband problems. Methods utilizing this fact forms the other group of estimators that consider the case with one or more harmonically related, two-dimensional sinusoids. These methods can, therefore, be seen as a generalization of the first group of methods. A few methods dealing with this case have been proposed; in [28] a ML-based method is proposed; in [29–31] subspace-based methods are proposed; in [32, 33] a correlation-based method is proposed; and in [34, 35] some spatio-temporal filtering methods based on the linearly constrained minimum variance (LCMV) beamformer [36] and the periodogram are proposed. Note that some of the above-mentioned methods considers time delay estimation and not DOA estimation, however, these two parameters are closely related.

In this paper, we also consider joint DOA and pitch estimation. Based on the harmonic model, we derive the exact and asymptotic CRBs for the joint DOA and pitch estimation problem. Moreover, we propose a non-linear least squares (NLS) method for joint DOA and pitch estimation. The proposed estimator is derived under the assumptions that the noise is white Gaussian, the array is a uniform linear array, the environment is anechoic, and the source of interest is located in the far-field of the array. When the assumptions hold, the proposed NLS estimator is also the maximum likelihood (ML) estimator as opposed to most of the existing joint DOA and pitch estimators [28–35]. Moreover, the proposed estimator is applicable in scenarios with any number of sensors, and it is easily generalized to support any array structure as opposed to the joint estimators in [29, 30]. Finally, we propose an approximate NLS (aNLS) method which is computationally more efficient.

The rest of the paper is organized as follows: in Section 2, we introduce the spatio-temporal harmonic signal model. Then, in Section 3, we derive the exact and asymptotic CRBs for the joint DOA and pitch estimation problem; the asymptotic bounds are used to motivate why the DOA and pitch should be estimated jointly. We derive the NLS and aNLS estimators for joint DOA and pitch estimation in Section 4, and the estimators are evaluated on synthetic as well as real-life signals in Section 5. Finally, Section 6 concludes our work.

2 Signal Model

In this paper, we consider joint estimation of the DOA, θ, and the pitch, ω_0, of a quasi-periodic source, also referred to as the source of interest (SOI), which is recorded using a N_s-element uniform linear array (ULA) in a noisy and anechoic environment. We assume that the noise is uncorrelated with the SOI. The ULA and the definition of θ are illustrated in Fig. E.1. Real-life examples of quasi-periodic sources are, e.g., voiced speech and musical instruments. We assume that the quasi-periodic source is in the far-field of the ULA. The signal measured on the n_sth sensor at time instance, n_t, for $n_\text{s} = 0, \ldots, N_\text{s} - 1$ and $n_\text{t} = 0, \ldots, N_\text{t} - 1$ is then given by

$$\begin{aligned} y_{n_\text{s}}(n_\text{t}) &= \beta_{n_\text{s}} s(n_\text{t} - f_\text{s}\tau_{n_\text{s}}) + w_{n_\text{s}}(n_\text{t}) \\ &= x_{n_\text{s}}(n_\text{t}) + w_{n_\text{s}}(n_\text{t}) \ , \end{aligned} \tag{E.1}$$

where β_{n_s} and τ_{n_s} are the attenuation and the delay of the wave generated by the SOI from sensor 0 to sensor n_s, respectively, f_s is the sampling frequency, $s(n_\text{t} - f_\text{s}\tau_{n_\text{s}})$ is the delayed quasi-periodic signal, and $w_{n_\text{s}}(n_\text{t})$ is the noise picked up by the n_sth sensor. Note that in the rest of the paper $(\cdot)_{n_\text{s}}$ means that the variable or constant is related to the n_sth sensor. Due to the array structure, we know that the delay is given by

$$\tau_{n_\text{s}} = n_\text{s}\frac{d \sin\theta}{c} \ , \tag{E.2}$$

where d is the inter-element spacing of the ULA, and c is the wave propagation velocity. Since the SOI is assumed to be quasi-periodic, we know that it can be modeled as a harmonic source,

$$s(n_\text{t}) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n_\text{t}} \ , \tag{E.3}$$

for $n_\text{t} = 0, \ldots, N_\text{t} - 1$, where L is the model order, $\alpha_l = A_l e^{j\phi_l}$, and $A_l > 0$ and ϕ_l are the real amplitude and phase of the lth harmonic. In case the desired signal

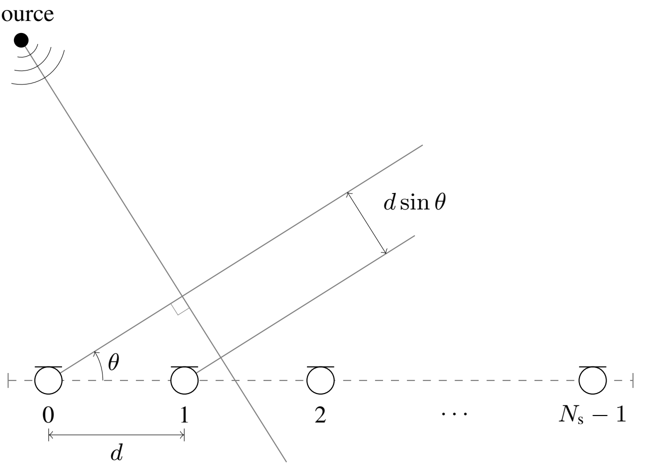


Fig. E.1: Illustration of the uniform linear array structure assumed in this paper.

has inharmonicities, the model can be extended to account for this [12, 37, 38]. Note that the signal model is complex as opposed to many real-life signals which are real. However, it is common to use complex signal representations since it leads to a simpler notation, and the complex model can easily be applied on real signals if we convert these to analytic signals using the Hilbert transform [6, 9]. In this paper, we consider the model order L as a known parameter (see, e.g., [6, 39] and the references therein for an overview of existing model order estimators).

Using the signal model in (E.3), the desired signal at sensor n_s can be written as

$$\begin{aligned} s(n_\mathrm{t} - f_\mathrm{s}\tau_{n_\mathrm{s}}) &= \sum_{l=1}^{L} \alpha_l e^{jl\omega_0(n_\mathrm{t} - f_\mathrm{s}\tau_{n_\mathrm{s}})} &\text{(E.4)}\\ &= \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n_\mathrm{t}} e^{-jl\omega_\mathrm{s} n_\mathrm{s}} \ , &\text{(E.5)} \end{aligned}$$

where $\omega_\mathrm{s} = \omega_0 f_\mathrm{s} \tau_1$ is the so-called spatial frequency. Note that the spatial frequency is dependent on the fundamental frequency, ω_0.

Additionally, we define a spatio-temporal matrix signal model, which is useful in the derivation of parameter estimators. The matrix model is defined as

$$\mathbf{Y}(n_\mathrm{t}) = \mathbf{X}(n_\mathrm{t}) + \mathbf{W}(n_\mathrm{t}) \ , \tag{E.6}$$

where

$$\mathbf{Y}(n_\mathrm{t}) = \begin{bmatrix} y_0(n_\mathrm{t}) & \cdots & y_0(n_\mathrm{t} - N_\mathrm{t} + 1) \\ \vdots & \ddots & \vdots \\ y_{N_\mathrm{s}-1}(n_\mathrm{t}) & \cdots & y_{N_\mathrm{s}-1}(n_\mathrm{t} - N_\mathrm{t} + 1) \end{bmatrix} , \tag{E.7}$$

with $\mathbf{X}(n_\mathrm{t})$ and $\mathbf{W}(n_\mathrm{t})$ being defined similarly to $\mathbf{Y}(n_\mathrm{t})$. The attenuated desired signal matrix $\mathbf{X}(n_\mathrm{t})$ can be rewritten using (E.5) as

$$\mathbf{X}(n_\mathrm{t}) = \boldsymbol{\beta} \sum_{l=1}^{L} \alpha_l(n_\mathrm{t}) \mathbf{z}_\mathrm{s}(l\omega_\mathrm{s}) \mathbf{z}_\mathrm{t}^T(l\omega_0) , \tag{E.8}$$

where

$$\alpha_l(n_\mathrm{t}) = \alpha_l e^{jl\omega_0 n_\mathrm{t}} , \tag{E.9}$$

$$\boldsymbol{\beta} = \mathrm{diag}\left\{ \begin{bmatrix} \beta_0 & \cdots & \beta_{N_\mathrm{s}-1} \end{bmatrix}^T \right\} , \tag{E.10}$$

$$\mathbf{z}_\mathrm{s}(\omega_\mathrm{s}) = \begin{bmatrix} 1 & e^{-j\omega_\mathrm{s}} & \cdots & e^{-j(N_\mathrm{s}-1)\omega_\mathrm{s}} \end{bmatrix}^T , \tag{E.11}$$

$$\mathbf{z}_\mathrm{t}(\omega_0) = \begin{bmatrix} 1 & e^{-j\omega_0} & \cdots & e^{-j(N_\mathrm{t}-1)\omega_0} \end{bmatrix}^T , \tag{E.12}$$

with diag$\{\cdot\}$ denoting the operator that transforms a vector into a diagonal matrix, and $(\cdot)^T$ denoting the transpose of a vector or matrix. Alternatively, the matrix model in (E.6) can be mapped to a vector model by stacking the columns of $\mathbf{Y}(n_\mathrm{t})$ as

$$\begin{aligned} \mathbf{y}(n_\mathrm{t}) &= \mathrm{vec}\{\mathbf{Y}(n_\mathrm{t})\} \\ &= \mathbf{x}(n_\mathrm{t}) + \mathbf{w}(n_\mathrm{t}) = \bar{\mathbf{Z}}\boldsymbol{\alpha}(n_\mathrm{t}) + \mathbf{w}(n_\mathrm{t}) , \end{aligned} \tag{E.13}$$

where vec$\{\cdot\}$ is the column-wise stacking operator, and

$$\bar{\mathbf{Z}} = \mathbf{B}\mathbf{Z} , \tag{E.14}$$

$$\mathbf{B} = \begin{bmatrix} \boldsymbol{\beta} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \boldsymbol{\beta} \end{bmatrix} , \tag{E.15}$$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}(\omega_0, \omega_\mathrm{s}) & \cdots & \mathbf{z}(L\omega_0, L\omega_s) \end{bmatrix} , \tag{E.16}$$

$$\mathbf{z}(l\omega_0, l\omega_\mathrm{s}) = \mathbf{z}_\mathrm{t}(l\omega_0) \otimes \mathbf{z}_\mathrm{s}(l\omega_\mathrm{s}) , \tag{E.17}$$

$$\boldsymbol{\alpha}(n_\mathrm{t}) = \begin{bmatrix} \alpha_1 e^{j\omega_0 n_\mathrm{t}} & \cdots & \alpha_L e^{jL\omega_0 n_\mathrm{t}} \end{bmatrix}^T , \tag{E.18}$$

with $\otimes$ denoting the Kronecker product operator. In summary, the objective considered in this paper is to estimate the DOA and the pitch jointly from spatio-temporal observed signal samples which can be modeled by (E.13).

3 Cramér-Rao Bounds

It is common practice to place a lower bound on the variance of unbiased estimators. This is useful while evaluating the performance of such estimators, and it provides insight into the nature of the estimation problem. There exists a multitude of such bounds among which the CRB is one of the most commonly used [40]. In this section, we derive exact and asymptotic expressions for the CRBs for the joint DOA and pitch estimation problem. Moreover, we show why it is beneficial to estimate the DOA and the pitch jointly by analyzing the asymptotic CRB expressions.

3.1 Exact Bounds

First, we derive the exact CRBs for the joint DOA and pitch estimation problem. Let

$$\bar{\mathbf{y}}(n_\text{t}) = \begin{bmatrix} y_0(n_\text{t}) & \cdots & y_{N_\text{s}-1}(n_\text{t}) \end{bmatrix}^T \tag{E.19}$$

be the observed signal vector from the N_s-element ULA at $n_\text{t} \in [0; N_\text{t}-1]$. We can also write the observation vector, $\bar{\mathbf{y}}(n_\text{t})$, as

$$\bar{\mathbf{y}}(n_\text{t}) = \bar{\mathbf{x}}(n_\text{t}) + \bar{\mathbf{w}}(n_\text{t}) \ , \tag{E.20}$$

where the noise vector, $\bar{\mathbf{w}}(n_\text{t})$, is defined similar to $\bar{\mathbf{y}}(n_\text{t})$ and

$$\begin{aligned} \bar{\mathbf{x}}(n_\text{t}) &= \begin{bmatrix} \beta_0 s(n_\text{t}-\tau_0) & \cdots & \beta_{N_\text{s}-1} s(n_\text{t}-\tau_{N_\text{s}-1}) \end{bmatrix}^T & \text{(E.21)} \\ &= \boldsymbol{\beta}\bar{\mathbf{s}}(n_\text{t}) \ , & \text{(E.22)} \\ \bar{\mathbf{s}}(n_\text{t}) &= \begin{bmatrix} s(n_\text{t}-\tau_0) & \cdots & s(n_\text{t}-\tau_{N_\text{s}-1}) \end{bmatrix}^T \ . & \text{(E.23)} \end{aligned}$$

We derive the CRBs under the assumption that the noise, $\bar{\mathbf{w}}(n_\text{t})$, is complex white Gaussian with zero mean and variance σ^2. Under this assumption, we can write the log-likelihood function of the observed signal as

$$\begin{aligned} \ln p(\bar{\mathbf{y}};\boldsymbol{\psi}) = &-N\ln(\pi\sigma^2) & \text{(E.24)} \\ &- \frac{1}{\sigma^2}\sum_{n_\text{t}=0}^{N_\text{t}-1} \left[\bar{\mathbf{y}}(n_\text{t}) - \boldsymbol{\beta}\bar{\mathbf{s}}(n_\text{t})\right]^H \left[\bar{\mathbf{y}}(n_\text{t}) - \boldsymbol{\beta}\bar{\mathbf{s}}(n_\text{t})\right] \ , \end{aligned}$$

where

$$\boldsymbol{\psi} = \begin{bmatrix} \omega_0 & \theta & A_1 & \cdots & A_L & \phi_1 & \cdots & \phi_L \end{bmatrix}^T \ , \tag{E.25}$$

The Fisher information matrix (FIM) for the joint DOA and pitch estimation problem is given by

$$\mathbf{I}(\boldsymbol{\psi}) = -\text{E}\left\{\frac{\partial^2 \ln p(\bar{\mathbf{y}};\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\right\} \ . \tag{E.26}$$

If we assume that the covariance matrix of the noise signal does not depend on the parameter vector, $\boldsymbol{\psi}$, the FIM is given by

$$\mathbf{I}(\boldsymbol{\psi}) = \frac{2}{\sigma^2}\text{Re}\left\{\sum_{n_\text{t}=0}^{N_\text{t}-1}\mathbf{D}_{n_\text{t}}^H(\boldsymbol{\psi})\boldsymbol{\beta}^2\mathbf{D}_{n_\text{t}}(\boldsymbol{\psi})\right\} , \tag{E.27}$$

where Re$\{\cdot\}$ denotes the real part of a complex number, and $\mathbf{D}_{n_\text{t}}(\boldsymbol{\psi})$ is the gradient matrix for time instance n_t defined as

$$\begin{aligned}\mathbf{D}_{n_\text{t}}(\boldsymbol{\psi}) = \big[\mathbf{d}_{n_\text{t}}(\omega_0) \quad \mathbf{d}_{n_\text{t}}(\theta) \quad \mathbf{d}_{n_\text{t}}(A_1) \\ \cdots \quad \mathbf{d}_{n_\text{t}}(A_L) \quad \mathbf{d}_{n_\text{t}}(\phi_1) \quad \cdots \quad \mathbf{d}_{n_\text{t}}(\phi_L)\big] .\end{aligned} \tag{E.28}$$

Note that the columns of $\mathbf{D}_{n_\text{t}}(\boldsymbol{\psi})$ can be interpreted as the gradient vectors with respect to each of the unknown parameters. The gradient vector with respect to the pitch, $\mathbf{d}_{n_\text{t}}(\omega_0)$, is defined as

$$\mathbf{d}_{n_\text{t}}(\omega_0) = \frac{\partial\bar{\mathbf{s}}(n_\text{t})}{\partial\omega_0} , \tag{E.29}$$

and the vectors $\mathbf{d}_{n_\text{t}}(\theta)$, $\mathbf{d}_{n_\text{t}}(A_l)$ and $\mathbf{d}_{n_\text{t}}(\phi_l)$ are defined similar to $\mathbf{d}_{n_\text{t}}(\omega_0)$ for $l = 1, \ldots, L$. The individual entries of the gradient vectors are given by

$$\begin{aligned}\left[\mathbf{d}_{n_\text{t}}(\omega_0)\right]_{n_\text{s}} =& \sum_{l=1}^{L} jlA_l\left(n_\text{t} - f_\text{s}n_\text{s}\frac{d\sin\theta}{c}\right) \\ & \times e^{jl\omega_0\left(n_\text{t}-f_\text{s}n_\text{s}\frac{d\sin\theta}{c}\right)+j\phi_l} , \end{aligned}\tag{E.30}$$

$$\begin{aligned}\left[\mathbf{d}_{n_\text{t}}(\theta)\right]_{n_\text{s}} =& -\sum_{l=1}^{L} jlA_l\omega_0 f_\text{s}n_\text{s}\frac{d\cos\theta}{c} \\ & \times e^{jl\omega_0\left(n_\text{t}-f_\text{s}n_\text{s}\frac{d\sin\theta}{c}\right)+j\phi_l} , \end{aligned}\tag{E.31}$$

$$\left[\mathbf{d}_{n_\text{t}}(A_l)\right]_{n_\text{s}} = e^{jl\omega_0\left(n_\text{t}-f_\text{s}n_\text{s}\frac{d\sin\theta}{c}\right)+j\phi_l} , \tag{E.32}$$

$$\left[\mathbf{d}_{n_\text{t}}(\phi_l)\right]_{n_\text{s}} = jA_l e^{jl\omega_0\left(n_\text{t}-f_\text{s}n_\text{s}\frac{d\sin\theta}{c}\right)+j\phi_l} , \tag{E.33}$$

for $n_\text{s} = 0, \ldots, N_\text{s} - 1$. The exact CRB for the kth parameter in $\boldsymbol{\psi}$ is defined as the (k, k)th element of the inverse FIM, i.e.,

$$\text{CRB}\left([\boldsymbol{\psi}]_k\right) = \left[\mathbf{I}^{-1}(\boldsymbol{\psi})\right]_{kk} . \tag{E.34}$$

3.2 Asymptotic Bounds

The exact CRB expressions are rather complicated, and it is difficult to see how the different parameters and the sample lengths influence the different CRBs. Furthermore, it is hard to see from the exact CRB expressions if there are any benefits of estimating the DOA and pitch jointly compared to estimating them separately. Therefore, we also derive simpler asymptotic CRBs for the joint DOA and pitch estimation problem.

The asymptotic bounds are derived under the assumption that the sensors in the ULA are closely spaced such that $\boldsymbol{\beta} \approx \mathbf{I}$. First, we introduce a new variable,

$$\Delta(x, y) = \sum_{n_\mathrm{t}=0}^{N_\mathrm{t}-1} \mathrm{Re}\left\{\mathbf{d}_{n_\mathrm{t}}^H(x)\mathbf{d}_{n_\mathrm{t}}(y)\right\} \tag{E.35}$$

$$= \Delta(y, x) \; . \tag{E.36}$$

For $N_\mathrm{s} \to \infty$ and $N_\mathrm{t} \to \infty$, we know that the frequency spaced sinusoids are orthogonal. For large N_s and N_t, it follows that

$$\begin{aligned}\Delta(\omega_0,\omega_0) &\approx \Bigg[\frac{N_\mathrm{t}(N_\mathrm{t}-1)(2N_\mathrm{t}-1)}{6}N_\mathrm{s} \\ &\quad +N_\mathrm{t}\zeta^2\sin^2\theta\frac{N_\mathrm{s}(N_\mathrm{s}-1)(2N_\mathrm{s}-1)}{6} \\ &\quad -\frac{N_\mathrm{t}(N_\mathrm{t}-1)}{2}\zeta\sin\theta N_\mathrm{s}(N_\mathrm{s}-1)\Bigg]\sum_{l=1}^{L} l^2A_l^2\ , \end{aligned} \tag{E.37}$$

$$\begin{aligned}\Delta(\omega_0,\theta) &\approx \Bigg[-\frac{N_\mathrm{t}(N_\mathrm{t}-1)}{2}\omega_0\zeta\cos\theta\frac{N_\mathrm{s}(N_\mathrm{s}-1)}{2} \\ &\quad +N_\mathrm{t}\omega_0\zeta^2\frac{\sin 2\theta}{2}\frac{N_\mathrm{s}(N_\mathrm{s}-1)(2N_\mathrm{s}-1)}{6}\Bigg]\sum_{l=1}^{L} l^2A_l^2, \end{aligned} \tag{E.38}$$

$$\Delta(\omega_0,A_l) \approx 0 \tag{E.39}$$

$$\Delta(\omega_0,\phi_l) \approx \left[\frac{N_\mathrm{t}(N_\mathrm{t}-1)}{2}N_\mathrm{s} - N_\mathrm{t}\zeta\sin\theta\frac{N_\mathrm{s}(N_\mathrm{s}-1)}{2}\right]lA_l^2, \tag{E.40}$$

$$\Delta(\theta,\theta) \approx N_\mathrm{t}\omega_0^2\zeta^2\cos\theta\frac{N_\mathrm{s}(N_\mathrm{s}-1)(2N_\mathrm{s}-1)}{6}\sum_{l=1}^{L} l^2A_l^2, \tag{E.41}$$

$$\Delta(\theta,A_l) \approx 0\ , \tag{E.42}$$

$$\Delta(\theta,\phi_l) \approx -N_\mathrm{t}\omega_0\zeta\cos\theta\frac{N_\mathrm{s}(N_\mathrm{s}-1)}{2}lA_l^2\ , \tag{E.43}$$

$$\Delta(A_p,A_q) = \begin{cases} N_\mathrm{t}N_\mathrm{s}, & p=q \\ (\approx)0, & p\neq q\ , \end{cases} \tag{E.44}$$

$$\Delta(A_p,\phi_q) = \begin{cases} 0, & p=q \\ (\approx)0, & p\neq q\ , \end{cases} \tag{E.45}$$

$$\Delta(\phi_p,\phi_q) = \begin{cases} N_\mathrm{t}N_\mathrm{s}A_l^2, & p=q \\ (\approx)0, & p=q\ , \end{cases} \tag{E.46}$$

where $\zeta = \frac{f_s d}{c}$. Furthermore, we know that [41]

$$\begin{bmatrix}\mathbf{A} & \mathbf{U}\\ \mathbf{V} & \mathbf{B}\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{C}^{-1} & -\mathbf{C}^{-1}\mathbf{U}\mathbf{B}^{-1}\\ -\mathbf{B}^{-1}\mathbf{V}\mathbf{C}^{-1} & \mathbf{B}^{-1}\mathbf{V}\mathbf{C}^{-1}\mathbf{U}\mathbf{B}^{-1}+\mathbf{B}^{-1}\end{bmatrix}, \tag{E.47}$$

with $\mathbf{C} = \mathbf{A} - \mathbf{U}\mathbf{B}^{-1}\mathbf{V}$. We now apply (E.47) on the FIM with the expressions in

(E.37)-(E.46) and with

$$\mathbf{A} = \begin{bmatrix} \Delta(\omega_0,\omega_0) & \Delta(\omega_0,\theta) \\ \Delta(\theta,\omega_0) & \Delta(\theta,\theta) \end{bmatrix} , \tag{E.48}$$

$$\mathbf{U} = \begin{bmatrix} 0 & \Delta(\omega_0,\phi_1) & \cdots & 0 & \Delta(\omega_0,\phi_L) \\ 0 & \Delta(\theta,\phi_1) & \cdots & 0 & \Delta(\theta,\phi_L) \end{bmatrix} \tag{E.49}$$

$$= \mathbf{V}^H , \tag{E.50}$$

$$\begin{aligned}\mathbf{B} = \text{diag}\{[\Delta(A_1,A_1) \quad \Delta(\phi_1,\phi_1) \quad \cdots \\ \Delta(A_L,A_L) \quad \Delta(\phi_L,\phi_L)]\} .\end{aligned} \tag{E.51}$$

The asymptotic CRBs of the DOA and the pitch can then be found from the diagonal elements of the matrix, $\mathbf{C}^{-1}$. Here, we only derive the asymptotic CRBs for these two parameters while the derivations for the other parameters are left to the interested reader. Some tedious manipulations yield

$$\text{CRB}(\omega_0) \approx \frac{6}{N_\text{t}^3 N_\text{s}} \text{PSNR}^{-1} , \tag{E.52}$$

$$\begin{aligned}\text{CRB}(\theta) \approx \Bigg[\left(\frac{c}{\omega_0 f_\text{s} d \cos\theta} \right)^2 \frac{6}{N_\text{t} N_\text{s}^3} \\ + \left(\frac{\tan\theta}{\omega_0} \right)^2 \frac{6}{N_\text{t}^3 N_\text{s}} \Bigg] \text{PSNR}^{-1} ,\end{aligned} \tag{E.53}$$

where

$$\text{PSNR} = \frac{\sum_{l=1}^{L} l^2 A_l^2}{\sigma^2} \tag{E.54}$$

is the so-called pseudo signal-to-noise ratio. In Fig. E.2, we see that the asymptotic bounds indeed approaches the exact bounds for large N_ss and N_ts. To obtain the results in Fig. E.2, we used the following set up: the pitch was $f_0 = 100$ Hz, the DOA was $\theta = 20°$, the model order was $L = 4$, the variance of the noise was $\sigma^2 = 0.1$, the sampling frequency was $f_\text{s} = 2$ kHz, the wave propagation speed was $c = 340$ m/s, and the inter-element spacing was $d = 2c/f_\text{s}$. Furthermore, the number of sensors was $N_\text{s} = 20$ for the simulations with varying N_t, and the number of samples was $N_\text{t} = 20$ for the simulations with varying N_s.

3.3 Motivation for Joint DOA and Pitch Estimation

By investigating the asymptotic CRB expressions, it can be seen that the bound for ω_0 is decreasing cubically in N_t and linearly in N_s. The bound for θ consists of two terms; one of the terms is linear in N_t and cubical in N_s, and vice versa for the other term. Moreover, it can be seen that it is beneficial to estimate the DOA and the pitch jointly

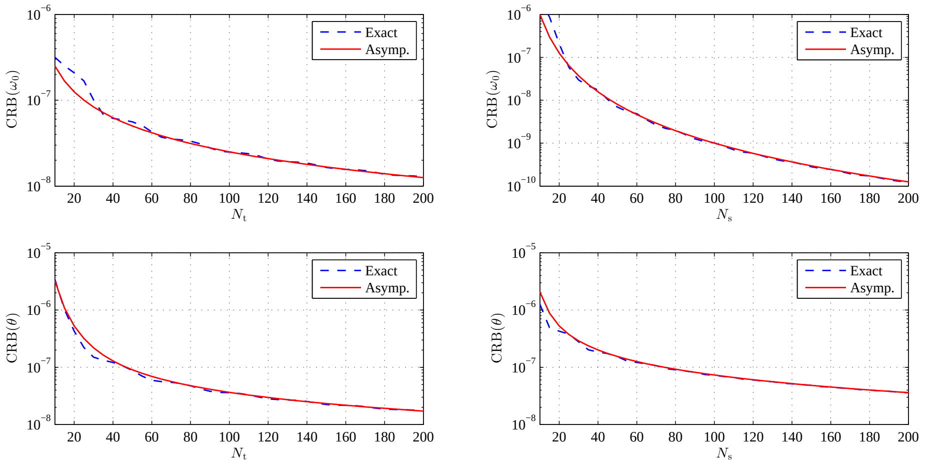


Fig. E.2: Plot of the exact and asymptotic Cramér-Rao bounds for (top) the pitch and (bottom) the DOA of the joint DOA and pitch estimation problem as a function of (left) N_t and (right) N_s.

rather then separately. First, we can see from the asymptotic DOA bound in (E.53) that the CRB is decreased by taking the harmonic signal structure into account as opposed to if we estimated the DOA of a single sinusoid since the bound depends on the PSNR. Moreover, we can see from the asymptotic bound in (E.52) that the CRB of the pitch can be decreased linearly by increasing the number of sensors, N_s.

The DOA and the pitch could also be estimated separately using a two-step procedure where we 1) estimate the DOA and extract the signal impinging from the estimated DOA, and 2) estimate the pitch from the extracted signal. Similarly, we could also estimate the pitch first, extract the signal with the estimated pitch, and then estimate the DOA of the extracted signal. We will term such estimation methods as cascaded methods. The cascaded methods, however, will most likely increase the CRBs of the parameters to be estimated in the second step. The cause of the CRB increase is the signal extraction occurring in the first step of the cascaded methods, since the extraction is often performed by a filter which, in general, does not span or contains the subspace spanned by the gradient matrix, $\mathbf{D}_{n_\text{t}}(\boldsymbol{\psi})$.

4 Joint DOA and Pitch Estimation

In this section, we propose two estimators that jointly estimate the DOA and the pitch of a periodic source that is sampled by a ULA. The methods are based on nonlinear least-squares (NLS), and they are derived under a white Gaussian noise assumption.

4.1 Nonlinear Least-Squares Method

First, we derive the NLS method for joint DOA and pitch estimation. The NLS method is derived under the assumption that the noise is white Gaussian. If the noise is indeed white Gaussian, the proposed NLS method resembles the maximum likelihood (ML) estimator, i.e., it will attain the CRB. The proposed NLS method may even provide accurate estimates when the noise is not white Gaussian, as the NLS method for a single sinusoid derived for white Gaussian noise is asymptotically efficient even for colored noise [42].

In this paper, the attenuation matrix $\boldsymbol{\beta}$ is considered as known, i.e., the joint NLS estimates of the DOA and pitch are found by solving

$$\left\{\hat{\theta}, \hat{\omega}_0\right\} = \arg \min_{\boldsymbol{\alpha}, \{\theta, \omega_0\} \in \Theta \times \Omega} \left\|\mathbf{y} - \bar{\mathbf{Z}}\boldsymbol{\alpha}\right\|_2^2 , \tag{E.55}$$

with $\|\cdot\|_2$ denoting the ℓ_2-norm. Minimizing (E.55) with respect to the complex amplitude vector, $\boldsymbol{\alpha}$, yields

$$\hat{\boldsymbol{\alpha}} = (\bar{\mathbf{Z}}^H \bar{\mathbf{Z}})^{-1} \bar{\mathbf{Z}}^H \mathbf{y} . \tag{E.56}$$

If we then insert (E.56) into (E.55), we get that

$$\left\{\hat{\theta}, \hat{\omega}_0\right\} = \arg \max_{\{\theta, \omega_0\} \in \Theta \times \Omega} \mathbf{y}^H \bar{\mathbf{Z}} \left(\bar{\mathbf{Z}}^H \bar{\mathbf{Z}}\right)^{-1} \bar{\mathbf{Z}}^H \mathbf{y} , \tag{E.57}$$

The above estimator is referred to as the NLS estimator. If we keep only the highest order terms, the complexity of the estimator per point in the search grid $\Theta \times \Omega$ is $\mathcal{O}(NL^2 + L^3)$ where $N = N_\text{t} N_\text{s}$. On basis of (E.57), we define the NLS cost-functions as

$$\begin{aligned} J_\text{NLS}(\theta, \omega_0) &= \left\|\bar{\mathbf{Z}}^H \mathbf{y}\right\|^2_{\left(\bar{\mathbf{Z}}^H \bar{\mathbf{Z}}\right)^{-1}} \\ &= \text{Tr}\left\{\bar{\mathbf{Z}}^H \mathbf{y}\mathbf{y}^H \bar{\mathbf{Z}} \left(\bar{\mathbf{Z}}^H \bar{\mathbf{Z}}\right)^{-1}\right\} , \end{aligned} \tag{E.58}$$

with $\|\cdot\|^2_\mathbf{W}$ denotes the weighted ℓ_2-norm where $\mathbf{W}$ is the weighting matrix. Instead of only using a single-data snapshot, $\mathbf{y}$, in the cost-function in (E.58), we could replace $\mathbf{y}$ by

$$\mathbf{y}_{n_\text{s}}(n_\text{t}) = \text{vec}\left\{\begin{bmatrix} y_{n_\text{s}}(n_\text{t}) & \cdots & y_{n_\text{s}}(n_\text{t} - M'_\text{t}) \\ \vdots & \ddots & \vdots \\ y_{n_\text{s}+M'_\text{s}}(n_\text{t}) & \cdots & y_{n_\text{s}+M'_\text{s}}(n_\text{t} - M'_\text{t}) \end{bmatrix}\right\} , \tag{E.59}$$

with $M'_\text{s} = M_\text{s} - 1$, $M'_\text{t} = M_\text{t} - 1$, $M_\text{s} \leq N_\text{s}$, and $M_\text{t} \leq N_\text{t}$ in (E.58). If we then take the expected value, we get

$$\text{E}\left\{\|\bar{\mathbf{Z}}^H \mathbf{y}_{n_\text{s}}(n_\text{t})\|^2_{(\bar{\mathbf{Z}}^H \bar{\mathbf{Z}})^{-1}}\right\} = \text{Tr}\left\{\bar{\mathbf{Z}}^H \mathbf{R} \bar{\mathbf{Z}} \left(\bar{\mathbf{Z}}^H \bar{\mathbf{Z}}\right)^{-1}\right\} . \tag{E.60}$$

That is, we can also estimate the DOA and pitch jointly by matching the signal model to the covariance matrix, $\mathbf{R}$, of $\mathbf{y}_{n_\text{s}}(n_\text{t})$ as

$$\left\{\hat{\theta}, \hat{\omega}_0\right\} = \arg \max_{\{\theta,\omega_0\}\in\Theta\times\Omega} \text{Tr}\left\{\bar{\mathbf{Z}}^H\mathbf{R}\bar{\mathbf{Z}}\left(\bar{\mathbf{Z}}^H\bar{\mathbf{Z}}\right)^{-1}\right\} . \tag{E.61}$$

The computational complexity per grid point for the expectation based estimator is $\mathcal{O}(L^2M + LM^2 + L^3)$ where $M = M_\text{t}M_\text{s}$. Note that even though this complexity looks worse than for the single snapshot NLS estimator it might not be the case in all scenarios since $M \leq N$.

In practice, we do not know the exact covariance matrix $\mathbf{R}$, but we can replace it by, e.g., the sample covariance matrix estimate defined as [19]

$$\hat{\mathbf{R}} = \sum_{m_\text{s}=0}^{N_\text{s}-M_\text{s}} \sum_{m_\text{t}=0}^{N_\text{t}-M_\text{t}} \frac{\mathbf{y}_{m_\text{s}}(n_\text{t}-m_\text{t})\mathbf{y}_{m_\text{s}}^H(n_\text{t}-m_\text{t})}{(N_\text{s}-M_\text{s}')(N_\text{t}-M_\text{t}')} . \tag{E.62}$$

The gradient of the cost-function, $J_\text{NLS}(\theta,\omega_0)$, is given by

$$\nabla J_\text{NLS}(\theta,\omega_0) = \begin{bmatrix} \frac{\partial J_\text{NLS}}{\partial\theta} & \frac{\partial J_\text{NLS}}{\partial\omega_0} \end{bmatrix}^T , \tag{E.63}$$

where

$$\frac{\partial J_\text{NLS}}{\partial\theta} = \mathbf{y}^H\left(\mathbf{G}_\theta\mathbf{P}^\perp + \mathbf{P}^\perp\mathbf{G}_\theta^H\right)\mathbf{y} , \tag{E.64}$$

$$\frac{\partial J_\text{NLS}}{\partial\omega_0} = \mathbf{y}^H\left(\mathbf{G}_{\omega_0}\mathbf{P}^\perp + \mathbf{P}^\perp\mathbf{G}_{\omega_0}^H\right)\mathbf{y}, \tag{E.65}$$

with

$$\mathbf{P}^\perp = (\mathbf{I} - \mathbf{P}), \tag{E.66}$$

$$\mathbf{P} = \bar{\mathbf{Z}}(\bar{\mathbf{Z}}^H\bar{\mathbf{Z}})^{-1}\bar{\mathbf{Z}}^H , \tag{E.67}$$

$$\mathbf{G}_\theta = \bar{\mathbf{Z}}\left(\bar{\mathbf{Z}}^H\bar{\mathbf{Z}}\right)^{-1}\mathbf{Y}_\theta^H\mathbf{B} , \tag{E.68}$$

$$\mathbf{G}_{\omega_0} = \bar{\mathbf{Z}}\left(\bar{\mathbf{Z}}^H\bar{\mathbf{Z}}\right)^{-1}\mathbf{Y}_{\omega_0}^H\mathbf{B} , \tag{E.69}$$

$$[\mathbf{Y}_\theta]_{pq} = -jq\omega_0\zeta\cos\theta k_{\text{s},p}e^{-jq\omega_0(\zeta\sin\theta k_{\text{s},p}+k_{\text{t},p})} , \tag{E.70}$$

$$[\mathbf{Y}_{\omega_0}]_{pq} = -jq\left(\zeta\sin\theta k_{\text{s},p} + k_{\text{t},p}\right)e^{-jq\omega_0(\zeta\sin\theta k_{\text{s},p}+k_{\text{t},p})} , \tag{E.71}$$

$k_{\text{s},p} = (p-1) \pmod{M_\text{s}}$ and $k_{\text{t},p} = \lfloor\frac{p-1}{M_\text{s}}\rfloor$. Note that $y \pmod{x}$ denotes that y is modulo x, and $\lfloor\cdot\rfloor$ is the floor operator. Using the gradient in (E.63), we can iteratively obtain refined estimates of the DOA and the pitch as

$$\begin{bmatrix} \hat{\theta}^{(i+1)} \\ \hat{\omega}_0^{(i+1)} \end{bmatrix} = \begin{bmatrix} \hat{\theta}^{(i)} \\ \hat{\omega}_0^{(i)} \end{bmatrix} + \delta\nabla J_\text{NLS} , \tag{E.72}$$

where i is the iteration index and $\delta > 0$ is a small constant which can be found using a line search algorithm.

4.2 Approximate Nonlinear Least-Squares Method

When the number of spatial and temporal samples are large, the harmonics are close to being orthogonal, i.e., [9]

$$\lim_{M\to\infty} \frac{1}{M}\mathbf{Z}^H\mathbf{Z} = \mathbf{I}\,. \tag{E.73}$$

Therefore, cf. (E.14) and the mixed-product property

$$(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = \mathbf{AC}\otimes\mathbf{BD}, \tag{E.74}$$

it can be shown that

$$\lim_{M\to\infty} \frac{1}{M}\bar{\mathbf{Z}}^H\bar{\mathbf{Z}} = \frac{\|\boldsymbol{\beta}\|_2^2}{M_\text{s}}\mathbf{I}\,. \tag{E.75}$$

Inserting (E.75) into (E.57) yields the approximate NLS (aNLS) estimator defined as

$$\left\{\hat{\theta}, \hat{\omega}_0\right\} = \arg \max_{\{\theta,\omega_0\}\in\Theta\times\Omega} \mathbf{y}^H\bar{\mathbf{Z}}\bar{\mathbf{Z}}^H\mathbf{y}\,. \tag{E.76}$$

That is, the aNLS cost-function is given by

$$J_\text{aNLS}(\theta,\omega_0) = \|\bar{\mathbf{Z}}^H\mathbf{y}\|_2^2\,. \tag{E.77}$$

The computational complexity per search grid point of the aNLS estimator is $\mathcal{O}(NL)$, i.e., it is only quadratic compared to the complexity of the NLS estimator which was cubic[1]. As for the NLS method, we also propose an alternative covariance-based estimator

$$\left\{\hat{\theta}, \hat{\omega}_0\right\} = \arg \max_{\{\theta,\omega_0\}\in\Theta\times\Omega} \text{Tr}\left\{\bar{\mathbf{Z}}^H\mathbf{R}\bar{\mathbf{Z}}\right\} \tag{E.78}$$

The computational complexity of evaluating the expectation based aNLS estimator in each point of the search grid is $\mathcal{O}(M^2L + ML^2)$. Note that the alternative aNLS cost-function in (E.78) can be interpreted as the output power of a periodogram-based filterbank when $\mathbf{B} = \mathbf{I}$.

The expressions for the partial derivatives of the aNLS cost-function are given by

$$\frac{\partial J_\text{aNLS}}{\partial\theta} = \mathbf{y}^H\left(\mathbf{B}\mathbf{Y}_\theta\bar{\mathbf{Z}}^H + \bar{\mathbf{Z}}\mathbf{Y}_\theta^H\mathbf{B}\right)\mathbf{y}\,, \tag{E.79}$$

$$\frac{\partial J_\text{aNLS}}{\partial\omega_0} = \mathbf{y}^H\left(\mathbf{B}\mathbf{Y}_{\omega_0}\bar{\mathbf{Z}}^H + \bar{\mathbf{Z}}\mathbf{Y}_{\omega_0}^H\mathbf{B}\right)\mathbf{y}\,. \tag{E.80}$$

We can then obtain refined aNLS estimates by using (E.79) and (E.80) in (E.72).

[1] Here, we consider all unknown variables as one variable when counting the order, i.e., $\mathcal{O}(NL)$ is considered as a second order term.

5 Experimental Results

To evaluate the proposed joint DOA and pitch estimators, we conducted simulations on both synthetic as well as real-life data. The results from these simulations are explained in the following subsections.

5.1 Statistical Evaluation

We conducted several series of Monte-Carlo simulations using synthetic data. In all of these simulations, the sampling frequency was $f_\text{s} = 8$ kHz, the speed of sound was assumed to be $c = 343.2$ m/s, the array was uniform and linear with $d = \frac{c}{f_\text{s}}$, there was no attenuation across the sensors such that $\mathbf{B} = \mathbf{I}$, and the desired signal was designed to be a harmonic signal with $f_0 = 243$ Hz, $\theta = 15°$, $L = 5$ and $\alpha_l = 1$. We estimated the pitch and DOA in each of the simulations using different estimators including the proposed; besides the proposed estimators, we used the multichannel maximum likelihood (MC-ML) and multichannel approximate maximum likelihood (MC-aML) estimators [43] for pitch estimation, and we used the steered response power (SRP) method, the steered response power with phase transform (SRP-PHAT) method [44, 45], and the broadband MVDR (bMVDR) beamformer [17] for DOA estimation. Finally, we used the position-pitch plane (PoPi) based estimator in [32], the subspace (Sub.) method in [30], and the LCMV filtering (LCMV) method in [35] for joint DOA and pitch estimation. For pitch estimation, we compare with the MC-ML and MC-aMLS estimators since these were shown to outperform the multi-channel pitch estimators in [46, 47]. Note that in our implementations of the SRP and bMVDR methods we use an FFT length of 256, and we integrate over all frequency indices whereas in the SRP-PHAT method we integrate over the frequency indices corresponding to the interval [200 Hz; $L \max\{f_{0,\text{grid}}\}$] with $\max\{f_{0,\text{grid}}\}$ being our maximum pitch candidate. Moreover, in our implementation of the bMVDR method, we used 20 blocks of length $\lfloor N_\text{t}/3 \rfloor$ to estimate the cross-spectral density. We used an FFT size of 1024 for the PoPi method, a block size of $N_\text{t}/2$ and a smoothing factor of 5 for the subspace method, and spatial and temporal filter lengths of $\max\{[2, \lfloor N_\text{s} \cdot 2/3 \rfloor]\}$ and $\lfloor N_\text{t}/4 \rfloor$, respectively, for the LCMV method.

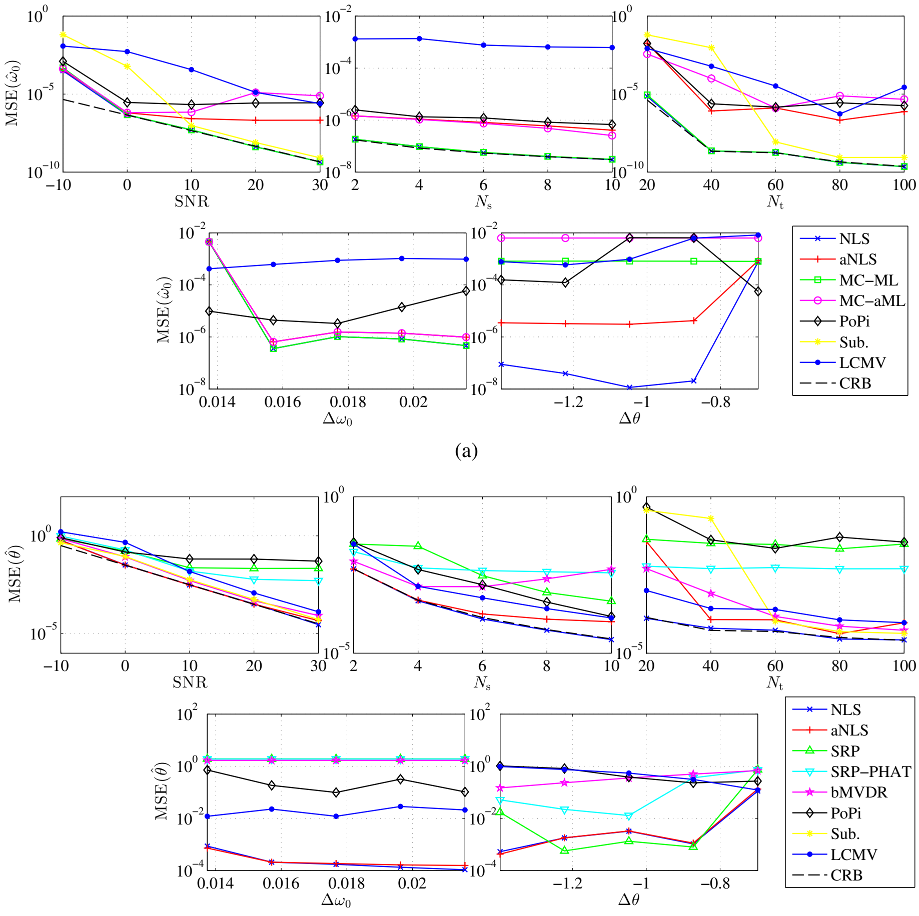
In each series of Monte-Carlo simulations, the performance of the estimators was measured in terms mean squared error (MSE). In the first series of Monte-Carlo simulations, we measured the estimation performance as a function of the signal-to-noise ratio (SNR) defined as

$$\text{SNR} = 10 \log_{10} \frac{\sum_{l=1}^{L} |\alpha_l|^2}{\sigma^2} \ . \tag{E.81}$$

The number of sensors were $N_\text{s} = 2$ and the number of temporal sample was $N_\text{t} = 80$. Then, we conducted another series of Monte-Carlo simulations where the estimation performance was evaluated versus the number of sensors N_s while the SNR was fixed to 10 dB and the number of temporal samples was $N_\text{t} = 60$. In the third series, we measured the performance as a function of the number of temporal samples N_t, and, here, the SNR was 30 dB while the number of sensors was $N_\text{s} = 2$. Finally, we conducted two series of Monte-Carlo simulations on synthetic data containing two harmonic sources each with five harmonics with unit amplitudes. In the first of these series, both sources had a DOA of 15°. One of the sources had a pitch of 243 Hz, while pitch of the other source was varied. The MSEs was then measured as the mean of the MSEs for the two sources. For this experiment, the number of sensors and samples was $N_\text{s} = 2$ and $N_\text{t} = 80$, respectively. In the other of these series, the pitch of both sources was 243 Hz, and the DOA of one of the sources was 15°, while the DOA of the other source was varying. The number of sensors and samples was $N_\text{s} = 8$ and $N_\text{t} = 60$, respectively, in this experiment.

The results from all of the series of Monte-Carlo simulations are depicted in Fig. E.3, and they reveal several interesting facts. First, we note that the proposed NLS estimator attains the CRB for both the DOA and pitch when the noise is white Gaussian and we have a single harmonic sources. This was also expected according to our previous claims. Moreover, the NLS estimator has a better or similar performance than all other methods in the comparison. The proposed aNLS estimator, however, is slightly biased and does therefore not attain the CRB, but in many scenarios it closely follows it. The PoPi, MC-aML, SRP and SPR-PHAT methods are also biased; therefore, as for the aNLS method, their performances do not necessarily improve by improving the estimation conditions, e.g., by increasing the SNR, N_t or N_s. Another key observation for the single-source experiments is that the aNLS method seems to outperform the MC-aML method in most scenarios in terms of the MSE of the pitch estimates.

In the first two-source scenario^2, the DOAs of the sources were the same while the pitch spacing was varying. The NLS, aNLS, MC-ML and MC-aML methods outperform the PoPi and LCMV methods for pitch estimation for pitch spacings above ≈ 0.0155 in this scenario. Moreover, the proposed NLS and aNLS estimators clearly outperforms all other methods for DOA estimation. It is expected that the SRP, SRP-PHAT, and bMVDR methods fail in this scenario, as the broadband methods can not resolve sources with the same DOA. In the other two-source scenario, the two sources had the same pitch, while the DOA spacing between the sources was altered. Here, we observe that the proposed methods outperforms all other methods for pitch estimation for DOA spacings below ≈ -0.87. We note that the MC-ML and MC-aML methods fail in this scenario, since they only conduct a one-dimensional search. For DOA estimation, the NLS, aNLS and SRP methods yields the best performance for DOA spacings below ≈ -0.87.



(a)

(b)

Fig. E.3: MSE of (a) pitch and (b) DOA estimates obtained in different scenarios. In all scenarios, 500 Monte-Carlo simulations were conducted for each experimental setup to estimate the MSE of the parameter estimates.

5.2 Real-life Examples

We also conducted some qualitative experiments to evaluate the performance of the proposed methods on real-life signals. These experiments were conducted in a meeting room. The floor plan of the room and the measurement setup are illustrated in Fig. E.4,

[2] The subspace method is not considered in these scenarios as it is only suited for estimating the parameters of a single source.

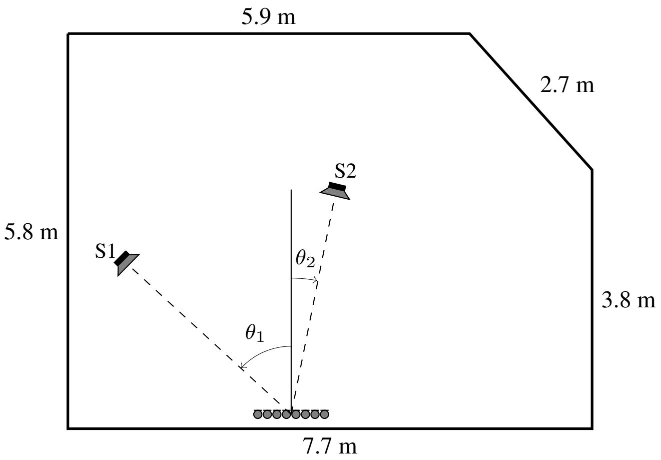


Fig. E.4: Floor plan of the meeting room used for the real-life experiments. The angle between the two speakers, 'S1' and 'S2', was $\theta_1 \approx 45°$ and $\theta_2 \approx -13°$, respectively.

while the height of the room was 2.64 m. In these simulations, the sampling frequency was $f_s = 44.1$ kHz, the speed of sound was assumed to be $c = 343.2$ m/s, the room reverberation time[3] at 1 kHz was $T_{60} \approx 0.53$ s, the array was uniform and linear with $d = 4$ cm and $N_s = 8$, we assume that there was no attenuation across the sensors such that $\mathbf{B} = \mathbf{I}$, and the desired signal was assumed to consist of $L = 8$ harmonics. The estimators used in these experiments were the same as in the previous simulations with synthetic data and they were set up similarly.

In the first of the real-life experiments, we played back an anechoic trumpet signal using the speaker 'S2'. The anechoic trumpet signal was generated by concatenating anechoic trumpet signal excerpts[4]. The played back trumpet signal was recorded using the ULA to obtain a multichannel trumpet signal with slight reverberation. From the recording, we estimated the pitch and the DOA of the trumpet signal using the estimators mentioned previously. In Fig. E.5, the estimates obtained from this experiment are depicted. We can see that all of the applied estimators for pitch estimation except the PoPi and LCMV methods seems to correctly estimate the pitch of the trumpet signal if we compare the estimates with the spectrogram. Regarding DOA estimation, we can see that all estimators obtain estimates relatively close to the true DOA except the bMVDR and LCMV methods which looks heavily biased. Note that the NLS and aNLS methods yield estimates close to the true parameter values even though recording contains reverberation and $\mathbf{B} \neq \mathbf{I}$ in practice. We then conducted a similar experiment where

[3] Here, the reverberation time is defined as time required for reflections of a direct sound to decay by 60 dB below the level of the direct sound.

[4] The excerpts were downloaded from `http://theremin.music.uiowa.edu/MIS.html`

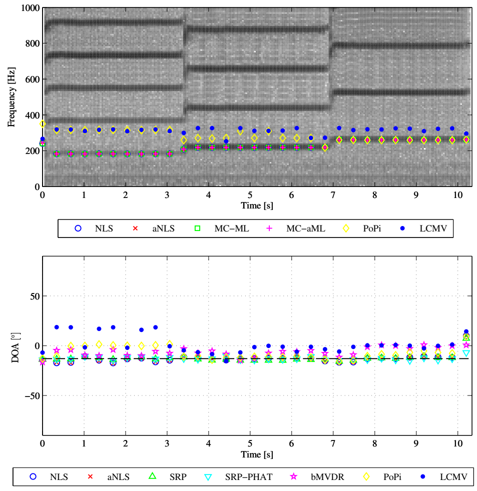


Fig. E.5: Estimation results from a real-life experiment with a single source; the source was a trumpet signal played back using ’S2’. The pitch and DOA estimates of the trumpet signal is depicted in the top and bottom plots, respectively.

we played back the same trumpet signal using ’S1’ and a speech signal using ’S2’, and the mixture was recorded using the array. The played back speech signal was a female speech excerpt taken from the Keele pitch database [48]. In this experiment, the speech signal was considered noise, so the objective was to estimate the DOA and the pitch of the trumpet signal. In Fig. E.6, the results from this experiment are shown. Again, we observe that the pitch estimators except for the PoPi and LCMV estimators seem to provide correct pitch estimates except at $\approx$ 7.3 s. For DOA estimation, it seems that the proposed methods outperform the other methods. The SRP-PHAT method provides heavily biased estimates, and the estimates of the bMVDR, PoPi, and LCMV methods seem erroneous. We note that the proposed methods provide good pitch and DOA estimates even though the noise is indeed not white Gaussian in this experiment. In summary, the proposed methods show comparable or better estimation performance than other state-of-the-art DOA and pitch estimators in our real-life experiments. Moreover, the results from these experiments indicate that the proposed methods are applicable on real-life signals, and that they are robust against reverberation as well as other noise types than white Gaussian noise. Note that the above observations based on our qualitative experiments may be different for, e.g., other sensor and source positions, and array structures, due to the complicated nature of reverberant signals.

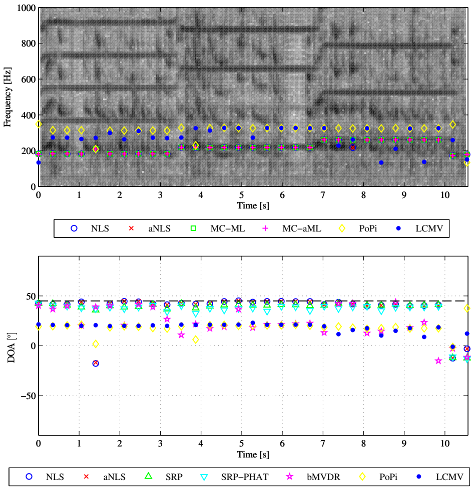


Fig. E.6: Estimation results from a real-life experiment with two sources; the sources were a trumpet signal and a speech signal played back using 'S1' and 'S2', respectively. The pitch and DOA estimates of the trumpet signal is depicted in the top and bottom plots, respectively.

6 Conclusion

In this paper, we have considered joint estimation of the DOA and the pitch of a harmonic source recorded using a ULA. First, we derived the exact and asymptotic Cramér-Rao bounds (CRBs) for the joint estimation problem. From the asymptotic bounds, it is clear that the DOA can be estimated more accurately by taking the harmonic structure into account compared to if we just estimated the DOA of, e.g., the fundamental tone. Moreover, these bounds reveal that the pitch can be estimated more accurately when multiple sensors are used. Then, we proposed two estimators for joint DOA and pitch estimation, namely the NLS and aNLS methods. The proposed estimators are maximum likelihood estimators when the noise is white Gaussian, the environment is anechoic, and the source of interest is in the far-field. We conducted numerous of simulations on synthetic data where the proposed methods and other state-of-the-art methods for DOA and pitch estimation were applied. The results show that the proposed methods attains the CRB with the aNLS being slightly biased. In general, the proposed methods outperform the other methods for both DOA and pitch estimation in terms of mean squared error. This is even the case in two-source scenarios where the noise is not white Gaussian only. The results obtained from the two-source scenarios also show that it is beneficial to estimate the DOAs and the pitches jointly when two sources are having the same DOA or pitch, since the methods estimating only one of these parameters may fail. Furthermore, we conducted experiments on real-life data. The results from these experiments indicate that the proposed methods has similar or better estimation performance than the other applied methods. Moreover, these experiments indicate that the the proposed methods are applicable on real-life signals, and that they are robust against reverberation and noise which is not white Gaussian.

References

[1] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[2] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[3] R. Kumaresan and D. W. Tufts, "Estimating the angles of arrival of multiple plane waves," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 19, no. 1, pp. 134–139, Jan. 1983.

[4] 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, Nov. 1991.

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

[6] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Education, Inc., 2005.

[7] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis - Principles, Algorithm, and Applications*. John Wiley & Sons, Inc., 2006.

[8] S. Makino, T. W. Lee, and H. Sawada, *Blind Speech Separation*. Springer, 2007.

[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] 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.

[11] 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, no. 5, pp. 1635–1644, Jul. 2007.

[12] V. Emiya, B. David, and R. Badeau, "A parametric method for pitch estimation of piano tones," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, Apr. 2007, pp. 249–252.

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

[14] 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.

[15] L. Du, T. Yardibi, J. Li, and P. Stoica, "Review of user parameter-free robust adaptive beamforming algorithms," *Digital Signal Processing*, vol. 19, no. 4, pp. 567–582, Jul. 2009.

[16] M. Viberg and P. Stoica, "A computationally efficient method for joint direction finding and frequency estimation in colored noise," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, vol. 2, Nov. 1998, pp. 1547–1551.

[17] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[18] ——, *Nonlinear Methods of Spectral Analysis*. Springer-Verlag, 1983, ch. Maximum-Likelihood Spectral Estimation.

[19] A. Jakobsson, S. L. Jr. Marple, and P. Stoica, "Computationally efficient two-dimensional Capon spectrum analysis," *IEEE Trans. Signal Process.*, vol. 48, no. 9, pp. 2651–2661, Sep. 2000.

[20] A. N. Lemma, A.-J. van der Veen, and E. F. Deprettere, "Analysis of joint angle-frequency estimation using ESPRIT," *IEEE Trans. Signal Process.*, vol. 51, no. 5, pp. 1264–1283, May 2003.

[21] T. Shu and X. Liu, "Robust and computationally efficient signal-dependent method for joint DOA and frequency estimation," *EURASIP J. on Advances in Signal Processing*, vol. 2008, no. 1, pp. 1–16, Apr. 2008.

[22] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Elsevier Signal Process.*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[23] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[24] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 4, pp. 823–831, Aug. 1985.

[25] G. C. Carter, "Coherence and time delay estimation," *Proc. IEEE*, vol. 75, no. 2, pp. 236–255, Feb. 1987.

[26] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, Apr. 1997, pp. 375–378.

[27] 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.

[28] 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.

[29] G. Liao, H. C. So, and P. C. Ching, "Joint time delay and frequency estimation of multiple sinusoids," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 5, May 2001, pp. 3121–3124.

[30] 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.

[31] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, "Joint DOA and multi-pitch estimation based on subspace techniques," *EURASIP J. on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–11, Jan. 2012.

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

[33] M. Képesi, L. Ottowitz, and T. Habib, "Joint position-pitch estimation for multiple speaker scenarios," May 2008, pp. 85–88.

[34] 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.

[35] 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.

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

[37] 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.*, Oct. 2005, pp. 283–286.

[38] M. G. Christensen, P. Vera-Candeas, S. D. Somasundaram, and A. Jakobsson, "Robust subspace-based fundamental frequency estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 2008, pp. 101–104.

[39] 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.

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

[41] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed. The John Hopkins University Press, 1996.

[42] 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, Aug. 1997.

[43] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 2012.

[44] 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.

[45] 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.

[46] L. Armani and M. Omologo, "Weighted autocorrelation-based f0 estimation for distant-talking interaction with a distributed microphone network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, May 2004, pp. 113–116.

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

[48] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. Eurospeech*, Sep. 1995, pp. 837–840.

Paper F

Non-Causal Time-Domain Filters for Single-Channel Noise Reduction

Jesper Rindom Jensen, Jacob Benesty, Mads Græsbøll Christensen and Søren Holdt Jensen

The paper has been published in the
IEEE Trans. on Audio, Speech and Language Processing, vol. 20(5), pp. 1526–1541, 2012.

The layout has been revised.

Abstract

In many existing time-domain filtering methods for noise reduction in, e.g., speech processing, the filters are causal. Such causal filters can be implemented directly in practice. However, it is possible to improve the performance of such noise reduction filtering methods in terms of both noise suppression and signal distortion by allowing the filters to be non-causal. Non-causal time-domain filters require knowledge of the future, and are therefore not directly implementable. If the observed signal is processed in blocks, however, the non-causal filters are implementable. In this paper, we propose such non-causal time-domain filters for noise reduction in speech applications. We also propose some performance measures that enable us to evaluate the performance of non-causal filters. Moreover, it is shown how some of the filters can be updated recursively. Using the recursive expressions, it is also shown that the output SNRs of the filters always increase as we increase the length of the filter when the desired signal is stationary. From both the theoretical and practical evaluations of the filters, it is clearly shown that the performance of time-domain filtering methods for noise reduction can be improved by introducing non-causality.

1 Introduction

Noise reduction is an important fundamental signal processing problem. In this paper, we consider generic noise reduction filters which are useful for enhancing any kind of desired signal. An example of a desired signal is speech which is commonly utilized in a multitude of applications such as telecommunications, teleconferencing, hearing-aids, and human-machine interfaces. In all these, the speech first has to be recorded using one or more microphones, and the speech will inevitably be corrupted by some degree of background noise. The noise could be, for example, other interfering speakers, fan noise, car noise, etc. Since the noise will reduce the speech quality and intelligibility, it will most likely have a detrimental impact on speech applications. In hearing-aids, for example, decreased speech quality can cause listener fatigue. It is therefore highly important to develop noise reduction methods to reduce the impact of the noise in various signal processing applications. Over the years, numerous noise reduction methods have been proposed. For an overview of speech related noise reduction methods, see, e.g., [1, 2] and the references therein. In general, we can divide these speech related noise reduction methods into three groups, i.e., spectral-subtractive algorithms [3], statistical-model-based algorithms [4–7], and subspace algorithms [8–11]. The references, [3–5, 8–10], refer to some of the pioneering work within these groups. Note that in the literature, noise reduction in speech applications is also termed speech enhancement.

Often, noise reduction methods rely on linear filtering. In such filtering methods, the noise reduction problem is formulated as a filter design problem. The goal of such filter design problems is to design a filter which attenuates the noise as much as possible while it only introduces an inconsiderable amount of distortion of the desired signal, e.g., speech. The filter can be derived directly in the time domain or in different transform domains. For example, it is possible to reduce the computational complexity by utilizing transform domain filters [12]. Two examples of transform domains are the Fourier [3, 9, 13, 14] and Karhunen-Loève [15, 16] domains. The filters can, though, be equivalently derived in all domains. In this paper, we consider time-domain filters only. Moreover, we restrict ourselves to the study of single-channel filters only.

Many existing time-domain filter designs for noise reduction are causal. In this paper, however, we propose novel non-causal filter designs, and we quantify the performance gain which can be obtained by exploiting non-causality. Note that we only consider the effects of introducing non-causality in the filter designs and not of introducing non-causality in the estimation of the signal and noise statistics since the statistics are assumed to be known exactly in most parts of the paper. The proposed filter designs are based on two different decompositions of the desired signal; three designs are based on an orthogonal decomposition [12], and one is based on a harmonic decomposition [17, 18]. The orthogonal decomposition based filters are suitable for enhancing any kind of desired signal since they are designed using the noise statistics, whereas the harmonic decomposition based filter is calculated from the statistics of the desired signal under the assumption that it is periodic. Periodicity or quasi-periodicity is a reasonable assumption for, e.g., short segments of voiced speech and musical instrument signals. The two decomposition approaches both have advantages and disadvantages as discussed in [19]. For example, the orthogonal decomposition based filters can be used for enhancing any kind of desired signal, however, they are sensible to non-stationary noise since it is difficult to estimate the noise statistics when the desired signal is present. The harmonic decomposition filter, on the other hand, is robust against non-stationary noise since it is based on the statistics of the desired signal, but it will cause distortion of the desired signal when the periodicity assumption does not hold exactly. It was shown in [19, 20] that the orthogonal and harmonic decomposition based filters are closely related, and that it is beneficial to use them jointly for speech enhancement.

In this paper, we generalize the mentioned decompositions such that they support the derivation of non-causal time-domain filters. Based on these generalized decompositions, we propose several performance measures suited for evaluation of non-causal filters. Moreover, we derive different non-causal orthogonal and harmonic decomposition based filters. Note that the causal filters proposed in [12, 17, 18, 21] can be seen as special cases of the proposed designs. For the two particular cases where the filter is causal and anti-causal, respectively, we derive expressions for recursive updates of the orthogonal decomposition based filters and the maximum output signal-to-noise ratio of these. From these recursive expressions, it can be shown that the maximum output SNR always increases if we increase the filter order when the desired signal is stationary. We quantify the performance gain that can be achieved by introducing non-causality in the filter design. To this end, we assume that the desired signal is periodic and, therefore, has a harmonic structure. Using this assumption, we can obtain exact closed-form expressions for the performance measures which we can use to precisely quantify the theoretical gains that can be achieved by using non-causal filters. Finally, we apply the non-causal filters to noise reduction of noisy speech signals to show the practical benefits of introducing non-causality in the filter design.

The rest of the paper is organized as follows. In Section 2, we introduce the signal model utilized in the paper, and we define the problem of designing non-causal time-domain filters for noise reduction. We then, in Section 3, describe the concept of linear non-causal filtering for noise reduction for two different signal decompositions. Based on the different decompositions, we propose several performance measures for non-causal noise reduction filters in Section 4. In Section 5, we propose new optimal non-causal noise reduction filters. We show, in Section 6, that some of the filters and their output signal-to-noise ratios can be updated recursively. In Section 7, we quantify the performance gain that can be obtained by introducing non-causality in the filter design. Finally, we conclude on the paper in Section 9.

2 Signal Model and Problem Statement

In this paper, we consider the benefits of introducing non-causality in optimal time-domain linear filters for noise reduction. The objective of such filters is to extract a zero-mean desired signal $x(n) \in \mathbb{R}$ from an observed signal $y(n) \in \mathbb{R}$ defined as

$$y(n) = x(n) + v(n) , \tag{F.1}$$

where $v(n) \in \mathbb{R}$ is additive noise and n denotes the discrete-time index. The observed signal $y(n)$ could, for example, be a microphone recording and the desired signal $x(n)$ could be clean speech. In the rest of the paper, we assume that the noise $v(n)$ is a zero-mean random process which is uncorrelated with the desired signal.

In some parts of the paper, we assume that the desired signal is quasi-periodic. This is a reasonable assumption for voiced speech segments. By assuming this specific signal structure, we can obtain closed-form expressions for certain performance measures related to optimal filters which are applied on the observed signal. Ultimately, the closed-form performance measures enable easy quantification of the performance gain which can be obtained by introducing non-causality in noise reduction filters. This will become clear from the later sections. When the desired signal is quasi-periodic, we can express it in terms of a harmonic model. The signal model in (F.1) then becomes

$$y(n) = \sum_{l=1}^{L} A_l \cos(l\omega_0 n + \phi_l) + v(n) , \tag{F.2}$$

where ω_0 is the fundamental frequency (aka. the pitch), L is the number of harmonics, A_l is the amplitude of the lth harmonic, and ϕ_l is the phase of the lth harmonic. In this paper, we consider the pitch, ω_0, and the model order, L, as known parameters. Numerous methods for estimation of these parameters exist [17, 18, 22–28]. Using Euler's formula, we can also write (F.2) as

$$y(n) = \sum_{l=1}^{L} \left(a_l e^{jl\omega_0 n} + a_l^* e^{-jl\omega_0 n} \right) + v(n) \ , \tag{F.3}$$

where $a_l = \frac{A_l}{2} e^{j\phi_l}$ is the complex amplitude of the lth harmonic, and $(\cdot)^*$ denotes the elementwise complex conjugate of a matrix/vector.

To make the notation simpler when deriving the optimal non-causal noise reduction filters, we stack consecutive samples of the observed signal $y(n)$ into a vector $\mathbf{y}(n_k) \in \mathbb{R}^{M\times 1}$ where $n_k = n + k$. The vector signal model is then given by

$$\mathbf{y}(n_k) = \mathbf{x}(n_k) + \mathbf{v}(n_k) \ , \tag{F.4}$$

where

$$\mathbf{y}(n_k) = \begin{bmatrix} y(n_k) & y(n_k - 1) & \cdots & y(n_k - M + 1) \end{bmatrix}^T \ , \tag{F.5}$$

with $(\cdot)^T$ denoting the matrix/vector transpose. Note that the definitions of the desired signal vector $\mathbf{x}(n_k)$ and the noise vector $\mathbf{v}(n_k)$ follow the definition of the observed signal vector $\mathbf{y}(n_k)$ in (F.5). Since the observed signal $x(n)$ and the noise $v(n)$ are uncorrelated by assumption, we can obtain a simple expression for the covariance matrix $\mathbf{R_y} \in \mathbb{R}^{M\times M}$ of $\mathbf{y}(n_k)$ as

$$\begin{aligned} \mathbf{R_y} &= \mathrm{E}[\mathbf{y}(n_k)\mathbf{y}^T(n_k)] \\ &= \mathbf{R_x} + \mathbf{R_v} \ , \end{aligned} \tag{F.6}$$

where E[$\cdot$] is the mathematical expectation operator, $\mathbf{R_x} = \mathrm{E}[\mathbf{x}(n_k)\mathbf{x}^T(n_k)]$ is the covariance matrix of $\mathbf{x}(n_k)$ and $\mathbf{R_v} = \mathrm{E}[\mathbf{v}(n_k)\mathbf{v}^T(n_k)]$ is the covariance matrix of $\mathbf{v}(n_k)$. When $x(n)$ is quasi-periodic, we can also model $\mathbf{R_x}$ as [29]

$$\begin{aligned} \mathbf{R_x} &\approx \mathbf{Z}_k(\omega_0)\mathbf{P}\mathbf{Z}_k^H(\omega_0) \\ &= \mathbf{Z}(\omega_0)\mathbf{P}\mathbf{Z}^H(\omega_0) \ , \end{aligned} \tag{F.7}$$

with $(\cdot)^H$ denoting the complex conjugate transpose operator, and

$$\mathbf{P} = \mathrm{diag}\left\{ \begin{bmatrix} |a_1|^2 & |a_1^*|^2 & \cdots & |a_L|^2 & |a_L^*|^2 \end{bmatrix} \right\} \ , \tag{F.8}$$

$$\mathbf{Z}_k(\omega_0) = \mathbf{Z}(\omega_0)\mathbf{S}(k) \ , \tag{F.9}$$

$$\mathbf{Z}(\omega_0) = \begin{bmatrix} \mathbf{z}(\omega_0) & \mathbf{z}^*(\omega_0) & \cdots & \mathbf{z}(L\omega_0) & \mathbf{z}^*(L\omega_0) \end{bmatrix} \ , \tag{F.10}$$

$$\mathbf{z}(l\omega_0) = \begin{bmatrix} 1 & e^{-jl\omega_0} & \cdots & e^{-jl\omega_0(M-1)} \end{bmatrix}^T \ , \tag{F.11}$$

$$\mathbf{S}(k) = \mathrm{diag}\left\{ \begin{bmatrix} e^{j\omega_0 k} & e^{-j\omega_0 k} & \cdots & e^{jL\omega_0 k} & e^{-jL\omega_0 k} \end{bmatrix} \right\} , \tag{F.12}$$

where diag$\{\cdot\}$ denotes the construction of a diagonal matrix from a vector.

The objective in traditional noise reduction methods is to find a "good" estimate of $x(n)$ or $\mathbf{x}(n)$ from the observed signal vector $\mathbf{y}(n)$. Within the field of speech enhancement research there is, in general, consensus on that "good" means the noise should be reduced as much as possible while the desired signal remains undistorted or nearly undistorted in the noise reduction process. In this paper, we consider another approach on noise reduction where we instead estimate $x(n)$ from $\mathbf{y}(n_k)$ where $k \in [0; M-1]$. That is, we introduce non-causality in the estimation procedure which, eventually, can increase the amount of noise reduction.

3 Noise Reduction Using Non-Causal Linear Filters

Filtering methods constitute a commonly used group of methods for noise reduction tasks such as speech enhancement. In the majority of such filtering methods, a finite impulse response (FIR) filter is applied on the observed signal vector. If the filter is allowed to be non-causal, we can, in general, write the noise reduction filtering operation as

$$\begin{aligned} \hat{x}_k(n) &= \sum_{m=-k}^{M-1-k} h_{m,k} y(n-m) \\ &= \mathbf{h}_k^T \mathbf{y}(n_k) , \end{aligned} \tag{F.13}$$

for $k \in [0; M-1]$ and where

$$\mathbf{h}_k = \begin{bmatrix} h_{-k,k} & h_{-k+1,k} & \cdots & h_{-k+M-1,k} \end{bmatrix}^T \tag{F.14}$$

and $\hat{x}_k(n)$ should be an estimate of $x(n)$. Traditionally, time-domain filters for noise reduction have been considered causal, i.e., they have been derived for $k = 0$ (see, e.g., [12] and the references therein). In this paper, however, we consider the general case where k can be any integer in the interval $[0; M-1]$. In practice, we can easily implement non-causal filters by doing block processing and allowing a small delay.

In the last couple of decades, several different causal filter designs have been proposed. The main difference between the designs is how the observed signal is decomposed. For the causal filter design problem, we have, for example, the classical, the orthogonal and the harmonic decompositions [12, 18]. Followingly, we redefine the orthogonal and harmonic decompositions by introducing non-causality.

3.1 Orthogonal Decomposition

Recently, it was proposed to design causal time-domain noise reduction filter based on an orthogonal decomposition of the desired signal [12, 21]. By using this approach, it is clear that some components of the signal vector $\mathbf{x}(n)$ actually act as interference when we estimate the desired signal $x(n)$. Here, we generalize this decomposition by introducing non-causality to enable the estimation of $x(n)$ from $\mathbf{x}(n_k)$. If we apply the orthogonal decomposition with respect to $x(n)$ on the signal vector $\mathbf{x}(n_k)$, we get

$$\begin{aligned}\mathbf{x}(n_k) &= x(n)\boldsymbol{\rho}_{\mathbf{x}x,k} + \mathbf{x}_{\mathrm{i}}(n_k)\\ &= \mathbf{x}_{\mathrm{d}}(n_k) + \mathbf{x}_{\mathrm{i}}(n_k)\ , \qquad \text{(F.15)}\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{\rho}_{\mathbf{x}x,k} &= \frac{\mathrm{E}[\mathbf{x}(n_k)x(n)]}{\mathrm{E}[x^2(n)]} \qquad \text{(F.16)}\\ &= \begin{bmatrix}\rho_x(k) & \rho_x(k-1) & \cdots & \rho_x(k-M+1)\end{bmatrix}^T\ ,\\ \rho_x(m) &= \frac{\mathrm{E}[x(n+m)x(n)]}{\mathrm{E}[x^2(n)]}\ . \qquad \text{(F.17)}\end{aligned}$$

Note that $\rho_x(m) = 1$ for $m = 0$. The elements in $\mathbf{x}_{\mathrm{d}}(n_k)$ in (F.15) are the parts of the elements in $\mathbf{x}(n_k)$ which are proportional to the desired signal $x(n)$ and $\mathbf{x}_{\mathrm{i}}(n_k)$ is the "interference" which is orthogonal to $\mathbf{x}_{\mathrm{d}}(n_k)$. If we insert (F.15) into (F.13), we get

$$\begin{aligned}\hat{x}_k(n) &= \mathbf{h}_k^T\mathbf{x}_{\mathrm{d}}(n_k) + \mathbf{h}_k^T\mathbf{x}_{\mathrm{i}}(n_k) + \mathbf{h}_k^T\mathbf{v}(n_k)\\ &= x_{\mathrm{fd},k}(n) + x_{\mathrm{ri},k}(n) + v_{\mathrm{rn},k}(n)\ , \qquad \text{(F.18)}\end{aligned}$$

where $x_{\mathrm{fd},k}(n) = \mathbf{h}_k^T\mathbf{x}_{\mathrm{d}}(n_k)$ is the filtered desired signal, $x_{\mathrm{ri},k}(n) = \mathbf{h}_k^T\mathbf{x}_{\mathrm{i}}(n_k)$ is the residual interference, and $v_{\mathrm{rn}}(n_k) = \mathbf{h}_k^T\mathbf{v}(n_k)$ is the residual noise. Since $\mathbf{x}_{\mathrm{d}}(n_k)$, $\mathbf{x}_{\mathrm{i}}(n_k)$ and $\mathbf{v}(n_k)$ are all orthogonal to each other, the variance of $\hat{x}_k(n)$ is given by

$$\sigma^2_{\hat{x}_k} = \sigma^2_{x_{\mathrm{fd},k}} + \sigma^2_{x_{\mathrm{ri},k}} + \sigma^2_{v_{\mathrm{rn},k}}\ , \qquad \text{(F.19)}$$

where

$$\begin{aligned}\sigma^2_{x_{\mathrm{fd},k}} &= \mathbf{h}_k^T\mathbf{R}_{\mathbf{x}_{\mathrm{d},k}}\mathbf{h}_k = \sigma_x^2\left(\mathbf{h}_k^T\boldsymbol{\rho}_{\mathbf{x}x,k}\right)^2\ , \qquad \text{(F.20)}\\ \sigma^2_{x_{\mathrm{ri},k}} &= \mathbf{h}_k^T\mathbf{R}_{\mathbf{x}_{\mathrm{i},k}}\mathbf{h}_k\ , \qquad \text{(F.21)}\\ \sigma^2_{v_{\mathrm{rn},k}} &= \mathbf{h}_k^T\mathbf{R}_{\mathbf{v}}\mathbf{h}_k\ , \qquad \text{(F.22)}\end{aligned}$$

$\mathbf{R}_{\mathbf{x}_{\mathrm{d},k}} = \sigma_x^2\boldsymbol{\rho}_{\mathbf{x}x,k}\boldsymbol{\rho}_{\mathbf{x}x,k}^T$ is the covariance matrix of $\mathbf{x}_{\mathrm{d}}(n_k)$, $\sigma_x^2 = \mathrm{E}[x^2(n)]$ is the variance of the desired signal, and $\mathbf{R}_{\mathbf{x}_{\mathrm{i},k}} = \mathrm{E}[\mathbf{x}_{\mathrm{i}}(n_k)\mathbf{x}_{\mathrm{i}}^T(n_k)]$ is the covariance matrix of the interference $\mathbf{x}_{\mathrm{i}}(n_k)$.

We can obtain the following error function for the orthogonal decomposition approach

$$e_k(n) = x_{\mathrm{fd},k}(n) + x_{\mathrm{ri},k}(n) + v_{\mathrm{rn},k}(n) - x(n)\ . \qquad \text{(F.23)}$$

Compared to the classical filtering approach, the orthogonal decomposition approach has an extra noise term, namely the residual interference $x_{\text{ri},k}(n)$ [12]. Moreover, the desired signal $x_{\text{fd},k}(n)$ is different from the desired signal in the classical filtering approach. The design goal is to minimize the effect of $x_{\text{ri},k}(n)$ and $v_{\text{rn},k}(n)$ while keeping the difference between $x_{\text{fd},k}(n)$ and $x(n)$ small. These goals can obviously be fulfilled by minimizing the error function in the mean square error (MSE) sense, possibly under some constraints.

3.2 Harmonic Decomposition

The harmonic decomposition approach to noise reduction filter design is a special case of the classical approach to linear filtering. In the harmonic decomposition, it is assumed that the desired signal is periodic and modeled by the harmonic model in (F.3). The harmonic model has previously been applied in numerous pitch estimation methods [18]. Many real-life signals such as audio and voiced speech are quasi-periodic which makes the harmonic decomposition approach useful in practice. By using the harmonic model, we can write the signal vector $\mathbf{x}(n_k)$ as

$$\mathbf{x}(n_k) = \mathbf{Z}(\omega_0)\mathbf{a}(n_k) = \mathbf{x}'_\text{d}(n_k) \ , \tag{F.24}$$

where

$$\begin{aligned}\mathbf{a}(n_k) = \Big[a_1 e^{j\omega_0(n+k)} \quad & a_1^* e^{-j\omega_0(n+k)} \quad \cdots \\ & a_L e^{jL\omega_0(n+k)} \quad a_L^* e^{-jL\omega_0(n+k)}\Big]^T \ .\end{aligned} \tag{F.25}$$

In this approach, there is no interference since the harmonic model enables us to use all information embedded in $\mathbf{x}(n_k)$ in the estimation of $x(n)$. Moreover, the desired signal $x(n)$ is the $(k+1)$th entry of the vector $\mathbf{Z}(\omega_0)\mathbf{a}(n_k)$, i.e., we can write it as

$$x(n) = \mathbf{z}_{\text{r},k}(\omega_0)\mathbf{a}(n_k) \ , \tag{F.26}$$

where $\mathbf{z}_{\text{r},k}(\omega_0)$ is the $(k+1)$th row of $\mathbf{Z}(\omega_0)$. An estimate of the desired signal $x(n)$ can be obtained by inserting (F.24) into (F.13). This yields

$$\begin{aligned}\hat{x}_k(n) &= \mathbf{h}_k^T \mathbf{x}'_\text{d}(n_k) + \mathbf{h}_k^T \mathbf{v}(n_k) \\ &= x'_{\text{fd},k}(n) + v_{\text{rn},k}(n) \ .\end{aligned} \tag{F.27}$$

where $x'_{\text{fd},k}(n) = \mathbf{h}_k^T \mathbf{x}'_\text{d}(n_k)$ is the filtered desired periodic signal. The orthogonality between the desired signal and the noise enables us to write the variance of $\hat{x}_k(n)$ as

$$\sigma^2_{\hat{x}_k} = \sigma^2_{x'_{\text{fd},k}} + \sigma^2_{v_{\text{rn},k}} \ . \tag{F.28}$$

Since the desired signal is assumed periodic in this approach, the variance of the filtered signal can also be written as

$$\begin{aligned}\sigma^2_{x'_{\mathrm{fd},k}} &= \mathbf{h}_k^T \mathbf{R}_{\mathbf{x}'_{\mathrm{d},k}} \mathbf{h}_k^T \\ &\approx \mathbf{h}_k^T \mathbf{Z}(\omega_0)\mathbf{P}\mathbf{Z}^H(\omega_0)\mathbf{h}_k \ ,\end{aligned} \tag{F.29}$$

where

$$\mathbf{R}_{\mathbf{x}'_{\mathrm{d},k}} = \mathrm{E}[\mathbf{x}'_{\mathrm{d}}(n_k)\mathbf{x}'^T_{\mathrm{d}}(n_k)] \approx \mathbf{Z}(\omega_0)\mathbf{P}\mathbf{Z}^H(\omega_0) \tag{F.30}$$

is the covariance matrix of $\mathbf{x}'_{\mathrm{d}}(n_k)$. We can obtain the following error function for harmonic decomposition approach

$$e_k(n) = x'_{\mathrm{fd},k}(n) + v_{\mathrm{rn},k}(n) - \mathbf{z}_{\mathrm{r},k}(\omega_0)\mathbf{a}(n_k) \ . \tag{F.31}$$

A filter which minimizes the effect of the noise, $v_{\mathrm{rn},k}(n)$, and the difference between $x'_{\mathrm{fd},k}(n)$ and $\mathbf{z}_{\mathrm{r},k}(\omega_0)\mathbf{a}(n_k)$ can then be designed by minimizing (F.31), possibly under some constraints.

4 Performance Measures

Recently, several performance measures for noise reduction tasks were proposed in [12, 21]. In this section, we generalize these performance measures to encompass non-causal filters. Note that while the measures are here derived for the orthogonal decomposition approach, they can easily be derived for the harmonic decomposition approach by replacing $\sigma^2_{x_{\mathrm{fd},k}}$ by $\sigma^2_{x'_{\mathrm{fd},k}}$ and $\sigma^2_{x_{\mathrm{ri},k}}$ by 0.

4.1 Noise Reduction

A common measure of noise reduction is the signal-to-noise ratio (SNR). Here, we consider two SNRs, i.e., the input SNR (iSNR) and the output SNR (oSNR). The iSNR is the SNR of the observed signal before filtering

$$\mathrm{iSNR} = \frac{\sigma_x^2}{\sigma_v^2} \ , \tag{F.32}$$

with $\sigma_v^2 = \mathrm{E}[v^2(n)]$ being the variance of the noise. The oSNR is defined as the SNR after filtering. When using the orthogonal decomposition, it is therefore given by

$$\mathrm{oSNR}(\mathbf{h}_k) = \frac{\sigma^2_{x_{\mathrm{fd},k}}}{\sigma^2_{x_{\mathrm{ri},k}} + \sigma^2_{v_{\mathrm{rn},k}}} \ . \tag{F.33}$$

Another measure is the noise reduction factor $\xi_{\text{nr}}(\mathbf{h}_k)$. This measure is defined as the ratio between the noise before and after noise reduction. That is, we can write the factor as

$$\xi_{\text{nr}}(\mathbf{h}_k) = \frac{\sigma_v^2}{\sigma_{x_{\text{ri}},k}^2 + \sigma_{v_{\text{rn}},k}^2} \ . \tag{F.34}$$

The noise reduction factor is expected to be larger than or equal to 1.

4.2 Signal Distortion

In many noise reduction methods, the desired signal is distorted in the process of noise reduction. One measure which quantifies this distortion is the desired signal reduction factor. This factor is defined as the ratio between the variances of the desired signal before and after filtering, respectively. The measure can also be written as

$$\xi_{\text{dsr}}(\mathbf{h}_k) = \frac{\sigma_x^2}{\sigma_{x_{\text{fd}},k}^2} \ . \tag{F.35}$$

If there is no distortion, the desired signal reduction factor will be 1. Otherwise, it will be different from 1. According to (F.20), this implies that we must require that

$$\mathbf{h}_k^T \boldsymbol{\rho}_{\mathbf{x}x,k} = 1 \ , \tag{F.36}$$

if a filter should be distortionless. This knowledge can, for example, be applied in the filter design by using it as a constraint.

When the desired signal is periodic, we can also consider the harmonic distortion incurred by the filter. The harmonic distortion measure was proposed in [19]. This measure is defined as the sum of the absolute differences between the powers of the sinusoids before and after noise reduction, i.e.,

$$\begin{aligned}\xi_{\text{hd}}(\mathbf{h}_k) &= 2\sum_{l}^{L} |P_l - P_{\text{f},k,l}| \\ &= 2\sum_{l}^{L} P_l |1 - \mathbf{h}_k^T \mathbf{z}(l\omega_0)\mathbf{z}^H(l\omega_0)\mathbf{h}_k| \ ,\end{aligned} \tag{F.37}$$

where $P_{\text{f},k,l}$ is the power of the lth harmonic after filtering using $\mathbf{h}_k$. The harmonic distortion measure will be zero when the harmonics are not distorted. Otherwise it will be larger than zero. Note that the harmonics might be distorted even though (F.36) is fulfilled.

5 Optimal Non-Causal Filters for Noise Reduction

In this section, we rederive some optimal orthogonal and harmonic decomposition based noise reduction filters to obtain non-causal filters. The corresponding causal filters were derived in [12, 17, 18, 21]. Note that all filters derived here except the harmonic decomposition based linearly constrained minimum variance (HDLCMV) filter are based on the orthogonal decomposition.

5.1 Maximum SNR

The maximum SNR filter, $\mathbf{h}_{\text{max},k}$, is a filter which maximizes the output SNR with respect to the estimation of $x(n)$. The output SNR is defined in (F.33). If we insert (F.20)–(F.22) into (F.33), we can also write the output SNR for an orthogonal decomposition based filter as

$$\text{oSNR}(\mathbf{h}_k) = \frac{\mathbf{h}_k^T \mathbf{R}_{\mathbf{x}_{\text{d},k}} \mathbf{h}_k}{\mathbf{h}_k^T \mathbf{R}_{\text{in},k} \mathbf{h}_k} , \tag{F.38}$$

where

$$\mathbf{R}_{\text{in},k} = \mathbf{R}_{\mathbf{x}_{\text{i},k}} + \mathbf{R}_{\mathbf{v}} \tag{F.39}$$

is the covariance matrix of the interference-plus-noise. The expression in (F.38) can also be recognized as a generalized Rayleigh quotient [30]. This quotient is maximized when the filter, $\mathbf{h}_k$, equals the eigenvector $\mathbf{u}_{\text{max},k}$ corresponding to the largest eigenvalue, $\lambda_{\text{max},k}$, of $\mathbf{R}_{\text{in},k}^{-1}\mathbf{R}_{\mathbf{x}_{\text{d},k}}$. Clearly, $\mathbf{R}_{\text{in},k}^{-1}\mathbf{R}_{\mathbf{x}_{\text{d},k}}$ is rank one, i.e.,

$$\begin{aligned} \lambda_{\text{max},k} &= \text{Tr}\left(\mathbf{R}_{\text{in},k}^{-1}\mathbf{R}_{\mathbf{x}_{\text{d},k}}\right) \\ &= \sigma_x^2 \boldsymbol{\rho}_{\mathbf{x}x,k}^T \mathbf{R}_{\text{in},k}^{-1} \boldsymbol{\rho}_{\mathbf{x}x,k} \\ &= \text{oSNR}(\mathbf{h}_{\text{max},k}) , \end{aligned} \tag{F.40}$$

with $\text{Tr}(\cdot)$ denoting the trace operator. An important observation from the above expression is that, in general, $\lambda_{\text{max},p} \neq \lambda_{\text{max},q}$ for $p \neq q$. That is, the output SNR may be different for different ks which means that we may be able to improve the oSNR by introducing non-causality in the filter design.

From (F.40), we can readily see that $\mathbf{u}_{\text{max},k}$ and thus $\mathbf{h}_{\text{max},k}$ are given by

$$\mathbf{h}_{\text{max},k} = \alpha_k \mathbf{R}_{\text{in},k}^{-1} \boldsymbol{\rho}_{\mathbf{x}x,k} , \tag{F.41}$$

where $\alpha_k \neq 0$ is some arbitrary scaling factor. As it will become clear soon, the only difference between the orthogonal decomposition based filters described in this paper, is the scaling factor α_k.

5.2 Wiener

In the orthogonal decomposition based Wiener (ODW) filter design, the filter is designed by minimizing the MSE. The MSE criterion, $J(\mathbf{h}_k)$, can be written as

$$\begin{aligned} J(\mathbf{h}_k) &= \mathrm{E}[e_k^2(n)] \\ &= \sigma_x^2 \left(\mathbf{h}_k^T \boldsymbol{\rho}_{\mathbf{x}x,k} - 1\right)^2 + \mathbf{h}_k^T \mathbf{R}_{\text{in},k} \mathbf{h}_k \ . \end{aligned} \tag{F.42}$$

The minimizer of $J(\mathbf{h}_k)$ is found by differentiating with respect to $\mathbf{h}_k$ and equating with zero. If we do this, we get the following expression for the non-causal orthogonal decomposition based Wiener filter

$$\mathbf{h}_{\text{ODW},k} = \sigma_x^2 \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x,k} \ . \tag{F.43}$$

If we note that $\mathbf{R}_{\mathbf{y}}$ can also be written as

$$\mathbf{R}_{\mathbf{y}} = \sigma_x^2 \boldsymbol{\rho}_{\mathbf{x}x,k} \boldsymbol{\rho}_{\mathbf{x}x,k}^T + \mathbf{R}_{\text{in},k} \ , \tag{F.44}$$

and if we apply the matrix inversion lemma on $\mathbf{R}_{\mathbf{y}}^{-1}$, we can obtain another expression,

$$\mathbf{h}_{\text{ODW},k} = \frac{\sigma_x^2}{1 + \lambda_{\text{max},k}} \mathbf{R}_{\text{in},k}^{-1} \boldsymbol{\rho}_{\mathbf{x}x,k} \ , \tag{F.45}$$

for the Wiener filter. It appears from this expression that the ODW filter indeed maximizes the output SNR since it is just a scaled version of the maximum SNR filter where the scaling factor, $\alpha_{\text{W},k}$, is given by

$$\alpha_{\text{ODW},k} = \frac{\sigma_x^2}{1 + \lambda_{\text{max},k}} \ . \tag{F.46}$$

The output SNR of the non-causal ODW filter is therefore given by

$$\text{oSNR}(\mathbf{h}_{\text{ODW},k}) = \text{oSNR}(\mathbf{h}_{\text{max},k}) \ . \tag{F.47}$$

5.3 Minimum Variance Distortionless Response

The minimum variance distortionless response (MVDR) filter (aka. the Capon filter) was proposed by Capon in the context of spatial filtering [31, 32]. Here, the MVDR filter, or orthogonal decomposition based MVDR (ODMVDR) filter as we term it, is used for temporal filtering, and it is designed on basis of the orthogonal decomposition. The ODMVDR filter is designed such that it minimizes the variances of both the residual interference, $x_{\text{ri},k}(n)$, and the residual noise, $v_{\text{rn},k}(n)$. Moreover, the ODMVDR filter is designed to be distortionless with respect to the desired signal. Such a filter design can be obtained by solving the following quadratic minimization problem

$$\min_{\mathbf{h}_k} \mathbf{h}_k^T \mathbf{R}_{\text{in},k} \mathbf{h}_k \quad \text{s.t.} \quad \mathbf{h}_k^T \boldsymbol{\rho}_{\mathbf{x}x,k} = 1 \ . \tag{F.48}$$

The well-known solution of this optimization problem is given by

$$\mathbf{h}_{\text{ODMVDR},k} = \frac{\mathbf{R}_{\text{in},k}^{-1}\boldsymbol{\rho}_{\mathbf{x}x,k}}{\boldsymbol{\rho}_{\mathbf{x}x,k}^{T}\mathbf{R}_{\text{in},k}^{-1}\boldsymbol{\rho}_{\mathbf{x}x,k}} \; . \tag{F.49}$$

It turns out that the ODMVDR filter can be equivalently expressed as

$$\mathbf{h}_{\text{ODMVDR},k} = \frac{\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\rho}_{\mathbf{x}x,k}}{\boldsymbol{\rho}_{\mathbf{x}x,k}^{T}\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\rho}_{\mathbf{x}x,k}} \; . \tag{F.50}$$

Another important expression for the ODMVDR filter is given by

$$\begin{aligned}\mathbf{h}_{\text{ODMVDR},k} &= \frac{\sigma_x^2}{\lambda_{\max,k}}\mathbf{R}_{\text{in},k}^{-1}\boldsymbol{\rho}_{\mathbf{x}x,k} \\ &= \frac{1+\lambda_{\max,k}}{\lambda_{\max,k}}\mathbf{h}_{\text{W},k} \; ,\end{aligned} \tag{F.51}$$

from which it is clear that the ODMVDR filter maximizes the output SNR. The scaling factor, $\alpha_{\text{ODMVDR},k}$, is given by

$$\alpha_{\text{ODMVDR},k} = \frac{\sigma_x^2}{\lambda_{\max,k}} \; . \tag{F.52}$$

That is, by using the $\mathbf{h}_{\text{ODMVDR},k}$, we can have maximum output SNR while not distorting the desired signal, $x(n)$. Since the ODMVDR filter is just a scaled version of the maximum SNR filter, its output SNR is given by

$$\text{oSNR}(\mathbf{h}_{\text{ODMVDR},k}) = \text{oSNR}(\mathbf{h}_{\max,k}) \; . \tag{F.53}$$

5.4 Harmonic LCMV

The last filter described in this section, is the HDLCMV filter. This filter design is inspired by the LCMV beamformer (aka. the Frost beamformer) proposed by Frost in the context of spatial filtering [33]. Here, we derive a non-causal HDLCMV filter for temporal filtering. The HDLCMV filter is designed to extract periodic signals modeled by (F.2), i.e., it is suited for extraction of signals such as voiced speech and musical instruments. The causal version of the HDLCMV filter was proposed in [17].

Since the non-causal HDLCMV filter is based on the harmonic decomposition, it utilizes all the information in $\mathbf{x}(n_k)$ to estimate $x(n)$. In the harmonic decomposition, there is no interference term as opposed to in the orthogonal decomposition where we have $\mathbf{x}_{\text{i}}(n_k)$. Therefore, in the harmonic decomposition based filter design, we only have to minimize the residual noise power, $\sigma_{v_{\text{rn},k}}^2$, without distorting the signal too much. The HDLCMV filter, in particular, is designed to minimize $\sigma^2_{v_{\text{rn}},k}$ without distorting the harmonics of the desired periodic signal, $x(n)$. Such a filter can be obtained by solving the following optimization problem

$$\min_{\mathbf{h}_k} \mathbf{h}_k^T \mathbf{R}_{\mathbf{v}} \mathbf{h}_k \quad \text{s.t.} \quad \mathbf{Z}_k^H \mathbf{h}_k = \mathbf{1} \Leftrightarrow \mathbf{Z}^H \mathbf{h}_k = \mathbf{z}_{\text{r},k}^H \ , \tag{F.54}$$

where $\mathbf{1} = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^T$. It can readily be verified that the constraint in (F.54) makes the filter distortionless with respect to both the desired signal reduction factor and the harmonic distortion measure, respectively, by applying the covariance matrix model in (F.35) and (F.37).

The well-known solution to the multiple constrained quadratic optimization problem in (F.54) is given by

$$\mathbf{h}_{\text{HDLCMV},k} = \mathbf{R}_{\mathbf{v}}^{-1} \mathbf{Z} \left(\mathbf{Z}^H \mathbf{R}_{\mathbf{v}}^{-1} \mathbf{Z} \right)^{-1} \mathbf{z}_{\text{r},k}^H \ . \tag{F.55}$$

In [19], it was shown that we can replace $\mathbf{R}_{\mathbf{v}}$ by $\mathbf{R}_{\mathbf{y}}$ in the above expression without changing the filter response. If we do this, we get the following equivalent expression for the HDLCMV filter

$$\mathbf{h}_{\text{HDLCMV},k} = \mathbf{R}_{\mathbf{y}}^{-1} \mathbf{Z} \left(\mathbf{Z}^H \mathbf{R}_{\mathbf{y}}^{-1} \mathbf{Z} \right)^{-1} \mathbf{z}_{\text{r},k}^H \ . \tag{F.56}$$

This expression is interesting since we can find the optimal HDLCMV filter without knowing the noise statistics which is often a requirement in noise reduction methods. On the other hand, we need to know the pitch, ω_0, and the number of harmonics, L, of the desired signal, $x(n)$. When the HDLCMV filter is applied to a noise corrupted periodic signal, the output SNR can be found by replacing $\sigma^2_{x_{\text{fd}},k}$ by $\sigma^2_{x'_{\text{fd}},k}$ and $\sigma^2_{x_{\text{ri}},k}$ by 0 in (F.33). If we do this, we get the following expression

$$\begin{aligned} \text{oSNR}(\mathbf{h}_{\text{HDLCMV},k}) &\approx \frac{\mathbf{h}_{\text{HDLCMV},k}^T \mathbf{Z} \mathbf{P} \mathbf{Z}^H \mathbf{h}_{\text{HDLCMV},k}}{\mathbf{h}_{\text{HDLCMV},k}^T \mathbf{R}_{\mathbf{v}} \mathbf{h}_{\text{HDLCMV},k}} \\ &= \frac{\sigma_x^2}{\mathbf{z}_{\text{r},k} \left(\mathbf{Z}^H \mathbf{R}_{\mathbf{v}}^{-1} \mathbf{Z} \right)^{-1} \mathbf{z}_{\text{r},k}^H} \ . \end{aligned} \tag{F.57}$$

6 Recursive Filter Updates and the Maximum Output SNR

We now show how the ODW and ODMVDR filters presented in the previous section can be updated recursively. As a by product of this result, we also show how the maximum output SNR can be updated recursively which, eventually, proofs that the maximum output SNR always increases when the filter order M is increased. Here, we only provide the recursion for $k = 0$, but it can be generalized for all ks. Note that the derived recursive expressions also holds for $k = M - 1$ due to prediction symmetry.

From (F.43) and (F.50) it is clear that the ODW and ODMVDR filters both depend on $\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\rho}_{\mathbf{x}x}$ for $k = 0$ where $\boldsymbol{\rho}_{\mathbf{x}x} = \boldsymbol{\rho}_{\mathbf{x}x,0}$. Therefore, to find recursive filter and output SNR expressions, we derive a recursive expression for $\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\rho}_{\mathbf{x}x}$. To simplify the derivations of the recursion, we introduce a slightly different notation. First, we define the length m observed signal vector as

$$\begin{aligned}\mathbf{y}_m(n) &= \begin{bmatrix} y(n) & y(n-1) & \cdots & y(n-m+1)\end{bmatrix}^T \\ &= \begin{bmatrix} \mathbf{y}_{m-1}^T(n) & y(n-m+1)\end{bmatrix}^T \ .\end{aligned} \tag{F.58}$$

Using the above expression, we can write the covariance matrix of the observed signal as

$$\begin{aligned}\mathbf{R}_m &= \mathrm{E}\left[\mathbf{y}_m(n)\mathbf{y}_m^T(n)\right] \\ &= \begin{bmatrix} \mathbf{R}_{m-1} & \mathbf{r}_{\mathrm{b},m-1} \\ \mathbf{r}_{\mathrm{b},m-1}^T & r(0)\end{bmatrix} \ ,\end{aligned} \tag{F.59}$$

where

$$\mathbf{r}_{b,m-1} = \begin{bmatrix} r(m-1) & r(m-2) & \cdots & r(1)\end{bmatrix}^T \ , \tag{F.60}$$

$$r(i) = \mathrm{E}[y(n)y(n-i)] \ , \quad i = 0, 1, \ldots, m-1 \ . \tag{F.61}$$

Using the new notation, we can write the Wiener-Hopf equations as

$$\mathbf{R}_m\mathbf{g}_m = \mathbf{p}_m \ , \tag{F.62}$$

where

$$\begin{aligned}\mathbf{p}_m &= \boldsymbol{\rho}_{\mathbf{x}x} \\ &= \begin{bmatrix} p(0) & p(1) & \cdots & p(m-1)\end{bmatrix}^T \\ &= \begin{bmatrix} \mathbf{p}_{m-1}^T & p(m-1)\end{bmatrix} \ .\end{aligned} \tag{F.63}$$

We know from backward linear prediction theory that

$$\mathbf{R}_{m-1}\mathbf{b}_{m-1} = \mathbf{r}_{\mathrm{b},m-1} \ , \tag{F.64}$$

where $\mathbf{b}_{m-1}$ is the length $(m-1)$ optimal linear backward predictor. Moreover, we know that

$$\mathbf{R}_m \begin{bmatrix} -\mathbf{b}_{m-1} \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ E_{m-1}\end{bmatrix} \ , \tag{F.65}$$

with E_{m-1} being the prediction error energy defined as

$$E_{m-1} = r(0) - \mathbf{r}_{\mathrm{b},m-1}^T \mathbf{b}_{m-1} \; . \tag{F.66}$$

Consider now the following expression

$$\begin{aligned} \mathbf{R}_m \begin{bmatrix} \mathbf{g}_{m-1} \\ 0 \end{bmatrix} &= \begin{bmatrix} \mathbf{R}_{m-1} & \mathbf{r}_{\mathrm{b},m-1} \\ \mathbf{r}_{\mathrm{b},m-1}^T & r(0) \end{bmatrix} \begin{bmatrix} \mathbf{g}_{m-1} \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{p}_{m-1} \\ \mathbf{r}_{\mathrm{b},m-1}^T \mathbf{g}_{m-1} \end{bmatrix} \; . \end{aligned} \tag{F.67}$$

If we use (F.64) on in the above expression we immediately see that

$$\mathbf{r}_{\mathrm{b},m-1} \mathbf{g}_{m-1} = \mathbf{b}_{m-1}^T \mathbf{p}_{m-1} \; . \tag{F.68}$$

Then we subtract (F.67) from (F.62) which yields

$$\mathbf{R}_m \left(\mathbf{g}_m - \begin{bmatrix} \mathbf{g}_{m-1} \\ 0 \end{bmatrix} \right) = \begin{bmatrix} \mathbf{0} \\ \gamma_{m-1} \end{bmatrix} \; , \tag{F.69}$$

where

$$\gamma_{m-1} = p(m-1) - \mathbf{b}_{m-1}^T \mathbf{p}_{m-1} \; . \tag{F.70}$$

If we multiply both sides of (F.65) with $\frac{\gamma_{m-1}}{E_{m-1}}$ and compare it to (F.69), we can obtain that

$$\mathbf{g}_m = \begin{bmatrix} \mathbf{g}_{m-1} \\ 0 \end{bmatrix} - \frac{\gamma_{m-1}}{E_{m-1}} \begin{bmatrix} \mathbf{b}_{m-1} \\ -1 \end{bmatrix} \; . \tag{F.71}$$

That is, if we use the above expression in connection with the Levinson-Durbin algorithm, we can calculate $\mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x}$ recursively. The resulting algorithm is depicted in Table F.1. Note that in Table F.1, the matrix $\mathbf{J}_m \in \mathbb{R}^{m \times m}$ is defined as

$$\mathbf{J}_m = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \iddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix} \; , \tag{F.72}$$

and κ_m can be interpreted as the reflection coefficient.

We can now use the algorithm in Table F.1 to recursively calculate the orthogonal decomposition based Wiener and MVDR filters for $k = 0$ using the definitions in (F.43)

Table F.1: Efficient and recursive computation of $\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\rho}_{\mathbf{x}x}$.

```
$E_0 = r(0)$
**for** $m = 1, \ldots, M$
    $\gamma_{m-1} = p(m-1) - \mathbf{b}_{m-1}^T \mathbf{p}_{m-1}$
    $\kappa_m = \frac{1}{E_{m-1}} \left[ r(m) - \mathbf{b}_{m-1}^T \mathbf{J}_{m-1} \mathbf{r}_{\mathrm{b},m-1} \right]$
    $\mathbf{g}_m = \begin{bmatrix} \mathbf{g}_{m-1} \\ 0 \end{bmatrix} - \frac{\gamma_{m-1}}{\kappa_{m-1}} \begin{bmatrix} \mathbf{b}_{m-1} \\ -1 \end{bmatrix}$
    $\mathbf{b}_m = \begin{bmatrix} 0 \\ \mathbf{b}_{m-1} \end{bmatrix} - \kappa_m \mathbf{J}_m \begin{bmatrix} \mathbf{b}_{m-1} \\ -1 \end{bmatrix}$
    $E_m = E_{m-1}(1 - \kappa_m^2)$
**end**
```

and (F.51), respectively. By doing this, we get the following recursive expressions for the filters

$$\begin{aligned} \mathbf{h}_{\mathrm{ODW},0,m} &= \sigma_x^2 \mathbf{g}_m \\ &= \sigma_x^2 \begin{bmatrix} \mathbf{g}_{m-1} \\ 0 \end{bmatrix} - \frac{\gamma_{m-1}}{\kappa_{m-1}} \begin{bmatrix} \mathbf{b}_{m-1} \\ -1 \end{bmatrix}, \qquad \text{(F.73)} \\ \mathbf{h}_{\mathrm{ODMVDR},0,m} &= \frac{\mathbf{g}_m}{\mathbf{p}_m^T \mathbf{g}_m} \\ &= \frac{\begin{bmatrix} \mathbf{g}_{m-1} \\ 0 \end{bmatrix} - \frac{\gamma_{m-1}}{\kappa_{m-1}} \begin{bmatrix} \mathbf{b}_{m-1} \\ -1 \end{bmatrix}}{\mathbf{p}_{m-1}^T \mathbf{h}_{m-1} + \frac{\gamma_{m-1}^2}{E_{m-1}}}, \qquad \text{(F.74)} \end{aligned}$$

where the third subscript on the filters denotes the filter order. By calculating the Wiener and MVDR filters using the recursive procedure in Table F.1, we can reduce the computational complexity significantly compared to when the filters are calculated directly using (F.43) and (F.50), respectively [34]. Similar recursive filter expressions can be found for the non-causal filters where k is between 0 and $M-1$.

Moreover, we can use the recursive algorithm developed in this section, to find a recursive expression for the maximum output SNR when using orthogonal decomposition based filters. Again, we only derive the recursive expression for $k = 0$ (and thereby also for $k = M-1$), but it can be generalized to different ks. First, we have to rewrite the expression for the maximum output SNR. It can be seen that the covariance matrix, $\mathbf{R}_{\mathbf{x}_{\mathrm{i},k}}$, of the interference vector, $\mathbf{x}_i(n_k)$, is given by

$$\mathbf{R}_{\mathbf{x}_{\mathrm{i},k}} = \mathbf{R}_{\mathbf{x}} - \sigma_x^2 \boldsymbol{\rho}_{\mathbf{x}x,k} \boldsymbol{\rho}_{\mathbf{x}x,k}^T . \qquad \text{(F.75)}$$

If we then insert (F.75) into (F.39) which is then inserted into (F.40), and if we use the matrix inversion lemma, we can show that

$$\begin{aligned}\text{oSNR}_{\text{max},k} &= \sigma_x^2 \boldsymbol{\rho}_{\mathbf{x}x,k}^T \left(\mathbf{R}_\mathbf{y} - \sigma_x^2 \boldsymbol{\rho}_{\mathbf{x}x,k} \boldsymbol{\rho}_{\mathbf{x}x,k}^T\right)^{-1} \boldsymbol{\rho}_{\mathbf{x}x,k} \\ &= \frac{1}{\left(\sigma_x^2 \boldsymbol{\rho}_{\mathbf{x}x,k}^T \mathbf{R}_\mathbf{y}^{-1} \boldsymbol{\rho}_{\mathbf{x}x,k}\right)^{-1} - 1} .\end{aligned} \tag{F.76}$$

We now consider the case where $k = 0$. In this case, we can use the recursive expressions in Table F.1 to write

$$\begin{aligned}\boldsymbol{\rho}_{\mathbf{x}x}^T \mathbf{R}_\mathbf{y}^{-1} \boldsymbol{\rho}_{\mathbf{x}x} &= \mathbf{p}_m^T \mathbf{g}_m \\ &= \mathbf{p}_m^T \left(\begin{bmatrix} \mathbf{g}_{m-1} \\ 0 \end{bmatrix} - \frac{\gamma_{m-1}}{\kappa_{m-1}} \begin{bmatrix} \mathbf{b}_{m-1} \\ -1 \end{bmatrix} \right) \\ &= \mathbf{p}_{m-1}^T \mathbf{g}_{m-1} + \frac{\gamma_{m-1}^2}{E_{m-1}} .\end{aligned} \tag{F.77}$$

If we substitute (F.77) back into (F.76) we get

$$\text{oSNR}_{\text{max},0,m} = \frac{1}{\left[\sigma_x^2 \left(\mathbf{p}_{m-1}^T \mathbf{g}_{m-1} + \frac{\gamma_{m-1}^2}{E_{m-1}}\right)\right]^{-1} - 1} . \tag{F.78}$$

From the above expression, we can readily see that the output SNR will always increase as we increase m when the desired signal is stationary.

7 Study of Output SNR and Distortion

In this section, we investigate the performance of all the non-causal filters proposed in this paper when the desired signal is periodic. The assumption of periodicity enables us to exactly quantify the gains which can be obtained by introducing non-causality in the filters since we can then model the requisite statistics with closed-form expressions. First, we conduct a study where we measure the performance of all the non-causal filters as a function of k. Then, we investigate the asymptotic behavior of the maximum output SNR for different ks.

7.1 Filter Performances for Small M

We now investigate the performance of the non-causal ODW, ODMVDR, and HDL-CMV filters in terms of output SNR and harmonic distortion when the filters are applied on periodic signals. First, we derive closed-form expressions for the performance measures under the assumption of periodicity. When the desired signal is periodic, we know that

$$\begin{aligned}\boldsymbol{\rho}_{\mathbf{x}x,k} &= \frac{\mathbf{R}_{\mathbf{x}}\mathbf{i}_k}{\mathbf{i}_k^T\mathbf{R}_{\mathbf{x}}\mathbf{i}_k} \\ &\approx \frac{\mathbf{Z}\mathbf{P}\mathbf{z}_{\mathrm{r},k}^H}{\sigma_x^2} \ ,\end{aligned} \tag{F.79}$$

where $\mathbf{i}_k \in \mathbb{R}^{M\times 1}$ is a vector of zeros except at the kth entry which is a 1. If we insert (F.79) into (F.40), we see that the output SNR for the ODW and ODMVDR filters is

$$\begin{aligned}\text{oSNR}(\mathbf{h}_{\text{ODW},k}) &= \text{oSNR}(\mathbf{h}_{\text{ODMVDR},k}) \\ &\approx \frac{\mathbf{z}_{\mathrm{r},k}\mathbf{P}\mathbf{Z}^H\mathbf{R}_{\text{in},k}^{-1}\mathbf{Z}\mathbf{P}\mathbf{z}_{\mathrm{r},k}^H}{\sigma_x^2}\end{aligned} \tag{F.80}$$

when they are applied on periodic signals. The output SNR for the HDLCMV filter on periodic signals are given in (F.57). To find expressions for the harmonic distortion of the ODW and ODMVDR filters, we need expression for the filters for periodic signals. These can be obtained by inserting (F.79) into (F.43) and (F.50) which yields

$$\mathbf{h}_{\text{ODW},k} \approx \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}\mathbf{P}\mathbf{z}_{\mathrm{r},k}^H \ , \tag{F.81}$$

$$\mathbf{h}_{\text{ODMVDR},k} \approx \frac{\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}\mathbf{P}\mathbf{z}_{\mathrm{r},k}^H}{\mathbf{z}_{\mathrm{r},k}\mathbf{P}\mathbf{Z}^H\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}\mathbf{P}\mathbf{z}_{\mathrm{r},k}^H} \ . \tag{F.82}$$

We can then obtain closed-form expression for the harmonic distortion of the ODW and ODMVDR filters by inserting (F.81) and (F.82) into (F.37)

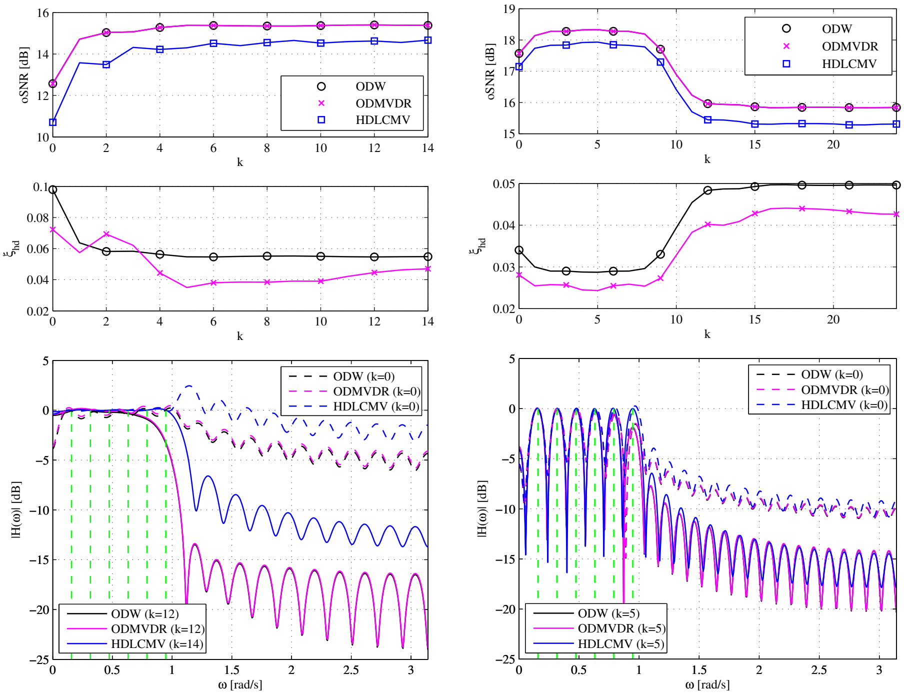
$$\xi_{\text{hd}}(\mathbf{h}_{\text{ODW}}) \approx 2\sum_{l=1}^{L} P_l \left| 1 - \left| \mathbf{z}_{\mathrm{r},k}\mathbf{P}\mathbf{Z}^H\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{z}(l\omega_0) \right|^2 \right| \ , \tag{F.83}$$

$$\xi_{\text{hd}}(\mathbf{h}_{\text{ODMVDR}}) \approx 2\sum_{l=1}^{L} P_l \left| 1 - \frac{\sigma_x^4 \left| \mathbf{z}_{\mathrm{r},k}\mathbf{P}\mathbf{Z}^H\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{z}(l\omega_0) \right|^2}{\left(\mathbf{z}_{\mathrm{r},k}\mathbf{P}\mathbf{Z}^H\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}\mathbf{P}\mathbf{z}_{\mathrm{r},k}^H \right)^2} \right| \ . \tag{F.84}$$

The harmonic distortion for the HDLCMV filter will always be zero due to its constraints.

Followingly, we have evaluated the performances of the ODW, ODMVDR, and HDLCMV filters in different scenarios. We evaluated the performances when the filters were applied for enhancement of a periodic signal, $x(n)$, in noise, $v(n)$. The periodic signal was constituted by $L = 6$ harmonic sinusoids with a pitch of $\omega_0 = 0.1578$. The amplitudes of the harmonics were chosen to be

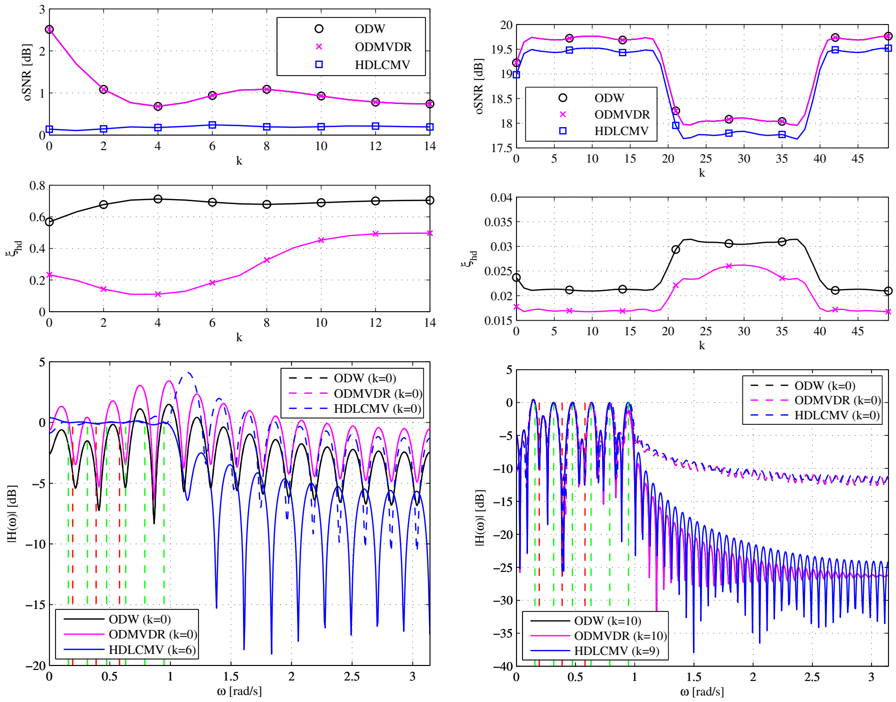
$$\begin{bmatrix} A_1 & \cdots & A_6 \end{bmatrix} = \begin{bmatrix} 1 & 0.8 & 0.5 & 0.35 & 0.2 & 0.1 \end{bmatrix} \ . \tag{F.85}$$



(a) (b)

Fig. F.1: Performance measures and filter responses of the ODW, ODMVDR, and HDLCMV filters for (a) $M = 30$ and (b) $M = 50$ when the noise is white Gaussian and the input SNR is 10 dB.

By using decreasing amplitudes, we can get insight into how the filters perform with respect to noise reduction of, for example, voiced speech. First, we evaluated the performances when the desired signal was corrupted by white Gaussian noise at an input SNR of 10 dB, and when the filter order was $M = 30$. In Fig. F.1a the results are shown for different values of k. From these results, it is clear that the output SNR can be improved significantly by changing k compared to the traditional approach where $k = 0$. For the ODW and ODMVDR filters, the output SNR can be improved by ≈ 3 dB by choosing $k = 12$, and for the HDLCMV filter an improvement of ≈ 4 dB is obtainable by choosing $k = 14$. It is important to note that we do not necessarily introduce additional harmonic distortion by improving the output SNR by changing k. In this case, the harmonic distortion is also lowered compared to $k = 0$ for both the ODW and ODMVDR filters when the output SNR is maximized in k. Also, in Fig. F.1a, we have plotted the responses of the filters, for both $k = 0$ and for the k that maximizes



(a)

(b)

Fig. F.2: Performance measures and filter responses of the ODW, ODMVDR, and HDLCMV filters for (a) $M = 30$ and (b) $M = 100$ when the noise is a sum of sinusoidal noise and white Gaussian. The ratio between the desired signal and the white noise is 10 dB and the input SNR is ≈ -0.09 dB.

the output SNR. It is clear from the filter responses, that the noise reduction can be improved significantly for $\omega > 1$ by choosing $k \neq 0$.

Then we conducted similar simulations, but with a filter order of $M = 50$. The results from these simulations are depicted in Fig. F.1b. Now, the gain that can be obtained by changing k is smaller. Compared to the case with $k = 0$, we can obtain a gain of ≈ 0.8 dB if we chose $k = 5$. Again, we can see that we can improve the output SNR and harmonic distortion simultaneously by changing k. We also plotted the filter responses. From these it is clear that we can obtain better noise reduction by choosing $k = 5$. This is especially so for high frequencies ($\omega > 1$).

We also conducted simulations where the noise was a sum of white Gaussian noise and sinusoidal noise. The sinusoidal noise source is used to investigate the impact of noise resembling voiced speech. In these simulations, the ratio between the desired signal and the white Gaussian noise was 10 dB. The sinusoidal noise source was constituted by 3 harmonic sinusoids having a pitch of 0.1932. The amplitudes of the three harmonics of the sinusoidal noise source were $\begin{bmatrix}1 & 0.9 & 0.3\end{bmatrix}$. The input SNR is therefore ≈ -0.09 dB in these simulations. Again, we conducted some simulations where the performances of the filters were evaluated for different ks. The results for a filter order of $M = 30$ are shown in Fig. F.2a. In this case, no improvement can be obtained for the ODW and ODMVDR filters compared to $k = 0$. For the HDLCMV filter a small improvement of ≈ 0.1 dB can be obtained by choosing $k = 6$ instead of $k = 0$. While the output SNR for the ODMVDR filter cannot be improved by changing k, its harmonic distortion can still be reduced. If we take a look on the filter responses, we can see that the HDLCMV filter for $k = 6$ provides significantly more noise reduction for $\omega > 1$ compared to when $k = 0$.

The simulations with sinusoidal noise were also conducted for a filter order of $M = 100$. The results from this simulation are depicted in Fig. F.2b. In these simulations, we see that the output SNR can be improved by ≈ 0.5 dB by changing k from 0 to 10 for the ODW and ODMVDR filters, and from 0 to 9 for the HDLCMV filter. We can see that the harmonic distortion of the ODW and ODMVDR filters are also improved by changing k. Again, it is clear from the frequency responses of the filters that we can obtain significantly more noise reduction for high frequencies ($\omega > 1$) by optimizing the output SNR over k.

From the results in Fig. F.1a-F.2b, we can conclude that the k that maximizes the output SNR is dependent on the filter length, the noise, the fundamental frequency and the number of harmonics. To the extend of our knowledge, there is no simple expression for this optimal k and, in practice, it therefore has to be estimated by maximizing over the estimated output SNRs for all ks in $[0; \lfloor (M-1)/2 \rfloor]$ at every time instance.

7.2 Filter Performances for Large M

We now consider the performances of the filters when we let M approach infinity. Recall that the maximum output SNR for the orthogonal decomposition based filters can be written as

$$\text{oSNR}_{\text{max},k} = \frac{1}{\left(\sigma_x^2 \boldsymbol{\rho}_{\mathbf{x}x,k}^T \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x,k}\right)^{-1} - 1} . \tag{F.86}$$

Inserting (F.79) in (F.86) and applying (F.79) in the left hand side of the denominator in (F.86) yields

$$\begin{aligned} \sigma_x^2 \boldsymbol{\rho}_{\mathbf{x}x,k}^T \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x,k} &= \frac{\mathbf{z}_{\text{r},k} \mathbf{P} \mathbf{Z}^H \left(\mathbf{Z}\mathbf{P}\mathbf{Z}^H + \mathbf{R}_{\mathbf{v}}\right)^{-1} \mathbf{Z}\mathbf{P}\mathbf{z}_{\text{r},k}^H}{\sigma_x^2} \\ &= \frac{\mathbf{z}_{\text{r},k} \mathbf{P}\mathbf{C}\mathbf{P}\mathbf{z}_{\text{r},k}^H}{\sigma_x^2} , \end{aligned} \tag{F.87}$$

where

$$\mathbf{C} = \mathbf{Z}^H \left(\mathbf{Z}\mathbf{P}\mathbf{Z}^H + \mathbf{R}_\mathbf{v}\right)^{-1} \mathbf{Z} \, . \tag{F.88}$$

When the noise is a summation of white Gaussian noise and sinusoidal interferers, it can be shown that [19]

$$\lim_{M\to\infty} \mathbf{C} = \mathbf{P}^{-1} \, . \tag{F.89}$$

If we combine (F.86), (F.87), and (F.89) we can see that

$$\lim_{M\to\infty} \text{oSNR}_{\text{max},k} = \infty \, . \tag{F.90}$$

That is, when M becomes very large and the noise is a sum of white Gaussian noise and sinusoidal noise, the maximum output SNR of the orthogonal decomposition filter will approach ∞ for all values of k. The same will be the case for the HDLCMV filter since it equals the ODMVDR filter for large M [19].

Followingly, we consider the asymptotic behavior of the harmonic distortion of the filters. Again, we assume that the desired signal is periodic. We know that the HDLCMV filter always has no harmonic distortion due to its constraints, so this filter is not considered in this investigation. The expression for the ODW and ODMVDR filters when the desired signal is periodic are given in (F.81) and (F.82). If we then let M approach infinity we can see that

$$\begin{aligned}\lim_{M\to\infty} \mathbf{h}_{\text{ODMVDR},k} = \lim_{M\to\infty} \mathbf{h}_{\text{HDLCMV},k} &= \mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}\mathbf{P}\mathbf{z}_{\text{r},k}^H \\ &= \mathbf{h}_{\text{ODW},k} \, .\end{aligned} \tag{F.91}$$

It can now be seen that the harmonic distortion of the ODW and ODMVDR filters approaches zero when M is increased. This can be seen by inserting (F.91) into (F.37), and by letting M approach infinity, which yields

$$\begin{aligned}\lim_{M\to\infty} \xi_{\text{hd}}(\mathbf{h}_{\text{ODW},k}) = \lim_{M\to\infty} \xi_{\text{hd}}(\mathbf{h}_{\text{ODMVDR},k}) &= \xi_{\text{hd}}(\mathbf{h}_{\text{HDLCMV},k}) \\ &= 0 \, .\end{aligned} \tag{F.92}$$

In summary, all filters show the same asymptotic performances for all ks both with respect to noise reduction and distortion. This motivates using the orthogonal and harmonic decomposition based filters jointly as considered in [19, 20] for $k = 0$ since they have complementary advantages and disadvantages. However, this will not be treated in this paper.

8 Example: Noise Reduction of Speech

Followingly, we demonstrate the applicability of the proposed non-causal filters on real-life signals. In particular, we consider noise reduction of speech recordings. First, we

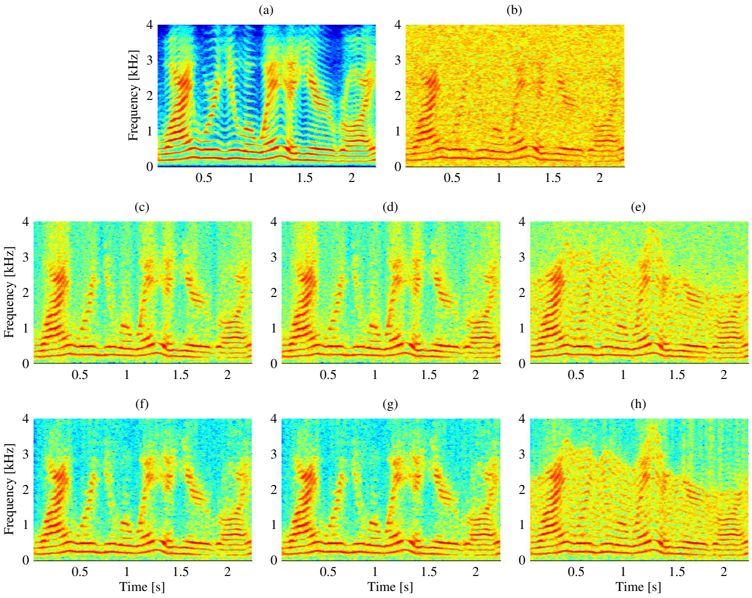


Fig. F.3: Spectrograms of (a) a clean female speaker signal, (b) a female speaker signal in white noise at an iSNR of 5 dB, (c) an enhanced signal obtained using a causal ODW filter, (d) an enhanced signal obtained using a causal ODMVDR filter, (e) an enhanced signal obtained using a causal HDLCMV filter, (f) an enhanced signal obtained using a non-causal OD Wiener filtering scheme, (g) an enhanced signal obtained using a non-causal ODMVDR filtering scheme, and (h) an enhanced signal obtained using a non-causal HDLCMV filtering scheme. The enhancement filters were all of length $M = 60$.

considered a 2.2 seconds long speech segment sampled at 8 kHz. The segment contains a female speaker uttering the sentence “Why where you away a year Roy?”. In Fig. F.3a, the spectrogram of the clean speech signal is plotted. As it can be seen from this spectrogram, the speech signal used in the first experiment on real-life speech contains voiced speech only. This was chosen to allow for the evaluation of the HDLCMV filter which is only applicable on (quasi-)periodic signals. We added white Gaussian noise to the speech signal at an average input SNR of 5 dB, and the spectrogram of the noisy signal is depicted in Fig. F.3b. The noisy signal was then enhanced using different causal and non-causal filtering schemes; we used causal and non-causal ODW, ODMVDR, and HDLCMV filtering schemes involving filters of length $M = 60$. In the filtering schemes, we used (F.43), (F.51) and (F.56) for different k; for the causal filtering schemes, k was set to 0 whereas, for the non-causal filtering schemes, k was chosen such that the estimated output SNRs of the filters were maximized at every time instance. Note that the applied filters require that the noise and/or signal statistics

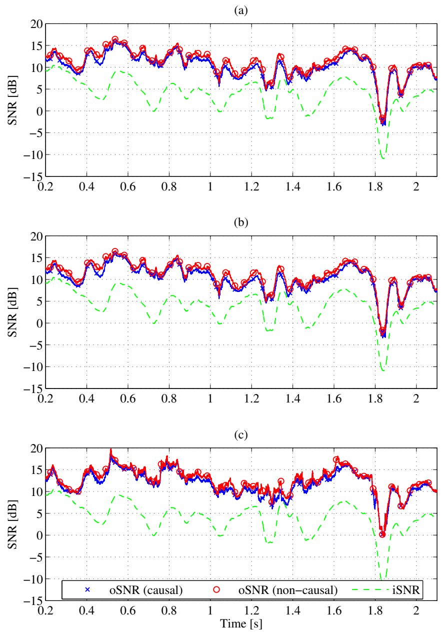


Fig. F.4: Estimated output SNRs over time for causal and non-causal (a) ODW, (b) ODMVDR, and (c) HDLCMV filtering schemes.

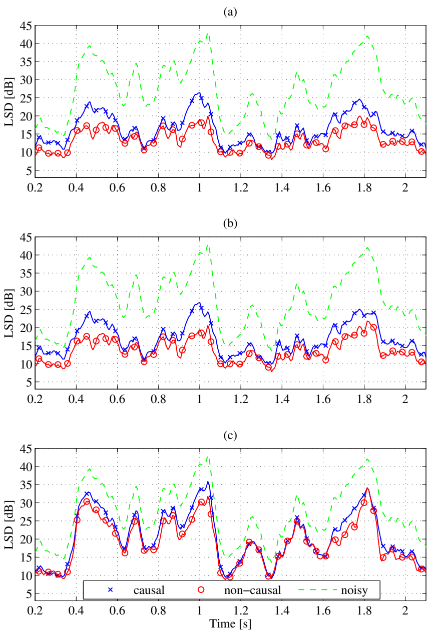
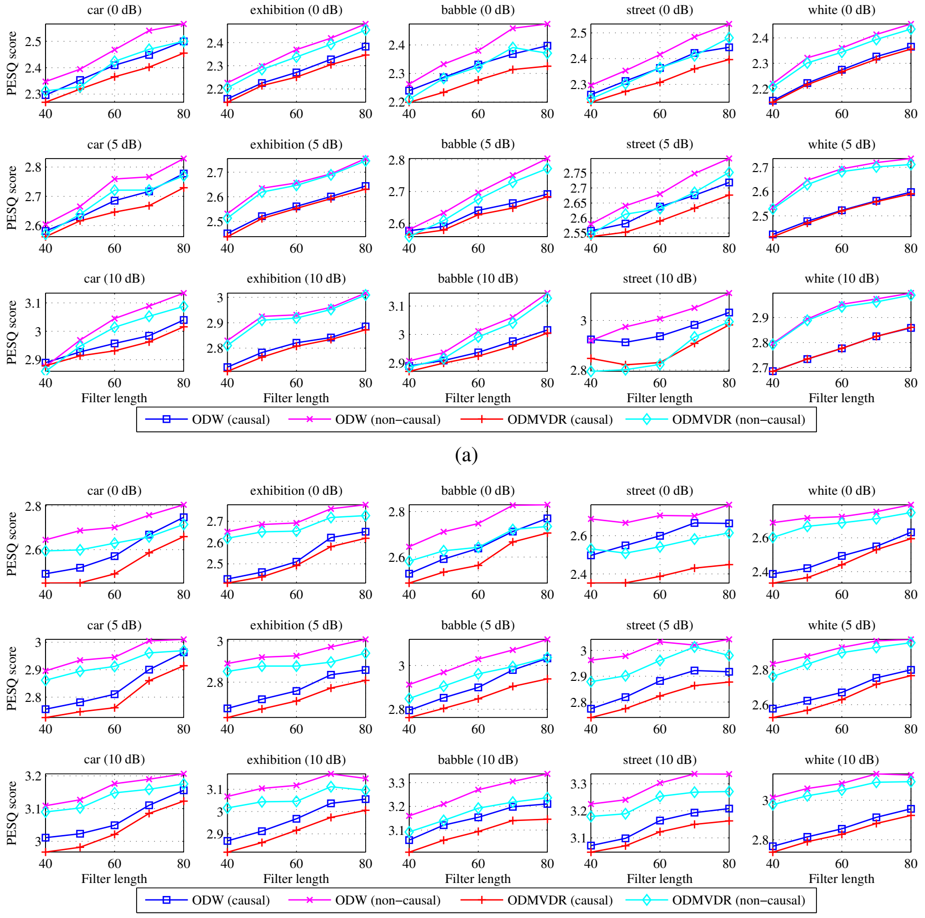


Fig. F.5: Estimated log-spectral distances over time for causal and non-causal (a) ODW, (b) ODMVDR, and (c) HDLCMV filtering schemes.

are known or estimated in practice which justifies that we require knowledge about the output SNRs. In all filtering schemes, we recalculated the filters and their output SNRs at every time instance, n, using the estimated observed signal, desired signal and noise statistics ($\hat{\mathbf{R}}_{\mathbf{y}}$, $\hat{\mathbf{R}}_{\mathbf{x}}$, and $\hat{\mathbf{R}}_{\mathbf{v}}$). The statistics were estimated from the previous $N = 400$ samples of the observed signal, desired signal and noise, respectively. We focus on comparing the performance of causal and non-causal filters in this section, so we assume that the noise signal is always available. In practice, we can estimate the noise statistics during silences by using a voice activity detector (VAD) if the noise is stationary, or we can estimate the noise statistics even in periods with voice activity using, e.g., [14, 35]. The ODW and ODMVDR filters were calculated using $\hat{\mathbf{R}}_{\mathbf{y}}$ and $\hat{\mathbf{R}}_{\mathbf{v}}$, whereas the HDLCMV filter was calculated using $\hat{\mathbf{R}}_{\mathbf{y}}$, the pitch estimated at every time instance, and a fixed harmonic model order of $L = 13$. We estimated the pitch using the orthogonality based subspace method in [17, 18] which is freely available online[1]. The model order, on the other hand, was chosen on basis of an inspection of

[1] http://www.morganclaypool.com/page/multi-pitch



(a)

(b)

Fig. F.6: PESQ scores for noisy (a) female and (b) male speech signals enhanced using the causal ODW filter, the non-causal maximum SNR ODW filtering scheme, the causal ODMVDR filter, and the non-causal maximum SNR ODMVDR filtering scheme. The scores were measured in different noise scenarios for different filter lengths.

the spectrogram in Fig. F.3a. Furthermore, in the calculations of the HDLCMV filter, we regularized the covariance matrix of the observed signal as in [36]

$$\hat{\mathbf{R}}_{\mathbf{y},\text{reg}} = (1-\gamma)\hat{\mathbf{R}}_{\mathbf{y}} + \gamma \frac{\text{Tr}\left\{\hat{\mathbf{R}}_{\mathbf{y}}\right\}}{M}\mathbf{I}\,. \tag{F.93}$$

The regularization is necessary due to estimation errors on the signal statistics and mismatch between the assumed harmonic model and the speech signal. We experienced that $\gamma = 0.7$ gives consistently good results in terms of oSNR and perceptual scores.

The spectrograms of the resulting enhanced signals obtained using the described simulation setup are depicted in Fig. F.3c-F.3h. It is clearly indicated by these spectrograms that the non-causal filtering schemes reduce the noise more than their causal counterparts. At the same time, the non-causal filtering schemes do not introduce additional distortion of the desired signal compared to the causal filters. To support the rather subjective observations on the noise reduction performances, we also estimated the output SNRs for both the non-causal filtering schemes and the causal filters at each time instance. The estimated output SNRs are depicted in Fig. F.4. As expected, the non-causal filtering schemes has higher output SNRs at every time instance compared to the causal filters. This is expected, since we maximize the output SNR at every time instance in the non-causal filtering schemes. It seems from the results in Fig. F.4 that the HDLCMV filter outperforms both the ODW and ODMVDR filters in terms of output SNR. This cannot be concluded, however, since the output SNRs for the orthogonal and harmonic decomposition are defined differently. In practice, the orthogonal decomposition based filters actually show superior noise reduction performances compared to the HDLCMV filter according to our listening experience. We also measured the log-spectral distance (LSD) between the clean signal and the enhanced signals over time, and the results are depicted in Fig. F.5. First of all, we can see from these results that the enhanced signals obtained using the non-causal filtering schemes have lower LSDs compared to the enhanced signals obtained using the causal filters at almost every time instance. That is, these results indicate that the non-causal filtering schemes have better distortion properties compared to the causal filters. Moreover, we can see from the results in Fig. F.5 that both the ODW and ODMVDR filters outperform the HDLCMV filter in terms of LSDs. This supports our previous claim on that, in practice, the ODW and ODMVDR filters introduce less distortion of the desired signal compared to the HDLCMV filter.

The results from the previous simulations indicate that we can achieve better noise reduction and distortion performances by using non-causal filtering schemes instead of non-causal filters. However, these results do not necessarily reflect the achievable perceptual improvement in performance. Therefore, we conducted another real-life experiment on speech where we considered enhancement of female (sp12.wav) and male (sp02.wav) speech signals in noise. The speech signals are part of the NOIZEUS speech corpus [37]. Note that since the utilized speech signals contain segments of unvoiced speech, we only evaluate the ODW and ODMVDR filters in this experiment. Using the ODW and ODMVDR non-causal filtering schemes and causal filters considered in the previous experiment, we conducted simulations where we enhanced the female and male speech signals in different noise scenarios, for different input SNRs, and for different filter lengths. The necessary statistics for this experiment were estimated as in the previous simulations. In each simulation, we measured the "Perceptual Evaluation of Speech Quality" (PESQ) scores [38] of the different enhanced signals compared to the clean speech signals using a freely available online toolbox[2]. The PESQ score is an objective measure which reflects the perceptual quality of a speech signal. That is, we can use PESQ scores to evaluate the practical applicability of the non-causal filtering schemes versus causal filters. The PESQ scores resulting from the simulations involving the female and male speech signals are shown in Fig. F.6. First, we observe from these results that the ODW non-causal filtering schemes and causal filters outperform the ODMVDR non-causal filtering schemes and causal filters, respectively. Moreover, we observe that, in the simulations with female speech, the non-causal filtering schemes outperform their causal counterparts in almost all scenarios. Only for some noise types (street and car), for a 10 dB iSNR, and for small filter lengths, the causal filters get similar or a slightly better PESQ scores compared to the non-causal filtering schemes. Finally, we observe that the non-causal OD Wiener and ODMVDR filtering schemes outperforms their causal versions in all of the considered scenarios with male speech. Furthermore, by listening to the enhanced signals, it is our experience that the non-causal filtering schemes indeed outperforms the corresponding causal filter versions in terms of noise reduction in most scenarios. The enhanced signals used in our informal listening test can be found at the demo website[3] for the paper.

9 Conclusion

In this paper, we proposed novel non-causal time-domain filters for noise reduction in, e.g., speech applications. The proposed filters are based on the orthogonal and harmonic signal decompositions. To enable the design of non-causal filters from these decompositions, we generalized the decompositions. We also proposed performance measures for evaluating non-causal time-domain filters based on the generalized decompositions. On a side note, we showed how the non-causal orthogonal decomposition based filters can be updated recursively when the filter order is increased. This was shown for the two particular cases where the filters are either causal or anti-causal. A by-product of these recursive updates is that we can also show how the output SNR is updated recursively which, eventually, proofs that the output SNR is always increased when we increase the filter length and the desired signal is stationary. We also conducted theoretical evaluations of the filters. In these evaluations, we assumed that the desired signal is periodic and thereby has a harmonic structure. By making this assumption, it is possible to obtain exact closed-form expressions for the performance measures of the filters. The theoretical evaluations showed that we can indeed improve both the output SNR and the harmonic distortion of the filters simultaneously by allowing the filters to be non-causal. Moreover, we applied the non-causal filters for noise reduction of noisy real-life speech signals. These simulations showed that the non-causal filters

[2] http://www.utdallas.edu/~loizou/speech/software.htm

[3] http://kom.aau.dk/~jrj/Demo/non_causal_filt/demo.html

can achieve more noise reduction compared to the causal filters in practice in terms of output SNR, log-spectral distance and PESQ scores.

References

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

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

[3] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.

[4] R. J. McAulay and M. L. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[5] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[6] K. V. Sørensen and S. V. Andersen, “Rayleigh mixture model-based hidden Markov modeling and estimation of noise in noisy speech signals,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 901–917, Mar. 2007.

[7] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, “Codebook-based bayesian speech enhancement for nonstationary environments,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 2, pp. 441–452, Feb. 2007.

[8] M. Dendrinos, S. Bakamidis, and G. Carayannis, “Speech enhancement from noise: A regenerative approach,” *Speech Communication*, vol. 10, no. 1, pp. 45 – 57, 1991.

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

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

[11] 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, p. 24, 2007.

[12] J. Benesty and J. Chen, *Optimal Time-Domain Noise Reduction Filters – A Theoretical Study*, 1st ed. Springer, 2011, no. VII.

[13] J. S. Lim, Ed., *Speech Enhancement*. Prentice-Hall, 1983.

[14] 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.

[15] J. Chen, J. Benesty, and Y. Huang, "Study of the noise-reduction problem in the Karhunen-Loève expansion domain," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 4, pp. 787–802, May 2009.

[16] J. Benesty, J. Chen, and Y. Huang, "Speech enhancement in the Karhunen-Loève expansion domain," *Synthesis Lectures on Speech and Audio Processing*, vol. 7, no. 1, pp. 1–112, 2011.

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

[18] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[19] 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.

[20] ——, "Joint filtering scheme for nonstationary noise reduction," 2012, accepted.

[21] J. Chen, J. Benesty, Y. Huang, and T. Gaensler, "On single-channel noise reduction in the time domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2011, pp. 277–280.

[22] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Elsevier Signal Process.*, vol. 80(9), pp. 1937–1944, 2000.

[23] K. W. Chan and H. C. So, "Accurate frequency estimation for real harmonic sinusoids," *IEEE Signal Process. Lett.*, vol. 11, no. 7, pp. 609–612, 2004.

[24] 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.

[25] V. Emiya, B. David, and R. Badeau, "A parametric method for pitch estimation of piano tones," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 2007, pp. 249–252.

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

[27] 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.

[28] 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, no. 5, pp. 1635–1644, Jul. 2007.

[29] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Education, Inc., 2005.

[30] J. N. Franklin, *Matrix Theory*. Prentice-Hall, 1968.

[31] J. Capon, “High-resolution frequency-wavenumber spectrum analysis,” *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[32] ——, *Nonlinear Methods of Spectral Analysis*. Springer-Verlag, 1983, ch. Maximum-Likelihood Spectral Estimation.

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

[34] C. Paleologu, J. Benesty, and S. Ciochina, “Sparse adaptive filters for echo cancellation,” *Synthesis Lectures on Speech and Audio Processing*, vol. 6, no. 1, pp. 1–124, 2010.

[35] D. Ealey, H. Kelleher, and D. Pearce, “Harmonic tunnelling: tracking non-stationary noises during speech,” in *Proc. Eurospeech*, 2001.

[36] F. van der Heijden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax, *Classification, Parameter Estimation and State Estimation - An Engineering Approach using MATLAB®*. John Wiley & Sons Ltd, 2004.

[37] 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.

[38] ITU-T, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” no. P.862, pp. 1–30, Feb. 2001.

Paper G

Enhancement of Single-Channel Periodic Signals in the Time-Domain

Jesper Rindom Jensen, Jacob Benesty, Mads Græsbøll Christensen and Søren Holdt Jensen

The paper has been published in the
IEEE Trans. on Audio, Speech and Language Processing, vol. 20(7), pp. 1948–1963, 2012.

The layout has been revised.

Abstract

Most state-of-the-art filtering methods for speech enhancement require an estimate of the noise statistics, but the noise statistics are difficult to estimate in practice when speech is present. Thus, non-stationary noise will have a detrimental impact on the performance of most speech enhancement filters. The impact of such noise can be reduced by using the signal statistics rather than the noise statistics in the filter design. For example, this is possible by assuming a harmonic model for the desired signal; while this model fits well for voiced speech, it will not be appropriate for unvoiced speech. That is, signal-dependent methods based on the signal statistics will introduce undesired distortion for some parts of speech compared to signal-independent methods based on the noise statistics. Since both the signal-independent and signal-dependent approaches to speech enhancement have advantages, it is relevant to combine them to reduce the impact of their individual disadvantages. In this paper, we give theoretical insights into the relationship between these different approaches, and these reveal a close relationship between the two approaches. This justifies joint use of such filtering methods which can be beneficial from a practical point of view. Our experimental results confirm that both signal-independent and signal-dependent approaches have advantages and that they are closely-related. Moreover, as a part of our experiments, we illustrate the practical usefulness of combining signal-independent and signal-dependent enhancement methods by applying such methods jointly on real-life speech.

1 Introduction

Human speech is frequently encountered in several signal processing applications such as telecommunications, teleconferencing, hearing-aids, and human-machine interfaces. Before the speech can be utilized in such applications, it must be picked up by one or more microphones. Unfortunately, the desired signal (in this case speech) will always, to a certain degree, be corrupted by noise which is present when sampling the signal. The noise will most likely have a detrimental impact on speech applications since it may degrade the speech quality and intelligibility. In hearing-aids, for example, a decreased speech quality (i.e., a high noise level) can cause listener fatigue. Therefore, it is of great importance to develop methods for reducing the noise of speech recordings before the speech is utilized in any relevant application. Such methods are typically termed noise reduction methods or enhancement methods. In the past few decades, developing such methods have been a major challenge. For an overview of existing enhancement methods, we refer to, e.g., [1, 2]. In general, we can divide speech enhancement methods into three groups, i.e., spectral-subtractive algorithms [3], statistical-model-based algorithms [4, 5], and subspace algorithms [6–8]. The references, [3–8], refer to some of the pioneering work within each of the groups.

A common approach used in speech enhancement is linear filtering. In this approach, the speech enhancement problem is formulated as a filter design problem. That is, a filter should be designed such that it reduces the noise level of the observed signal as much as possible while not introducing any noticeable distortion of the speech. The design of such a filter can be performed either directly in the time domain or in some transform domain. This could for example be in the frequency [3, 7, 9] or in the Karhunen-Loève expansion (KLE) domains [10]. The advantage of filtering in transform domains can, for example, be a reduced computational complexity. Filters derived in transform domains, however, can also be derived equivalently in other domains and vice versa. In this paper, we consider time-domain filters for single-channel recordings which can also be extended to other domains according to the previous discussion. Typically, time-domain filters are designed by minimizing some error function like in the classical Wiener filter design [11]. The first step in the design is therefore to define the error function.

In the vast majority of filtering methods for speech enhancement, the filter is designed from the statistics of the observed signal and the noise. We term this the signal-independent filter design approach. In practice, however, the noise signal is not directly available, and the noise statistics could, for example, be estimated during silence periods where only the noise is present. The main advantage of this approach is that it is completely independent of the statistics of the desired speech signal since it only uses the observed signal and the noise statistics, and it is well-known that the speech structure changes drastically over time. However, the signal-independent filter approach will not be influenced by this, since it does not rely on the statistics of the desired signal. Non-stationary noise, on the other hand, will have a detrimental impact on this filter design approach since the noise statistics are difficult to estimate when speech is present.

Recently, a signal-dependent filter design approach has been proposed [12]. By signal-dependent, we mean that the filter is calculated using the statistics of the desired signal and without using the statistics of the noise. The desired signal is assumed to be periodic in this approach and is therefore well-modeled by a sum of harmonically related sinusoids. This type of harmonic modeling has been used extensively within speech processing. Due to the periodicity assumption, the filter in [12] ends up being driven only by the pitch, the harmonic model order, and the statistics of the observed signal. In this paper, the pitch and the number of harmonics will be treated as known parameters, and we refer the interested reader to [13–22] and the references therein for an overview of methods for estimation of these parameters when they are unknown. Since the signal-dependent approach does not depend directly on the noise statistics, it will be robust against non-stationary noise as opposed to the signal-independent filter design approach. However, the harmonic model will only be appropriate for voiced speech segments. For unvoiced speech segments, the signal-dependent approach will therefore introduce some distortion of the speech signal due to model mismatch.

As highlighted in the previous discussion that the signal-independent and signal-dependent filter design approaches have complementary advantages and disadvantages. Therefore, it is highly relevant to investigate if these approaches can be combined to obtain the advantages of both while reducing the impact of their disadvantages. As a first step in this direction, we here provide further insight into the relationship between the signal-independent and signal-dependent filter design approaches in this paper. More specifically, we consider the relationship between two recently proposed filter designs, namely the orthogonal decomposition based minimum variance distortionless response (ODMVDR) filter [23], and the harmonic decomposition linearly constrained minimum variance (HDLCMV) filter [12, 21]. The ODMVDR filter is signal-independent whereas the HDLCMV filter is signal-dependent. Moreover, we present some closed-form performance measures for filters designed using both the signal-independent and signal-dependent design approaches when the desired signal is periodic. A new performance measure for the harmonic distortion is also proposed. The closed-form expressions for the performance measures enable easy comparison of the filters. Finally, in the experimental part of the paper, we propose a filtering scheme where the ODMVDR and HDLCMV filters are used jointly. By doing this, we can, to some extend, have the individual advantages of both a signal-independent and a signal-dependent filtering approach.

The remainder of the paper is organized as follows. In Section 2, we define the signal model which forms the basis of the paper. Then, in Section 3, we introduce the notion of using filtering for enhancement purposes for different signal decompositions. Based on this, we briefly introduce two recently proposed optimal filter designs for enhancement in Section 4. In Section 5, we perform a theoretical study of the two filters, and we show that there is a clear link between them. When the desired signal is periodic, we can obtain closed-form expression for the filter performance measures which we describe in Section 6. In the experimental part of the paper, in Section 7, we compare the ODMVDR and HDLCMV filters through simulations, and we propose and evaluate a scheme where the ODMVDR and HDLCMV filters are used jointly for speech enhancement. Finally, we conclude on the paper in Section 8.

2 Signal Model

In this paper, we consider the performance and the relationship of recent optimal filter designs for enhancement of a zero-mean desired signal, $x(n) \in \mathbb{R}^{1\times 1}$, buried in additive noise, $v(n) \in \mathbb{R}^{1\times 1}$, where n denotes the discrete-time index. That is, the objective is to recover $x(n)$ from a mixture signal given by

$$y(n) = x(n) + v(n) \; . \tag{G.1}$$

The mixture signal, $y(n) \in \mathbb{R}^{1\times 1}$, could be a microphone recording and the desired signal could be a speech signal. We assume that the noise, $v(n)$, is a zero-mean random process uncorrelated with the desired signal, $x(n)$. Specifically, we consider the special scenario where $x(n)$ is quasi-periodic which is a reasonable assumption for voiced speech segments. Considering this special scenario enables us to provide closed-form solutions for the enhancement performance measures, and it enables us to investigate the relationship between different optimal filter designs. These observations will become clear from the later sections.

By assuming quasi-periodicity, we can rewrite the signal model in (G.1) as

$$y(n) = \sum_{l=1}^{L} A_l \cos(l\omega_0 n + \phi_l) + v(n) \; , \tag{G.2}$$

where ω_0 is the pitch, L is the number of harmonics, A_l is the amplitude of the lth harmonic, and ϕ_l is the phase of the lth harmonic. For many signals, the harmonic model does not fit exactly due to inharmonicity, but we can cope with this by modifying the signal model in several ways (see, e.g., [21] and the references therein). However, inharmonicity is out of the scope of this paper, and it will not be discussed any further. Without loss of generality, we can also write the signal model in (G.2) as

$$y(n) = \sum_{l=1}^{L} \left(a_l e^{jl\omega_0 n} + a_l^* e^{-jl\omega_0 n} \right) + v(n) \; , \tag{G.3}$$

with $a_l = \frac{A_l}{2} e^{j\phi_l}$ being the complex amplitude of the lth harmonic, and $(\cdot)^*$ denotes the element-wise complex conjugate of a matrix/vector.

The observed data can be stacked into a vector, $\mathbf{y}(n) \in \mathbb{R}^{M\times 1}$, which enables us to do block processing. The vector signal model is given by

$$\mathbf{y}(n) = \mathbf{x}(n) + \mathbf{v}(n) \; , \tag{G.4}$$

where

$$\mathbf{y}(n) = \begin{bmatrix} y(n) & y(n-1) & \cdots & y(n-M+1) \end{bmatrix}^T \; , \tag{G.5}$$

with $(\cdot)^T$ denoting the matrix/vector transpose, and the definitions of $\mathbf{x}(n) \in \mathbb{R}^{M\times 1}$ and $\mathbf{v}(n) \in \mathbb{R}^{M\times 1}$ resemble the definition of $\mathbf{y}(n)$. Since we have assumed that $x(n)$ and $v(n)$ are uncorrelated, we can obtain the following simple expression for the covariance matrix, $\mathbf{R_y} \in \mathbb{R}^{M\times M}$, of the observed signal

$$\mathbf{R_y} = \mathrm{E}[\mathbf{y}(n)\mathbf{y}^T(n)] = \mathbf{R_x} + \mathbf{R_v} \; , \tag{G.6}$$

where $\mathrm{E}[\cdot]$ is the expectation operator, $\mathbf{R_x} \in \mathbb{R}^{M\times M}$ is the covariance matrix of $\mathbf{x}(n)$ and $\mathbf{R_v} \in \mathbb{R}^{M\times M}$ is the covariance matrix of $\mathbf{v}(n)$. Under the assumption of $x(n)$ being quasi-periodic, we know that $\mathbf{R_x}$ can be modeled by [24]

$$\mathbf{R_x} \approx \mathbf{Z}(\omega_0)\mathbf{P}\mathbf{Z}^H(\omega_0) \; , \tag{G.7}$$

where $(\cdot)^H$ denotes the complex conjugate transpose operator, and

$$\mathbf{P} = \mathrm{diag}\left\{\begin{bmatrix}|a_1|^2 & |a_1^*|^2 & \cdots & |a_L|^2 & |a_L^*|^2\end{bmatrix}\right\} , \tag{G.8}$$

$$\mathbf{Z}(\omega_0) = \begin{bmatrix}\mathbf{z}(\omega_0) & \mathbf{z}^*(\omega_0) & \cdots & \mathbf{z}(L\omega_0) & \mathbf{z}^*(L\omega_0)\end{bmatrix} , \tag{G.9}$$

$$\mathbf{z}(l\omega_0) = \begin{bmatrix}1 & e^{-jl\omega_0} & \cdots & e^{-jl\omega_0(M-1)}\end{bmatrix}^T , \tag{G.10}$$

with diag$\{\cdot\}$ denoting the construction of a diagonal matrix from a vector. In the remainder of the paper, we denote $\mathbf{Z}(\omega_0)$ as $\mathbf{Z}$ to get a simpler notation.

A common goal in different enhancement algorithms is then to find a "good" estimate of $x(n)$ or $\mathbf{x}(n)$. Often, in enhancement problems, "good" means that the noise reduction should be significant while the desired signal remains nearly undistorted. In this paper, we focus on two recently proposed filtering methods which estimate $x(n)$ from an observation vector, $\mathbf{y}(n)$, of length M.

3 Enhancement by Linear Filtering

Linear filters have been widely used for enhancement purposes. For example, enhancement performed by applying a finite impulse response (FIR) filter to the observed signal vector, $\mathbf{y}(n)$. The filtering operation can be written as

$$\hat{x}(n) = \sum_{m=0}^{M-1} h_m y(n-m) = \mathbf{h}^T \mathbf{y}(n) , \tag{G.11}$$

where

$$\mathbf{h} = \begin{bmatrix}h_0 & h_1 & \cdots & h_{M-1}\end{bmatrix}^T \tag{G.12}$$

and $\hat{x}(n)$ should be an estimate of $x(n)$. The output of the filter is often decomposed into a filtered desired signal part and a filtered noise part to facilitate the filter design. We here describe three different decompositions of the filter output: the classical, the orthogonal, and the harmonic decompositions.

3.1 Classical Decomposition

In most classical filtering methods for signal enhancement, the filter output is decomposed as

$$\hat{x}(n) = \mathbf{h}^T \mathbf{x}(n) + \mathbf{h}^T \mathbf{v}(n) = x_{\mathrm{f}}(n) + v_{\mathrm{rn}}(n), \tag{G.13}$$

where $x_{\mathrm{f}}(n) \triangleq \mathbf{h}^T \mathbf{x}(n)$ is the signal after filtering and $v_{\mathrm{rn}}(n) \triangleq \mathbf{h}^T \mathbf{v}(n)$ is the residual noise. The goal in the filter design is then two-fold. First, the noise should be attenuated significantly by filtering. Second, the distortion of the desired signal introduced by the filter should be low. Numerous filter designs have been proposed according to these design criteria. A common approach is to minimize the mean-square error (MSE) between the desired signal and the enhanced signal, where the error is defined as

$$e(n) = x(n) - \hat{x}(n) \; . \tag{G.14}$$

In [23], however, it was claimed and shown that this approach can be inappropriate since only some of the information embedded in $\mathbf{x}(n)$ is useful for the estimation of $x(n)$.

3.2 Orthogonal Decomposition

Recently, it has been proposed to design an enhancement filter based on an orthogonal decomposition of the desired signal since some components of $\mathbf{x}(n)$ interfere with the estimation of the desired signal $x(n)$ [23]. Using the orthogonal decomposition, the clean signal can be rewritten as

$$\mathbf{x}(n) = x(n)\boldsymbol{\rho}_{\mathbf{x}x} + \mathbf{x}_{\text{i}}(n) = \mathbf{x}_{\text{d}}(n) + \mathbf{x}_{\text{i}}(n) \; , \tag{G.15}$$

where

$$\begin{aligned} \boldsymbol{\rho}_{\mathbf{x}x} &= \frac{\text{E}[\mathbf{x}(n)x(n)]}{\text{E}[x^2(n)]} \qquad\qquad (G.16) \\ &= \begin{bmatrix} 1 & \rho_x(1) & \cdots & \rho_x(M-1) \end{bmatrix}^T \; , \\ \rho_x(m) &= \frac{\text{E}[x(n-m)x(n)]}{\text{E}[x^2(n)]} \; . \qquad\qquad (G.17) \end{aligned}$$

Note that $\mathbf{x}_{\text{d}}(n)$ is the part of $\mathbf{x}(n)$ being proportional to the desired signal $x(n)$ and $\mathbf{x}_{\text{i}}(n)$ is the "interference" being orthogonal to $\mathbf{x}_{\text{d}}(n)$. Inserting (G.15) into (G.13) yields

$$\hat{x}(n) = \mathbf{h}^T\mathbf{x}_{\text{d}}(n) + \mathbf{h}^T\mathbf{x}_{\text{i}}(n) + \mathbf{h}^T\mathbf{v}(n) \; . \tag{G.18}$$

It can be shown that the variance of $\hat{x}(n)$ is given by [23]

$$\sigma_{\hat{x}}^2 = \sigma_{x_{\text{fd}}}^2 + \sigma_{x_{\text{ri}}}^2 + \sigma_{v_{\text{rn}}}^2 \; , \tag{G.19}$$

where

$$\sigma_{x_{\text{fd}}}^2 = \mathbf{h}^T\mathbf{R}_{\mathbf{x}_{\text{d}}}\mathbf{h} = \sigma_x^2(\mathbf{h}^T\boldsymbol{\rho}_{\mathbf{x}x})^2 \; , \tag{G.20}$$

$$\sigma_{x_{\text{ri}}}^2 = \mathbf{h}^T\mathbf{R}_{\mathbf{x}_{\text{i}}}\mathbf{h} \; , \tag{G.21}$$

$$\sigma_{v_{\text{rn}}}^2 = \mathbf{h}^T\mathbf{R}_{\mathbf{v}}\mathbf{h} \; , \tag{G.22}$$

$\mathbf{R}_{\mathbf{x}_\text{d}} = \sigma_x^2 \boldsymbol{\rho}_{\mathbf{x}x} \boldsymbol{\rho}_{\mathbf{x}x}^T$ is the covariance matrix of $\mathbf{x}_\text{d}(n)$, $\sigma_x^2 = \text{E}[x^2(n)]$ is the variance of the desired signal, and $\mathbf{R}_{\mathbf{x}_\text{i}} = \text{E}[\mathbf{x}_\text{i}(n)\mathbf{x}_\text{i}^T(n)]$ is the covariance matrix of the interference, $\mathbf{x}_\text{i}(n)$.

The main difference between the classical approach and this approach is that we have two noise terms to minimize in this approach, namely $\sigma_{x_\text{ri}}^2$ and $\sigma_{v_\text{rn}}^2$. Moreover, the filtered desired signal is different in this approach since it does not include the interfering part of $\mathbf{x}(n)$ which is here considered as noise. Like in the previous approach, the filter should be designed such that the error in (G.14) is small (e.g., in the MSE sense) while there is no or only a little distortion of the desired signal.

3.3 Harmonic Decomposition

The harmonic model in (G.2) has been used in many pitch estimation methods [21]. In general, the model can be used for describing periodic signals as

$$\mathbf{x}(n) = \mathbf{Z}\mathbf{a}(n) = \mathbf{x}_\text{d}'(n) \; , \tag{G.23}$$

where

$$\mathbf{a}(n) = \begin{bmatrix} a_1 e^{j\omega_0 n} & a_1^* e^{-j\omega_0 n} & \cdots \\ & & a_L e^{jL\omega_0 n} & a_L^* e^{-jL\omega_0 n} \end{bmatrix}^T . \tag{G.24}$$

Note that in this approach there is no interference as opposed to in the orthogonal decomposition approach since all samples in $\mathbf{x}(n)$ can be fully used for describing the desired signal $x(n)$. This is due to the underlying harmonic signal model. Therefore, the vector, $\mathbf{x}_\text{d}'(n)$, describing the desired signal, $x(n)$, is simply equal to the signal vector, $\mathbf{x}(n)$, in this approach. The desired signal, $x(n)$, is equal to the first entry of the vector $\mathbf{Z}\mathbf{a}(n)$, i.e.,

$$x(n) = \mathbf{1}^T \mathbf{a}(n) \; , \tag{G.25}$$

where $\mathbf{1} = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^T$. Like in the orthogonal decomposition approach, we can insert (G.23) into (G.13) which yields the following estimate of $x(n)$

$$\hat{x}'(n) = \mathbf{h}^T \mathbf{x}_\text{d}'(n) + \mathbf{h}^T \mathbf{v}(n) \; . \tag{G.26}$$

If we exploit the orthogonality between $\mathbf{x}_\text{d}'(n)$ and $\mathbf{v}(n)$ in (G.26), we can write the variance of $\hat{x}'(n)$ as

$$\sigma_{\hat{x}'}^2 = \sigma_{x_\text{fd}'}^2 + \sigma_{v_\text{rn}}^2 \; , \tag{G.27}$$

where

$$\sigma_{x_\text{fd}'}^2 = \mathbf{h}^T \mathbf{R}_{\mathbf{x}_\text{d}'} \mathbf{h} = \mathbf{h}^T \mathbf{Z} \mathbf{P} \mathbf{Z}^H \mathbf{h} \; , \tag{G.28}$$

and $\sigma^2_{v_{\text{rn}}}$ is defined as in (G.22). Moreover, $\mathbf{R}_{\mathbf{x}'_{\text{d}}} = \text{E}\left[\mathbf{x}'_{\text{d}}(n)\mathbf{x}'^{T}_{\text{d}}(n)\right] = \mathbf{Z}\mathbf{P}\mathbf{Z}^H$ is the covariance matrix of $\mathbf{x}'_{\text{d}}(n)$.

Compared to the orthogonal decomposition approach, this approach only has one noise term, $\sigma^2_{v_{\text{rn}}}$. When this approach is used, the filter, $\mathbf{h}$, should therefore be designed such that it minimizes $\sigma^2_{v_{\text{rn}}}$ without distorting the $x(n)$ too much.

4 Optimal Filters for Enhancement

We consider two recently proposed filter designs for enhancement of single-channel signals: 1) the orthogonal decomposition MVDR filter [23] and 2) the harmonic decomposition LCMV filter [20]. Following, we will revisit the two filter designs.

4.1 Orthogonal Decomposition MVDR

Traditionally, the minimum variance distortionless response (MVDR) filter proposed by Capon [25, 26] has been derived and applied in the context of multichannel signals. Recently, however, the MVDR filter has also been applied for single-channel speech enhancement [23]. Here, we term the MVDR filter proposed in [23] as the orthogonal decomposition MVDR (ODMVDR) filter. The ODMVDR filter design is based on an orthogonal decomposition of the desired signal as described in Section 3.2. The filter is designed to minimize the sum of the residual interference variance, $\sigma^2_{x_{\text{ri}}}$, and the residual noise variance, $\sigma^2_{v_{\text{rn}}}$, while it should not distort the desired signal. That is,

$$\min_{\mathbf{h}} \mathbf{h}^T\mathbf{R}_{\text{in}}\mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^T\boldsymbol{\rho}_{\mathbf{x}x} = 1, \tag{G.29}$$

where $\mathbf{R}_{\text{in}} = \mathbf{R}_{\mathbf{x}_{\text{i}}} + \mathbf{R}_{\mathbf{v}}$ is the interference-plus-noise covariance matrix. The constraint comes from the measure of desired signal reduction (aka. speech reduction) for the orthogonal decomposition introduced in [23]

$$\xi_{\text{dsr}}(\mathbf{h}) = \frac{\sigma_x^2}{\sigma^2_{x_{\text{fd}}}} = \frac{1}{(\mathbf{h}^T\boldsymbol{\rho}_{\mathbf{x}x})^2} \,. \tag{G.30}$$

When $\xi_{\text{dsr}}(\mathbf{h}) = 1$ there is no desired signal reduction (or distortion if you will) while it is expected to be greater than 1 when there is a reduction. That is, to make the filter distortionless according to this measure, we must require that $\mathbf{h}^T\boldsymbol{\rho}_{\mathbf{x}x} = 1$ which exactly corresponds to the constraint in (G.29).

The well-known solution to the quadratic optimization problem in (G.29) is given by

$$\mathbf{h}_{\text{ODMVDR}} = \frac{\mathbf{R}_{\text{in}}^{-1}\boldsymbol{\rho}_{\mathbf{x}x}}{\boldsymbol{\rho}^T_{\mathbf{x}x}\mathbf{R}_{\text{in}}^{-1}\boldsymbol{\rho}_{\mathbf{x}x}} = \frac{\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\rho}_{\mathbf{x}x}}{\boldsymbol{\rho}^T_{\mathbf{x}x}\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\rho}_{\mathbf{x}x}} \,. \tag{G.31}$$

In practice, the correlation vector, $\boldsymbol{\rho}_{\mathbf{x}x}$, in (G.31) is replaced by

$$\begin{aligned}\boldsymbol{\rho}_{\mathbf{x}x} &= \frac{\mathrm{E}[\mathbf{y}(n)y(n)] - \mathrm{E}[\mathbf{v}(n)v(n)]}{\sigma_y^2 - \sigma_v^2} \\ &= \frac{\sigma_y^2\boldsymbol{\rho}_{\mathbf{y}y} - \sigma_v^2\boldsymbol{\rho}_{\mathbf{v}v}}{\sigma_y^2 - \sigma_v^2} \,, \end{aligned} \tag{G.32}$$

where σ_y^2 is the variance of $y(n)$, σ_v^2 is the variance of $v(n)$, and $\boldsymbol{\rho}_{\mathbf{y}y}$ and $\boldsymbol{\rho}_{\mathbf{v}v}$ are defined similarly to $\boldsymbol{\rho}_{\mathbf{x}x}$ in (G.16). The evaluation of the performance of the ODMVDR filter follows from later sections.

4.2 Harmonic Decomposition LCMV

Like the MVDR filter, the linearly constrained minimum variance (LCMV) filter proposed by Frost [27] has mainly been used in multichannel settings. Recently, however, an LCMV filtering method for enhancement of periodic signals was proposed which is applicable on single-channel signals [12, 20]. Following, we recast the LCMV design procedure from [20] such that it is more general and compliant with the harmonic decomposition in Section 3.3. This design procedure is somewhat similar to that of the ODMVDR filter.

In the harmonic decomposition LCMV (HDLCMV) filter, it is assumed that the desired signal is periodic. When the desired signal is periodic and modeled by (G.3), all information in $\mathbf{x}(n)$ can be used in the estimation of $x(n)$ which, in general, is not the case in the orthogonal decomposition approach where there will be some interference, $\mathbf{x}_{\mathrm{i}}(n)$. Therefore, we only need to care about minimizing the residual noise power, $\sigma_{v_{\mathrm{rn}}}^2$, in the harmonic decomposition approach without introducing too much desired signal distortion. The HDLCMV filter, in particular, is designed such that the residual noise variance, $\sigma_{v_{\mathrm{rn}}}^2$, is minimized while the desired signal, $x(n)$, is passed undistorted. This can also be casted as the following optimization problem

$$\min_{\mathbf{h}} \mathbf{h}^T\mathbf{R}_{\mathbf{v}}\mathbf{h} \quad \text{s.t.} \quad \mathbf{Z}^H\mathbf{h} = \mathbf{1} \,. \tag{G.33}$$

To verify that the constraint in (G.33) makes the filter distortionless, we consider the desired signal reduction measure for the harmonic decomposition approach which is given by

$$\xi'_{\mathrm{dsr}}(\mathbf{h}) = \frac{\sigma_x^2}{\sigma_{x'_{\mathrm{fd}}}^2} = \frac{\sigma_x^2}{\mathbf{h}^T\mathbf{Z}\mathbf{P}\mathbf{Z}^H\mathbf{h}} \,. \tag{G.34}$$

It can be seen that when the signal is periodic, the desired signal variance is given by $\sigma_x^2 = \mathbf{1}^T\mathbf{P}\mathbf{1}$. That is, the filter will indeed be distortionless with respect to the distortion measure in (G.34) if it is designed such that $\mathbf{Z}\mathbf{h} = \mathbf{1}$. It can also be shown that the constraint in (G.33) ensures that the individual harmonics are not distorted [24].

If we solve the quadratic optimization problem with multiple constraints in (G.33), we get

$$\mathbf{h}_{\text{HDLCMV}} = \mathbf{R}_{\mathbf{v}}^{-1}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{R}_{\mathbf{v}}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1} \, . \tag{G.35}$$

In the Appendix, we have shown that replacing $\mathbf{R}_{\mathbf{v}}$ by $\mathbf{R}_{\mathbf{y}}$ does not change the filter response. If we utilize this, we can also write the HDLCMV filter as

$$\mathbf{h}_{\text{HDLCMV}} = \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1} \, . \tag{G.36}$$

We can see from this expression that if $x(n)$ is periodic, the pitch, ω_0, is known, and the number of harmonics, L, is known, we only need the statistics, $\mathbf{R}_{\mathbf{y}}$, of the observed signal to design the HDLCMV filter. This is a key difference from the design of the ODMVDR filter for which we also need to know either the statistics of the desired signal, $\boldsymbol{\rho}_{\mathbf{x}x}$, or of the noise, $\boldsymbol{\rho}_{\mathbf{v}v}$.

5 Relation between the ODMVDR and HDLCMV Filters

Although the ODMVDR and HDLCMV filters were derived under different constraints, we show in this section that there is a clear link between the filters. For this analysis, we assume that the noise is a sum of interfering sinusoids and white Gaussian noise such that

$$\mathbf{R}_{\mathbf{v}} = \mathbf{Z}_{\text{sn}}\mathbf{P}_{\text{sn}}\mathbf{Z}_{\text{sn}}^H + \sigma_{\text{wn}}^2\mathbf{I} \, , \tag{G.37}$$

where $\mathbf{Z}_{\text{sn}}$ and $\mathbf{P}_{\text{sn}}$ are the steering and power matrices of the sinusoidal noise source, and σ_{wn}^2 is the variance of the white Gaussian noise. The matrices are defined similarly to (G.8) and (G.9).

It is clear from (G.16) that $\boldsymbol{\rho}_{\mathbf{x}x}$ corresponds to the first column of $\mathbf{R}_{\mathbf{x}}$ normalized with respect to the signal variance, σ_x^2. That is, without loss of generality, we can also write $\boldsymbol{\rho}_{\mathbf{x}x}$ as

$$\boldsymbol{\rho}_{\mathbf{x}x} = \frac{\mathbf{R}_{\mathbf{x}}\mathbf{i}}{\mathbf{i}^T\mathbf{R}_{\mathbf{x}}\mathbf{i}} \, , \tag{G.38}$$

where $\mathbf{i} = \begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix}^T \in \mathbb{R}^{M\times 1}$. Under the periodicity assumption, we can rewrite this expression by inserting (G.7) into (G.38)

$$\boldsymbol{\rho}_{\mathbf{x}x} = \frac{\mathbf{Z}\mathbf{P}\mathbf{Z}^H\mathbf{i}}{\mathbf{i}^T\mathbf{Z}\mathbf{P}\mathbf{Z}^H\mathbf{i}} = \frac{\mathbf{Z}\mathbf{P}\mathbf{1}}{\mathbf{1}^T\mathbf{P}\mathbf{1}} = \frac{\mathbf{Z}\mathbf{P}\mathbf{1}}{\sigma_x^2} \, . \tag{G.39}$$

If we substitute this expression for $\boldsymbol{\rho}_{\mathbf{x}x}$ back into the expression for the ODMVDR filter in (G.31), we get that

$$\begin{aligned}\mathbf{h}_{\text{ODMVDR}} &= \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}\mathbf{P}\mathbf{1}\left(\mathbf{1}^T\mathbf{P}\mathbf{Z}^H\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}\mathbf{P}\mathbf{1}\right)^{-1}\sigma_x^2 \\ &= \sigma_x^2\mathbf{B}\mathbf{P}\left(\mathbf{1}^T\mathbf{P}\mathbf{C}\mathbf{P}\mathbf{1}\right)^{-1}\mathbf{1}\,, \end{aligned} \tag{G.40}$$

where $\mathbf{B} = \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}$ and $\mathbf{C} = \mathbf{Z}^H\mathbf{B}$. Note that using the same notation, the HDLCMV filter can be written as

$$\mathbf{h}_{\text{HDLCMV}} = \mathbf{B}\mathbf{C}^{-1}\mathbf{1}\,. \tag{G.41}$$

At a first glance, the filters in (G.40) and (G.41) do not look similar. However, by using the matrix inversion lemma on $\mathbf{C}$, we see that it can be rewritten as

$$\begin{aligned}\mathbf{C} &= \mathbf{Z}^H\left(\mathbf{Z}\mathbf{P}\mathbf{Z}^H + \mathbf{R}_{\mathbf{v}}\right)^{-1}\mathbf{Z} \\ &= \mathbf{Z}^H\left[\mathbf{R}_{\mathbf{v}}^{-1} - \mathbf{R}_{\mathbf{v}}^{-1}\mathbf{Z}\left(\mathbf{P}^{-1} + \mathbf{Z}^H\mathbf{R}_{\mathbf{v}}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{R}_{\mathbf{v}}^{-1}\right]\mathbf{Z} \\ &= \mathbf{D} - \mathbf{D}\left(\mathbf{P}^{-1} + \mathbf{D}\right)^{-1}\mathbf{D}\,, \end{aligned} \tag{G.42}$$

where $\mathbf{D} = \mathbf{Z}^H\mathbf{R}_{\mathbf{v}}^{-1}\mathbf{Z}$. If we also use the matrix inversion lemma on $\mathbf{D}$, we get that

$$\mathbf{D} = \frac{1}{\sigma_{\text{wn}}^2}\mathbf{Z}^H\mathbf{Z} - \frac{1}{\sigma_{\text{wn}}^2}\mathbf{Z}^H\mathbf{Z}_{\text{sn}}\left(\mathbf{P}_{\text{sn}}^{-1} + \frac{1}{\sigma_{\text{wn}}^2}\mathbf{Z}_{\text{sn}}^H\mathbf{Z}_{\text{sn}}\right)^{-1}\mathbf{Z}_{\text{sn}}^H\mathbf{Z}\,. \tag{G.43}$$

Moreover, if we then assume that the frequencies of the sinusoidal noise sources are different from the harmonic frequencies, and if we let $M \to \infty$, we can write [21]

$$\lim_{M\to\infty}\frac{1}{M}\mathbf{Z}^H\mathbf{Z} = \mathbf{I}\,, \tag{G.44}$$

$$\lim_{M\to\infty}\frac{1}{M}\mathbf{Z}^H\mathbf{Z}_{\text{sn}} = \mathbf{0}\,. \tag{G.45}$$

Thus, for large M, we can approximate $\mathbf{C}$ as

$$\mathbf{C} \approx \sigma_v^{-2}\left[M\mathbf{I} - M^2\left(\sigma_v^2\mathbf{P}^{-1} + M\mathbf{I}\right)^{-1}\right]\,. \tag{G.46}$$

Furthermore, it turns out that we can approximate the (p,q)th element of $\mathbf{C}$ as

$$[\mathbf{C}]_{pq} \approx \begin{cases}\dfrac{M}{\sigma_v^2 + P_qM}\,, & \text{for } p = q \\ 0\,, & \text{for } p \neq q\end{cases}\,. \tag{G.47}$$

When M is large and $P_q M \gg \sigma_v^2$, the expression for the qth diagonal element of $\mathbf{C}$ can be further simplified as $[\mathbf{C}]_{qq} \approx P_q^{-1}$. In this case, we can write

$$\mathbf{C} \approx \mathbf{P}^{-1} \,. \tag{G.48}$$

If we insert this approximation for $\mathbf{C}$ in (G.40), we readily obtain that

$$\begin{aligned}\lim_{M\to\infty} \mathbf{h}_{\text{ODMVDR}} &= \mathbf{BP1} \\ &= \mathbf{h}_{\text{HDLCMV}} \,. \end{aligned} \tag{G.49}$$

Thus, when the desired signal is periodic, the noise is a summation of interfering sinusoids and white Gaussian noise, and the filter order M is large, then the ODMVDR and HDLCMV filters are approximately identical. This observation is important since it justifies the joint use of the two filters for enhancement of quasi-periodic signals. The two different filters are based on different knowledge, i.e., the noise and signal statistics, respectively. Depending on which statistics are available, the appropriate filter can be applied. In the experimental part of the paper, we also investigate the relation between the filters for small Ms.

6 Performance Measures

In [23], a number of performance measures for enhancement methods were introduced. In this section, we exploit the periodicity of the desired signal to derive closed-form expressions for the performance measures for each of the filters described in Section 4.

6.1 Noise Reduction

The most fundamental measure of the performance of enhancement algorithms is the signal-to-noise ratio (SNR). In general, we can consider two SNRs, namely the input SNR (iSNR) and the output SNR (oSNR). The iSNR is defined as the SNR of the observed signal before filtering, i.e.,

$$\text{iSNR} = \frac{\sigma_x^2}{\sigma_v^2} \,. \tag{G.50}$$

The oSNR, on the other hand, is the SNR after noise reduction. That is, when using the orthogonal decomposition, it is obtained as

$$\text{oSNR}^{\text{OD}}(\mathbf{h}) = \frac{\sigma_{x_{\text{fd}}}^2}{\sigma_{x_{\text{ri}}}^2 + \sigma_{v_{\text{rn}}}^2} = \frac{\sigma_x^2 \left(\mathbf{h}^T \boldsymbol{\rho}_{\mathbf{x}x}\right)^2}{\mathbf{h}^T \mathbf{R}_{\text{in}} \mathbf{h}} \,. \tag{G.51}$$

where $(\cdot)^{\text{OD}}$ denotes that the measure is applicable when using the orthogonal decomposition. We can then obtain a closed-form expression for the oSNR of the ODMVDR filter when the desired signal is periodic by inserting (G.39) and (G.40) into (G.51). This yields

$$\text{oSNR}^{\text{OD}}(\mathbf{h}_{\text{ODMVDR}}) = \frac{\mathbf{1}^T\mathbf{P}\mathbf{Z}^H\mathbf{R}_{\text{in}}^{-1}\mathbf{Z}\mathbf{P}\mathbf{1}}{\sigma_x^2} \ . \tag{G.52}$$

When the harmonic decomposition is utilized, the oSNR is given as

$$\text{oSNR}^{\text{HD}}(\mathbf{h}) = \frac{\sigma_{x'_{\text{fd}}}^2}{\sigma_{v_{\text{rn}}}^2} = \frac{\mathbf{h}^T\mathbf{Z}\mathbf{P}\mathbf{Z}^H\mathbf{h}}{\mathbf{h}^T\mathbf{R}_{\mathbf{v}}\mathbf{h}} \ , \tag{G.53}$$

where $(\cdot)^{\text{HD}}$ denotes that the measure is applicable when using the harmonic decomposition. A closed-form expression for the oSNR of the HDLCMV filter is then found by inserting (G.41) into (G.53), which yields

$$\text{oSNR}^{\text{HD}}(\mathbf{h}_{\text{HDLCMV}}) = \frac{\sigma_x^2}{\mathbf{1}^T\left(\mathbf{Z}^H\mathbf{R}_{\mathbf{v}}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1}} \ . \tag{G.54}$$

Yet another performance measure related to the noise reduction, is the so-called noise reduction factor, $\xi_{\text{nr}}(\mathbf{h})$. This factor is defined as the ratio between the noise in the observed signal and the noise remaining in the signal after filter. That is, when the orthogonal decomposition is used, the noise reduction factor is given by

$$\begin{aligned}\xi_{\text{nr}}^{\text{OD}}(\mathbf{h}) &= \frac{\sigma_v^2}{\sigma_{x_{\text{ri}}}^2 + \sigma_{v_{\text{rn}}}^2} \\ &= \frac{\sigma_v^2}{\mathbf{h}^T\mathbf{R}_{\text{in}}\mathbf{h}} \ .\end{aligned} \tag{G.55}$$

The noise reduction factor is expected to be larger than or equal to 1, since $\xi_{\text{nr}}(\mathbf{h}) < 1$ would imply that the noise is amplified through the filtering. If we insert the expression for the ODMVDR filter into (G.40), we get that

$$\xi_{\text{nr}}^{\text{OD}}(\mathbf{h}_{\text{ODMVDR}}) = \frac{\sigma_v^2\mathbf{1}^T\mathbf{P}\mathbf{Z}^H\mathbf{R}_{\text{in}}^{-1}\mathbf{Z}\mathbf{P}\mathbf{1}}{\sigma_x^4} \ . \tag{G.56}$$

If the harmonic decomposition is used instead, the noise reduction factor is obtained as

$$\xi_{\text{nr}}^{\text{HD}}(\mathbf{h}) = \frac{\sigma_v^2}{\sigma_{v_{\text{rn}}}^2} = \frac{\sigma_v^2}{\mathbf{h}^T\mathbf{R}_{\mathbf{v}}\mathbf{h}} \ . \tag{G.57}$$

This gives the following noise reduction factor for the HDLCMV filter

$$\xi_{\text{nr}}^{\text{HD}}(\mathbf{h}_{\text{HDLCMV}}) = \frac{\sigma_v^2}{\mathbf{1}^T\left(\mathbf{Z}^H\mathbf{R}_{\mathbf{v}}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1}} \ . \tag{G.58}$$

Note that if we know the pitch, ω_0, the number of harmonics, L, the powers of the harmonics, P_l, and the noise statistics, $\mathbf{R}_{\mathbf{v}}$, we can calculate the output SNRs and the noise reduction factors for the two filters.

6.2 Signal Distortion

A common and unwanted side-effect of most enhancement procedures is that they also attenuate the desired signal in the process of attenuating the noise. The desired signal attenuation can also be considered as distortion. The amount of distortion can be quantified by the speech reduction factor measure [23]. Here, the measure will be termed the desired signal reduction factor since we do not consider speech only. The reduction factor is defined as the ratio between the variance of the desired signal and the variance of the desired signal after filtering. That is, when the orthogonal decomposition is used, the factor is given by

$$\begin{aligned}\xi_{\text{dsr}}^{\text{OD}}(\mathbf{h}) &= \frac{\sigma_x^2}{\sigma_{x_{\text{fd}}}^2} \\ &= \frac{1}{\left(\mathbf{h}^T \boldsymbol{\rho}_{\mathbf{x}x}\right)^2} \ .\end{aligned} \quad \text{(G.59)}$$

If distortion occurs, the noise reduction factor will be greater or less than one (expectedly greater than one) and it will equal 1 otherwise. Therefore, if a filter should be distortionless, we must require that

$$\mathbf{h}^T \boldsymbol{\rho}_{\mathbf{x}x} = 1 \ . \quad \text{(G.60)}$$

The ODMVDR filter was derived exactly under this constraint, i.e.,

$$\xi_{\text{dsr}}^{\text{OD}}(\mathbf{h}_{\text{ODMVDR}}) = 1 \ , \quad \text{(G.61)}$$

which can also be easily verified. Similarly, for the harmonic decomposition approach, the desired signal distortion is defined as

$$\xi_{\text{dsr}}^{\text{HD}}(\mathbf{h}) = \frac{\sigma_x^2}{\sigma_{x'_{\text{fd}}}^2} = \frac{\sigma_x^2}{\mathbf{h}^T \mathbf{Z} \mathbf{P} \mathbf{Z}^H \mathbf{h}} \ . \quad \text{(G.62)}$$

The HDLCMV filter is designed to be distortionless when the desired signal is periodic, i.e.,

$$\xi_{\text{dsr}}^{\text{HD}}(\mathbf{h}_{\text{HDLCMV}}) = 1 \ . \quad \text{(G.63)}$$

This result can easily be verified. On a side note, it can be seen that the HDLCMV filter is also distortionless with respect to the desired signal reduction measure for the orthogonal decomposition approach since

$$\begin{aligned}\mathbf{h}_{\text{HDLCMV}}^T \boldsymbol{\rho}_{\mathbf{x}x} &= \frac{\mathbf{1}^T \left(\mathbf{Z}^H \mathbf{R}_{\mathbf{y}}^{-1} \mathbf{Z}\right)^{-1} \mathbf{Z}^H \mathbf{R}_{\mathbf{y}}^{-1} \mathbf{Z} \mathbf{P} \mathbf{1}}{\sigma_x^2} \\ &= \frac{\mathbf{1}^T \mathbf{P} \mathbf{1}}{\sigma_x^2} = 1 \ .\end{aligned} \quad \text{(G.64)}$$

This emphasizes the strong link between the two filters.

We also propose a new distortion measure, namely the harmonic distortion. The harmonic distortion is the sum of the differences between the powers of the harmonics before and after filtering which can also be written as

$$\begin{aligned}\xi_{\text{hd}}(\mathbf{h}) &= 2\sum_{l=1}^{L} |P_l - P_{\text{f},l}| \\ &= 2\sum_{l=1}^{L} P_l |1 - \mathbf{h}^T \mathbf{z}(l\omega_0)\mathbf{z}^H(l\omega_0)\mathbf{h}| \,, \end{aligned} \tag{G.65}$$

where $P_l = |a_l|^2$ and $P_{\text{f},l}$ is the power of the lth harmonic after filtering. This performance measure is defined in exactly the same way for both the orthogonal decomposition approach and the harmonic decomposition approach. The harmonic distortion will be equal to 0 when there is no distortion of the harmonics while it will be greater than 0 otherwise. A closed-form expression for the harmonic distortion of the ODMVDR filter can be obtained by inserting (G.40) into (G.65) which yields

$$\xi_{\text{hd}}(\mathbf{h}_{\text{ODMVDR}}) = 2\sum_{l=1}^{L} P_l \left| 1 - \frac{\sigma_x^4 \left| \mathbf{1}^T \mathbf{P} \mathbf{Z}^H \mathbf{R}_{\mathbf{y}}^{-1} \mathbf{z}(l\omega_0) \right|^2}{\left(\mathbf{1}^T \mathbf{P} \mathbf{Z}^H \mathbf{R}_{\mathbf{y}}^{-1} \mathbf{Z} \mathbf{P} \mathbf{1} \right)^2} \right| \,. \tag{G.66}$$

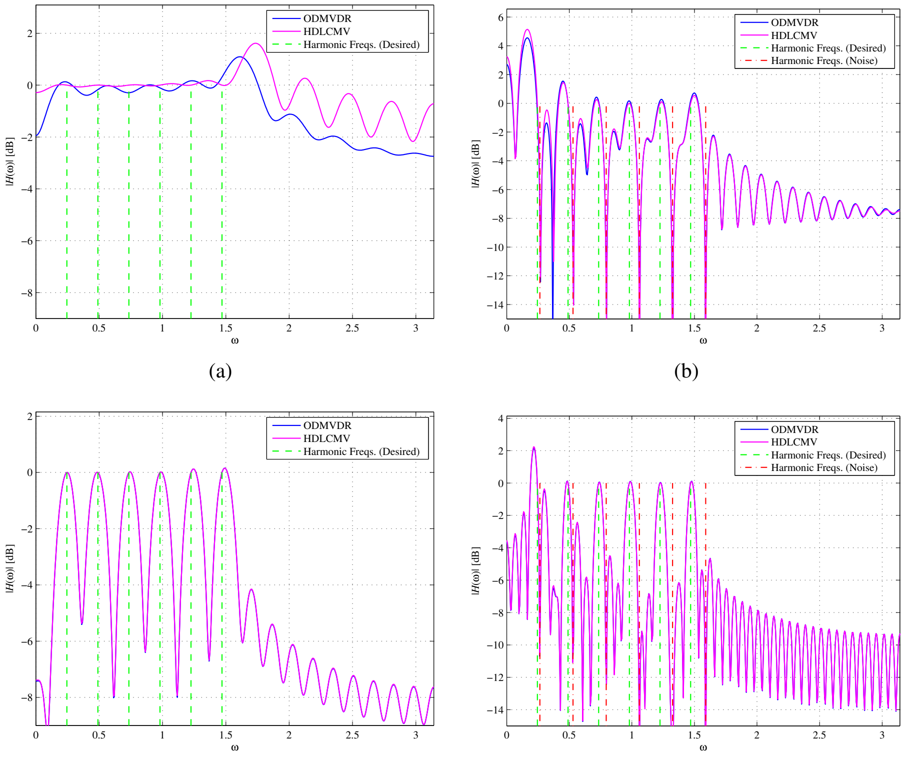
It is clear from the above expression that the harmonic distortion of the ODMVDR filter will be close to 0 when M is large. The HDLCMV filter is derived under the constraints that the harmonics should not be distorted, i.e.,

$$\xi_{\text{hd}}(\mathbf{h}_{\text{HDLCMV}}) = 0 \,, \tag{G.67}$$

which is readily verified by inserting (G.41) into (G.65).

7 Experimental Results

In the previous sections, we presented two single-channel filtering methods which can be used for extraction of periodic sources. These are the ODMVDR and HDLCMV filters. We showed that there is a clear link between the filters and that they are even equivalent in some special scenarios. To illustrate the link, we compare the responses of the filters in this section. The link between the filters suggests that they can be used jointly which can be useful in practice as we illustrate and account for in the application example later in this section. Furthermore, we defined some performance measures for both of the methods given that the underlying desired signal is periodic and modeled by (G.3). In this section, we will also study these measures through theoretical simulations.



(a) (b)

(c) (d)

Fig. G.1: Magnitude responses of the ODMVDR and HDLCMV filters of order (a) $M = 20$ and (c) $M = 40$ designed for a periodic signal corrupted by white Gaussian noise, and of order (b) $M = 50$ and (d) $M = 100$ when the noise also contained an interfering periodic signal.

7.1 Qualitative Comparison of Filter Responses

In this theoretical experiment, we compared the ODMVDR and HDLCMV filters in terms of their filter responses in different scenarios. The signal and noise statistics were assumed to be known in this experiment, i.e., we assumed that the desired signal was constituted by a sum of $L = 6$ harmonic sinusoids with a pitch of $\omega_0 = 0.245$. Each of the sinusoids was assumed to have a unit amplitude ($A_l = 1$).

In the first part of the experiment, we compared the ODMVDR and HDLCMV filters in (G.31) and (G.36), respectively, when white Gaussian noise, $v_{\text{wn}}(n)$, was added to the desired signal, $x(n)$, at an iSNR of 10 dB. When the filter length was set to $M = 20$, we obtained the filter responses depicted in Fig. G.1a. We observe from the plot that the filters have poor noise reduction capabilities due to the relatively short filter length. Furthermore, we can see that the filters have different magnitude responses. By careful inspection, we note that the HDLCMV filter has unit gains at the harmonic frequencies as a result of its constraints which is not the case for the ODMVDR filter. When we increase the filter length to $M = 40$, we get the responses in Fig. G.1c. In accordance with the theoretical discussion in Section 5, we observe that the filters become equivalent when the filter order becomes large.

In the second part of the experiment, the noise was a summation of white Gaussian noise, $v_{\text{wn}}(n)$, and sinusoidal noise, $v_{\text{sn}}(n)$, containing 6 harmonics with unit amplitudes. The pitch of the sinusoidal noise source was 0.247. The ratio between the desired signal and the white Gaussian noise was 10 dB resulting in an iSNR of -0.41 dB. First, we designed ODMVDR and HDLCMV filters of length $M = 50$, and the resulting responses are shown in Fig. G.1b. The filter responses are close, and they both seem to extract the desired signal while attenuating both the sinusoidal noise, $v_{\text{sn}}(n)$, and the white noise, $v_{\text{wn}}(n)$. When we increase the filter order, the filters become almost equivalent, as can be seen from Fig. G.1d. This was also expected in the sinusoidal noise scenario according to Section 5.

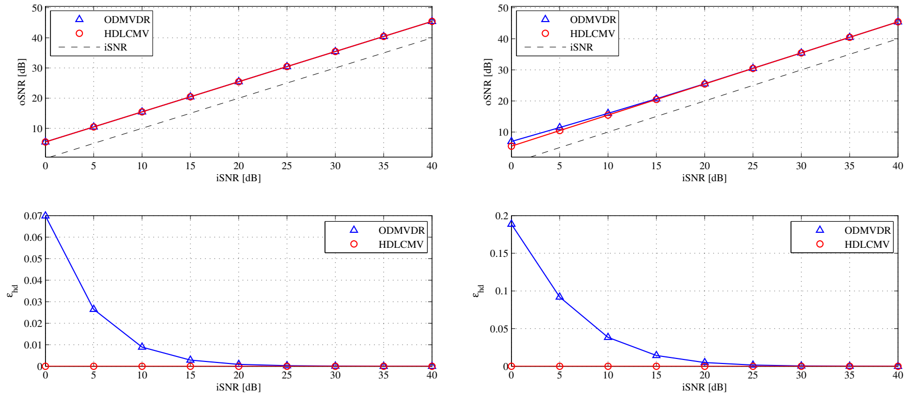
7.2 Evaluation of the Filter Performances

The second experiment was about evaluation of the performance of the ODMVDR and HDLCMV filters in different scenarios. The performance measures considered in this section were the output SNR and the harmonic distortion. As in the first experiment, this experiment was conducted with exact statistics, i.e., without synthetic data samples. In all simulations, the desired signal, $x(n)$, was a periodic signal containing $L = 6$ harmonic sinusoids. We conducted simulations with both unit amplitude harmonics ($A_l = 1$) and harmonics with decreasing amplitudes

$$\begin{bmatrix} A_1 & \cdots & A_6 \end{bmatrix}^T = \begin{bmatrix} 1 & 0.8 & 0.5 & 0.35 & 0.2 & 0.1 \end{bmatrix}^T . \tag{G.68}$$

By using decreasing amplitudes, we believe that we get a slightly better insight into the performance of the filters when the desired signal is speech which often has decreasing harmonic amplitudes. In all of the simulations in this experiment, the pitch of the desired signal was $\omega_0 = 0.245$.

First, we measured the performance of the two filters as a function of the iSNR. In this simulation, the filter length was $M = 30$, and the desired signal, $x(n)$, was corrupted by white Gaussian noise. For the scenario with unit amplitude harmonics, we obtained the results depicted in Fig. G.2a. Both filters improved the SNR by approximately 6 dB for all iSNRs. However, the ODMVDR filter had a little distortion of the harmonics at low iSNRs. For decreasing harmonic amplitudes, we got the results in Fig. G.2b. Note that in this scenario, the ODMVDR filter has a slightly higher oSNR than the HDLCMV filter at low iSNRs. However, the higher oSNR comes at the cost of distortion of the harmonics.



(a) (b)

Fig. G.2: Performance of the filters for $M = 30$ as a function of the iSNR when the harmonics has (a) unit amplitudes and (b) decreasing amplitudes, respectively, and the noise is white Gaussian.

Next, we compared the performance of the filters as a function of the filter length. In these simulations, the desired signal, $x(n)$, was corrupted by white Gaussian noise at an iSNR of 10 dB. First, the performance comparison was conducted for unit harmonic amplitudes resulting in the plot in Fig. G.3a. While the oSNRs of the filters are close, the ODMVDR filter has a little harmonic distortion. We also conducted the comparison for decreasing harmonic amplitudes as seen in Fig. G.3b. Here we see a larger difference in performance. For all filter lengths, the oSNR of the ODMVDR filter is greater than that of the HDLCMV filter. However, there is also some harmonic distortion introduced by the ODMVDR filter. Note that the step-wise increase in the oSNR in Fig. G.3a and Fig. G.3b is caused by the orthogonality (or the lack thereof) between the harmonics which is evident from (G.54) when the noise is white Gaussian.

Furthermore, we conducted simulations where the noise was a sum of white Gaussian noise, $v_{\text{wn}}(n)$, and sinusoidal noise, $v_{\text{sn}}(n)$. The variance, $\sigma^2_{v_{\text{sn}}}$, of the sinusoidal noise source was normalized with respect to the variance, σ^2_x, of the desired signal such that they had the same power. White Gaussian noise was also added to the desired signal resulting in the following iSNR

$$\text{iSNR} = \frac{\sigma^2_x}{\sigma^2_{v_{\text{sn}}} + \sigma^2_{v_{\text{wn}}}} . \tag{G.69}$$

Note that since the sinusoidal noise source has the same variance as the desired signal, the iSNR will always be smaller than or equal to zero (in dB) in these simulations according to the above equation. First, for the sinusoidal noise scenario, we compared the filter performances as a function of the iSNR when the filter order was $M = 50$.

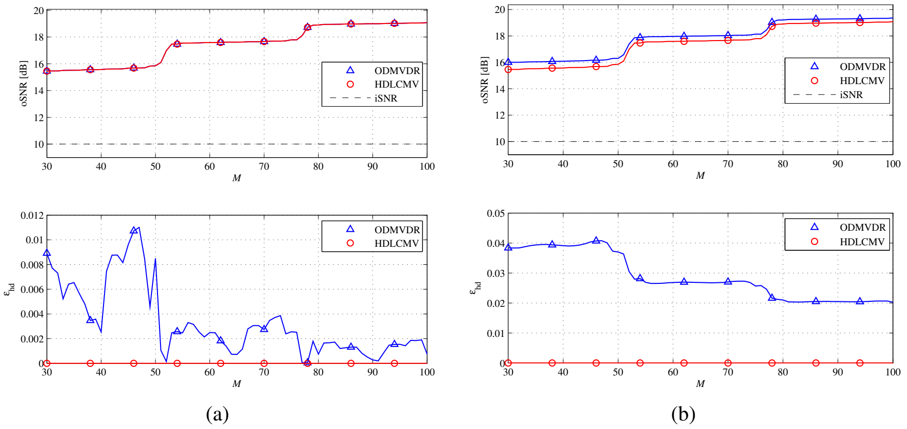
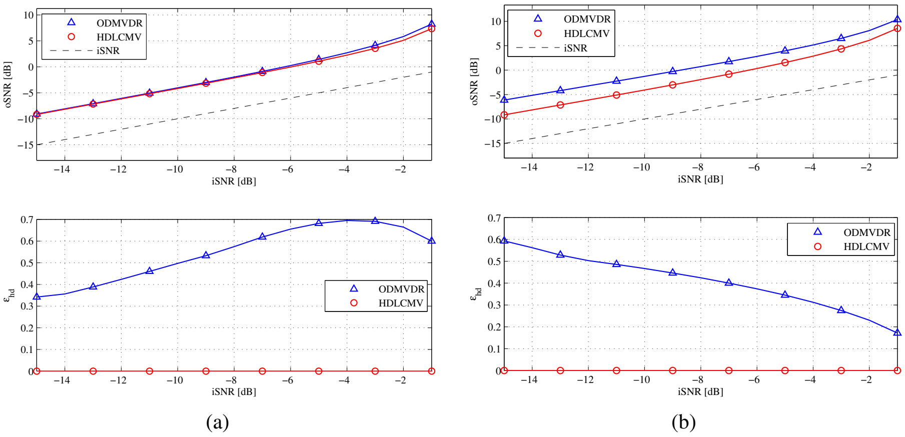


Fig. G.3: Performance of the filters as a function of M when the harmonics has (a) unit amplitudes and (b) decreasing amplitudes, respectively, and the noise is white Gaussian.

The result for unit harmonic amplitudes are given in Fig. G.4a. The oSNRs of the filters are relatively close, but with the largest difference when the white noise variance, $\sigma^2_{v_{\mathrm{wn}}}$, is largest. For all iSNRs, the ODMVDR filter has more harmonic distortion compared to the scenario with white Gaussian noise only. When decreasing harmonic amplitudes were considered (see Fig. G.4b), the difference in oSNRs between the filters was more pronounced with the ODMVDR having the highest oSNR for all iSNRs. The ODMVDR filter, however, also had more harmonic distortion in this case.

In the sinusoidal noise scenario, we also compared the performances as a function of the filter length, and the results are depicted in Fig. G.5a and Fig. G.5b, respectively. As in the previous simulations, we observe that the oSNR of the ODMVDR filter is in general higher than the oSNR of the HDLCMV filter. However, the difference between the filters decreases when M increases. The harmonic distortion of the ODMVDR filter is more significant in this simulation compared to the white Gaussian noise only scenario, but it decreases as we increase M.

Finally, we compared the filter performances as a function of the pitch spacing $\Delta\omega_0$ between the desired signal and the sinusoidal noise source. In this simulation, the filter order was $M = 100$. The results are given in Fig. G.6a and Fig. G.6b, respectively. For both unit and decreasing amplitudes, the oSNRs of the two filters are not much different for all source spacings, but with the ODMVDR having a slightly better oSNR. Moreover, for both filters the oSNR increases as we increase the spacing of the harmonic sinusoidal sources. We also observe that for both types of amplitudes, the ODMVDR has much harmonic distortion in this case compared to the other simulations.



(a)

(b)

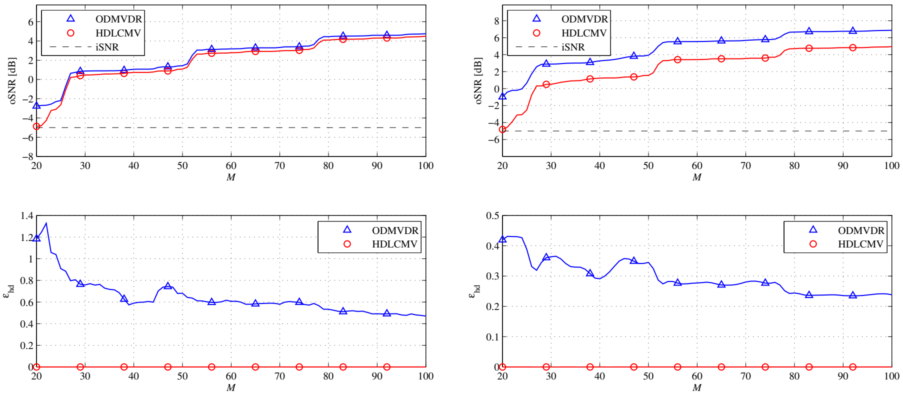
Fig. G.4: Performance of the filters for $M = 50$ as a function of the iSNR when the harmonics has (a) unit amplitudes and (b) decreasing amplitudes, respectively, and the noise is a sum of sinusoidal noise and white Gaussian noise.

7.3 Application Example: Using the ODMVDR and HDLCMV Filters Jointly for Speech Enhancement

In this experimental example, we show how the ODMVDR and HDLCMV can be applied jointly for enhancement of speech signals. For the experiment, we used a 2.2 second long speech segment sampled at 8 kHz. The segment contains a female speaker reading aloud the sentence “Why where you away a year Roy?” and it is plotted in Fig. G.7. Since the pitch is needed in the HDLCMV filter design, we estimated the pitch of the speech signal at all time instances using an orthogonality based subspace method [19, 21]. The pitch estimator is available from an online toolbox[1]. The pitch track resulting from the pitch estimation is also depicted in Fig. G.7, and it is used for later filter designs. Note that since we focus on speech enhancement rather than pitch estimation in this paper, we estimated the pitch directly from the clean speech signal, $x(n)$. The spectrogram of the speech signal, $x(n)$, is shown in Fig. G.8a.

First, we consider a scenario in which the speech signal is corrupted by babble noise at an average iSNR of 5 dB. The babble noise was taken from the AURORA database [28]. The spectrogram of the noisy signal is depicted in Fig. G.8b. We then enhanced the noisy signal using three different filtering setups, i.e., using the ODMVDR filter only, using the HDLCMV filter only, and using the ODMVDR and HDLCMV filters jointly. The joint filtering method is proposed since using only either the ODMVDR or the HDLCMV filter has drawbacks. For example, the ODMVDR method is sensitive to

[1] http://www.morganclaypool.com/page/multi-pitch

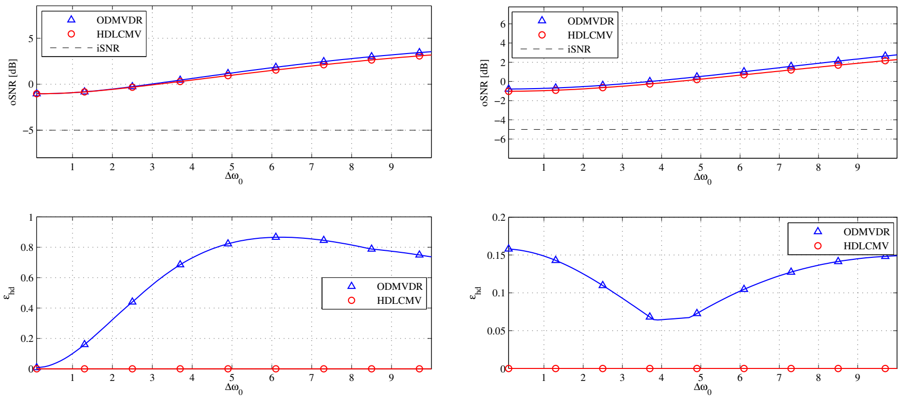


(a) (b)

Fig. G.5: Performance of the filters as a function of M when the harmonics has (a) unit amplitudes and (b) decreasing amplitudes, respectively, and the noise is a sum of sinusoidal noise and white Gaussian noise.

non-stationary noise, since it requires that knowledge about the noise statistics which we do not always have access to in practice. This is not an issue for the HDLCMV filter, but, on the other hand, it will introduce some distortion of speech signals because the harmonic model does not hold exactly. Furthermore, the HDLCMV filter has, in general, more constraints than the ODMVDR filter, and it will therefore most likely have a lower oSNR compared to the ODMVDR filter. The joint use of the filters can be justified by their close relationship described in Section 5. In the joint filtering scheme, we first use the HDLCMV filter to obtain a rough estimate of the speech signal. The rough speech estimate is then subtracted from the observed signal to obtain an estimate of the noise signal. We estimate the noise statistics from the estimated noise signal, and the noise statistics are used for designing the ODMVDR filter. Finally, the ODMVDR filter is applied for enhancement of the observed signal. By using the ODMVDR filter for the enhancement rather than the HDLCMV filter, we expect to remove some of the distortion introduced by the HDLCMV filter in practice. Moreover, we expect to obtain more noise reduction, since the ODMVDR filter is less constrained compared to the HDLCMV filter.

In all the filtering setups, the filters were updated for each time instance. The update was conducted by recalculating the filters from the signal and noise statistics ($\hat{\mathbf{R}}_{\mathbf{y}}$ and $\hat{\mathbf{R}}_{\mathbf{v}}$) estimated from the previous 400 samples (≈ 50 ms). Both $\hat{\mathbf{R}}_{\mathbf{y}}$ and $\hat{\mathbf{R}}_{\mathbf{v}}$ were used to calculate the ODMVDR filter. That is, we assumed that the noise signal was available in this simulation, albeit it is not the case in practice. The HDLCMV filter was updated using $\hat{\mathbf{R}}_{\mathbf{y}}$, the pitch estimates in Fig. G.7, and a model order of $L = 13$. The model



(a) (b)

Fig. G.6: Performance of the filters for $M = 100$ as a function of the source spacing $\Delta\omega_0$ when the harmonics has (a) unit amplitudes and (b) decreasing amplitudes, respectively, and the noise is a sum of sinusoidal noise and white Gaussian noise.

order was chosen by inspecting the spectrogram in Fig. G.8a since we do not consider model order estimation in this paper. Furthermore, in the calculations of the HDLCMV filter and the filters in the joint filtering setup, we regularized the covariance matrix using [29]

$$\hat{\mathbf{R}}_{\mathbf{y},\text{reg}} = (1-\gamma)\hat{\mathbf{R}}_{\mathbf{y}} + \gamma \frac{\text{Tr}\left\{\hat{\mathbf{R}}_{\mathbf{y}}\right\}}{M}\mathbf{I}\,, \tag{G.70}$$

where $\text{Tr}\{\cdot\}$ denotes the trace operator. The regularization is used to compensate for, e.g., numerical stability, model mismatch, and noisy statistics. Choosing $\gamma = 0.7$ was found to give the best results in terms of oSNR and perceptual scores. All filters were chosen to be of order $M = 100$.

The observed signal containing the speech signal and babble noise was then enhanced using the three filtering setups, and the spectrograms of the resulting enhanced signals are shown in Fig. G.9. The spectrograms indicate that the joint filtering method has better noise reduction abilities than when using either the ODMVDR or the HDLCMV filter only. Regarding distortion, the ODMVDR filter seems to outperform the joint filtering method. However, it is important to remember that the ODMVDR filter was designed using the noise signal, and it will therefore most likely have a worse performance in practice. To confirm the observations on the performances of the filters,

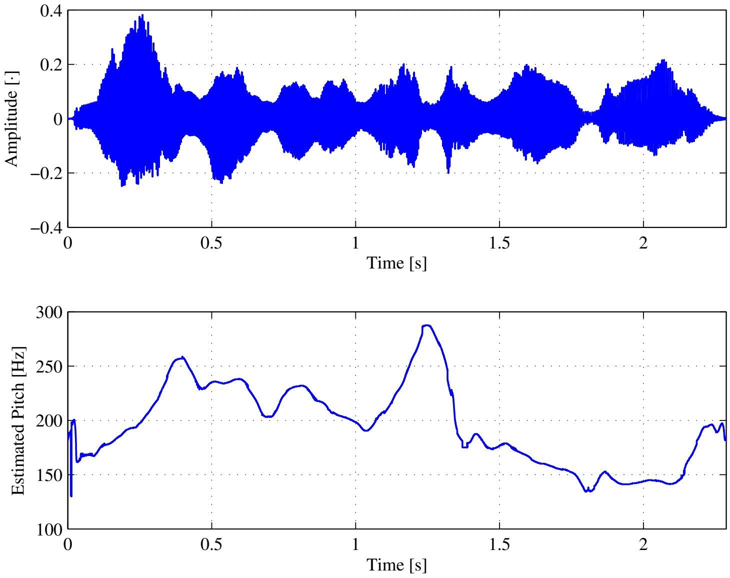


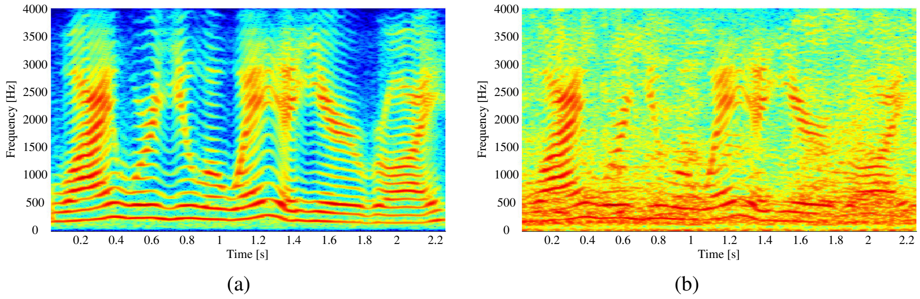
Fig. G.7: A plot of a female speech signal (top) and the pitch estimates associated with it (bottom).

we also measured the oSNRs associated with the enhanced signals in Fig. G.9 using

$$\text{oSNR}(\mathbf{h}) = \frac{\sigma_{x_\text{f}}^2}{\sigma_{v_\text{rn}}} = \frac{\mathbf{h}^T \mathbf{R}_\mathbf{x} \mathbf{h}}{\mathbf{h}^T \mathbf{R}_\mathbf{v} \mathbf{h}} \,. \tag{G.71}$$

Note that we here use the traditional oSNR measure, since, in practice, the interference term of the ODMVDR approach is relatively large which complicates the comparison of the oSNR measures in (G.51) and (G.53), respectively. The measured oSNRs are shown in Fig. G.10. These measurements show that both the ODMVDR and the joint filtering methods outperform the HDLCMV filtering method in terms of noise reduction. The ODMVDR and joint filtering methods have comparable noise reduction performance even though the joint filtering method is implemented without access to the noise signal directly. This justifies the use of the joint filtering method in practice as it is more tractable than the ODMVDR filtering method when the noise signal is not available.

The oSNR measure, however, does not quantify how much the filtering methods distort the desired signal. Therefore, we also evaluated the filtering methods in terms of "Perceptual Evaluation of Speech Quality" (PESQ) scores [30]. The PESQ score is an objective measure which reflects the perceptual quality of a speech signal. That is, the PESQ scores give a more complete picture of the performance of the filtering methods since the perceptual quality is affected both by noise reduction and distortion.



(a)

(b)

Fig. G.8: Spectrograms of (a) the clean speech signal in Fig. G.7 and (b) the speech signal in Fig. G.7 corrupted by babble noise at an iSNR of 5 dB.

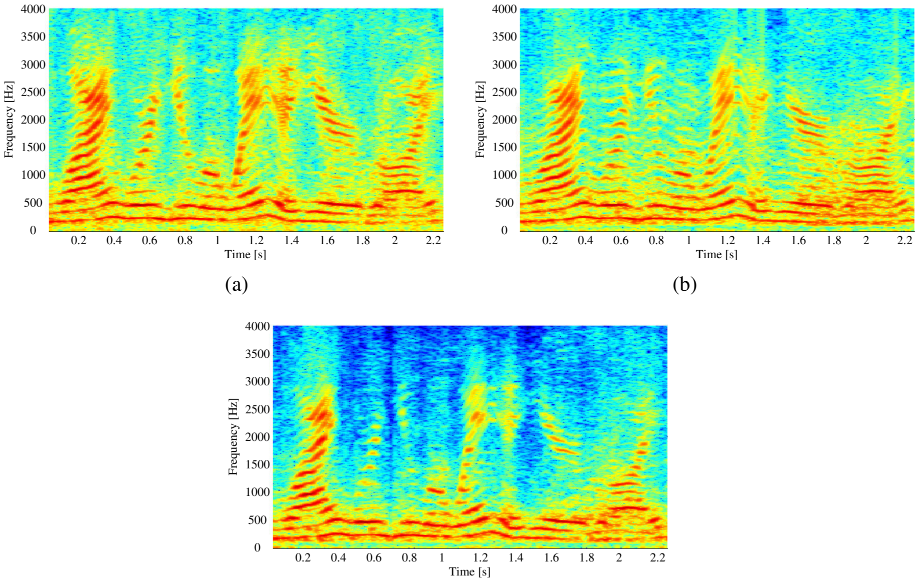
We compared the PESQ scores of noisy speech signal enhanced using the joint filtering method, the ODMVDR filtering method, the HDLCMV filtering method, a spectral subtraction based method [31], and a method using MMSE estimates of the spectral amplitudes [32]. Note that, in these simulations, we design the ODMVDR filter from the true noise signal, and it therefore only serves as a bound to the proposed joint filtering scheme.

Followingly, we describe how the different enhancement methods were set up for the PESQ score evaluations. In all of the filtering methods, i.e., the joint method, the ODMVDR method, and the HDLCMV method, the observed signal and noise statistics were calculated as in the previous experiment. The noise statistics were calculated directly from the noise signal, and they were only used for designing the ODMVDR filter. In the joint and HDLCMV filtering methods, the observed signal statistics were regularized as in the previous experiment. The model order was set to $L = \min([15, \lfloor \pi/\omega_0 \rfloor - 1])$ at each time instance when designing the HDLCMV filters. The speech signals used in these evaluations contained both voiced and unvoiced speech segments. However, the HDLCMV filter used in both the joint and HDLCMV filtering methods are designed for voiced speech segments only. Therefore, we updated the HDLCMV filter in these evaluations as follows; for voiced speech segments, the HDLCMV filter was designed as in (G.36), and for unvoiced speech segments, the filter was updated as

$$\mathbf{h}(n) = (1-\gamma)\mathbf{0} + \gamma\mathbf{h}(n-1) \ , \tag{G.72}$$

when $\|\mathbf{h}(n-1)\|_2 > 0.1$ with $\gamma = 0.95$ and $\mathbf{0}$ is a vector of zeros. The norm conditional update was introduced to avoid abrupt changes when transitioning between unvoiced/no speech and voiced speech. Both the spectral subtraction and the MMSE-based methods are available in the VOICEBOX toolbox[2] for MATLAB, in which they are implemented

[2] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html



(a)

(b)

(c)

Fig. G.9: Spectrograms of enhanced versions of the noisy signal in Fig. G.8b. The enhanced signals are obtained using (a) the ODMVDR filter only, (b) the HDLCMV filter only, and (c) the joint HDLCMV and ODMVDR filtering setup, respectively.

using noise power spectral density estimates based on optimal smoothing and minimum statistics [33]. We used the default settings given by the VOICEBOX toolbox for the spectral subtractions and MMSE methods.

For the PESQ score evaluations of the aforementioned enhancement methods, we used two female and two male speech excerpts each of length 4-6 seconds taken from the Keele database [34]. Since pitch estimation is not the main topic of this paper, we used the pitch estimates of the voiced parts of the speech excerpts from the Keele database for the design of the HDLCMV filters. Moreover, the pitch estimates in the Keele database are 0 when the speech is unvoiced or no voice is present. We exploited this to distinguish between voiced and unvoiced speech since the unvoiced/voiced speech detection problem is not considered here. The chosen speech excerpts were then buried in white Gaussian noise, car noise, babble noise, exhibition hall noise, and street noise. All noise sources except the white noise were taken from the AURORA database [28]. First, we applied the proposed joint filtering method on all four speech excerpts in all five noise scenarios for different filtering lengths when the iSNR was 5 dB. The PESQ scores averaged across the different noisy speech excerpts are shown in Fig. G.11a. We

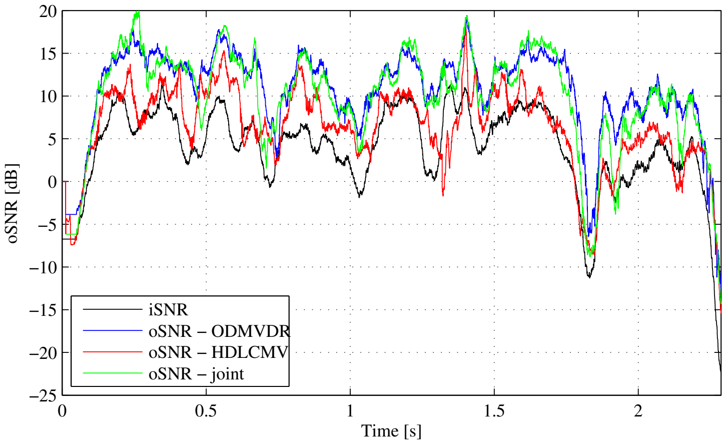
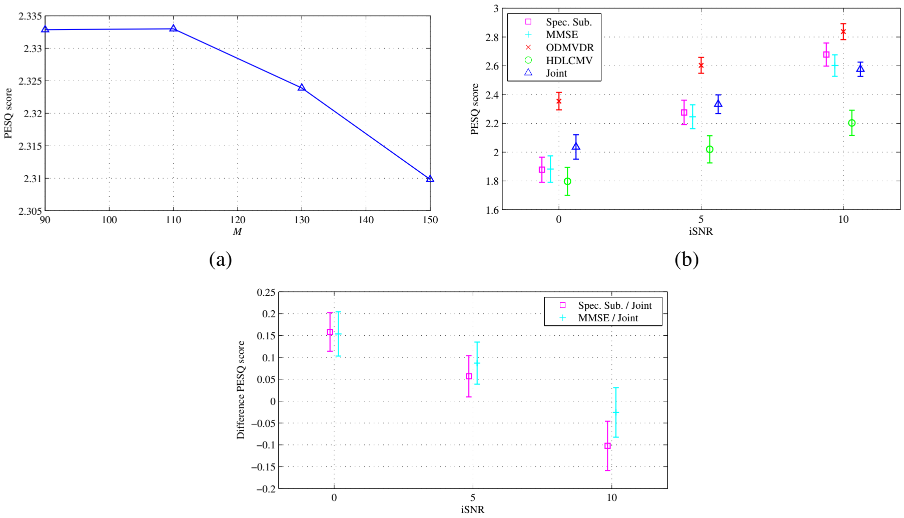


Fig. G.10: The estimated iSNR and oSNRs over time for the enhanced signals in Fig. G.9.

can see that the perceptual performance of the proposed joint filtering method peaks around $M = 110$. We then applied all of the enhancement methods of the comparison on all the speech excerpts in all of the different noise scenarios for different iSNRs. For these simulations, the filter length of the filtering-based enhancements methods was set to 110, and the PESQ results averaged over the different speech excerpts and noise scenarios are shown in Fig. G.11b with 95% confidence intervals. From these results, it seems that the joint filtering method outperforms the spectral subtraction and MMSE-based methods on average for relative low iSNRs (≤ 5 dB) and vice versa for a higher iSNR (10 dB). However, from these results, we cannot say this with 95% confidence due to overlapping confidence intervals, but it does not preclude that the observations are statistically significant since we can also consider the difference in PESQ scores. To investigate this further, we measured the average of the difference in PESQ scores between the proposed joint filtering scheme and the spectral subtraction and MMSE-based methods, respectively; the results from this investigation is plotted in G.11c with 95% confidence intervals. From these results, we can conclude with 95% confidence that the proposed joint filtering method outperforms the spectral subtraction and MMSE-based methods on average for iSNRs of 0 dB and 5 dB in terms of PESQ scores since the confidence intervals do not include 0. In practice, it is expected that the proposed joint filtering method only outperforms the other methods for relatively low iSNRs since the harmonic model assumption embedded in the proposed joint filtering design introduces a small amount of distortion due to model mismatch.



(a)

(b)

(c)

Fig. G.11: Average PESQ scores (a) for the joint filtering scheme as a function of M for an iSNR of 5 dB, and (b) for several enhancement methods as a function of the iSNR for $M = 110$ with 95% confidence intervals. In (c), the average differences in PESQ scores between the joint filtering scheme and the spectral subtraction and MMSE-based methods, respectively, are plotted with 95% confidence intervals.

8 Conclusion

In this paper, we considered two recent filter designs for speech enhancement, namely the ODMVDR and HDLCMV filters. The ODMVDR filter is not explicitly dependent of the desired signal since it is calculated from the observed signal and noise statistics. This makes it a general filtering method which is appropriate for enhancement of all types of speech (e.g., both voiced and unvoiced). However, the ODMVDR filter is vulnerable to non-stationary noise since the noise statistics are typically estimated during periods of silence. On the other hand, the HDLCMV filter is signal-dependent since it is designed using the observed signal and the desired signal statistics. In this filter, a harmonic model is assumed which enables the estimation of the signal statistics if the pitch and the number of harmonics are known. While this filter is robust against non-stationary noise, it will only be appropriate for voiced speech due to the harmonic model assumption. Since both filters have complementary advantages and disadvantages, we investigated the relationship between them in this paper. Our theoretical studies confirmed that the filters are indeed closely related. We also proposed some performance measures for both filters which are available in closed-form when the desired signal is periodic. We compared the performance measures in theoretical simulations. From these simulations, it was again clear that the methods are closely related, but each filter had its own advantages. For example, the ODMVDR filter has, in general, a slightly higher oSNR than the HDLCMV while the HDLCMV filter does not distort the harmonics as opposed to the ODMVDR filter. The close relationship between the filters inspired us to propose a filtering scheme where the ODMVDR and HDLCMV filters are used jointly. This scheme was applied on real speech signals in different noise scenarios. The results of these experiments showed that, for relatively low iSNRs (i.e., < 10 dB) , the joint filtering scheme outperforms some existing enhancement techniques in terms of average PESQ scores with 95% confidence.

9 Appendix: On Rewriting the HDLCMV Filter in Terms of the Observed Signal Covariance Matrix

In this appendix, we show that it makes no difference whether we use the noise covariance matrix, $\mathbf{R_v}$, or use the observed signal covariance matrix, $\mathbf{R_y}$, in (G.35). First, recall that the HDLCMV filter is given by

$$\mathbf{h}_{\text{HDLCMV}} = \mathbf{R}_\mathbf{v}^{-1}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{R}_\mathbf{v}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1}\ . \tag{G.73}$$

Note that in the following derivations we denote the HDLCMV filter as $\mathbf{h}$. If we use the covariance matrix model on $\mathbf{R_y}$, the noise covariance matrix can also be written as [24]

$$\mathbf{R_v} = \mathbf{R_y} - \mathbf{ZPZ}^H\ . \tag{G.74}$$

If we substitute (G.74) back into (G.73), we get that

$$\begin{aligned}\mathbf{h} &= \left(\mathbf{R_y} - \mathbf{ZPZ}^H\right)^{-1}\mathbf{Z}\left[\mathbf{Z}^H\left(\mathbf{R_y} - \mathbf{ZPZ}^H\right)^{-1}\mathbf{Z}\right]^{-1}\mathbf{1}\\ &= \mathbf{AZB1}\ ,\end{aligned} \tag{G.75}$$

where

$$\mathbf{A} = \left(\mathbf{R_y} - \mathbf{ZPZ}^H\right)^{-1}\ , \tag{G.76}$$

$$\begin{aligned}\mathbf{B} &= \left[\mathbf{Z}^H\left(\mathbf{R_y} - \mathbf{ZPZ}^H\right)^{-1}\mathbf{Z}\right]^{-1}\\ &= \left(\mathbf{Z}^H\mathbf{AZ}\right)^{-1}\ .\end{aligned} \tag{G.77}$$

Applying the matrix inversion lemma on $\mathbf{A}$ yields

$$\mathbf{A} = \mathbf{R}_\mathbf{y}^{-1} + \mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}\left(\mathbf{P}^{-1} - \mathbf{Z}^H\mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{R}_\mathbf{y}^{-1}\ . \tag{G.78}$$

If we insert this expression for $\mathbf{A}$ back into (G.77), we get

$$\mathbf{B} = \left(\mathbf{Z}^H \mathbf{R}_\mathbf{y}^{-1} \mathbf{Z}\right)^{-1} - \mathbf{P} \,. \tag{G.79}$$

We can then rewrite the HDLCMV filter expression by inserting (G.78) and (G.79) into (G.75) which yields

$$\begin{aligned} \mathbf{h} &= \mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1} - \mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}\mathbf{P}\mathbf{1} \\ &+\mathbf{R}_\mathbf{y}^{-1}\left(\mathbf{P}^{-1} - \mathbf{Z}^H\mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1} \\ &-\mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}\left(\mathbf{P}^{-1} - \mathbf{Z}^H\mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}\mathbf{P}\mathbf{1} \,. \end{aligned} \tag{G.80}$$

After some algebra, it turns out that the somewhat complex expression for the filter in (G.80) can be reduced to

$$\mathbf{h} = \mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1} \,. \tag{G.81}$$

That is, there is no difference between using the noise covariance matrix, $\mathbf{R}_\mathbf{v}$, and the observed signal covariance matrix, $\mathbf{R}_\mathbf{y}$, in (G.73).

References

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

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

[3] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.

[4] R. J. McAulay and M. L. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[5] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[6] M. Dendrinos, S. Bakamidis, and G. Carayannis, “Speech enhancement from noise: A regenerative approach,” *Speech Communication*, vol. 10, no. 1, pp. 45 – 57, 1991.

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

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

[9] J. S. Lim, Ed., *Speech Enhancement*. Prentice-Hall, 1983.

[10] J. Chen, J. Benesty, and Y. Huang, "Study of the noise-reduction problem in the Karhunen-Loève expansion domain," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 4, pp. 787–802, May 2009.

[11] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. M.I.T. Press, 1949.

[12] 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.

[13] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Elsevier Signal Process.*, vol. 80(9), pp. 1937–1944, 2000.

[14] K. W. Chan and H. C. So, "Accurate frequency estimation for real harmonic sinusoids," *IEEE Signal Process. Lett.*, vol. 11, no. 7, pp. 609–612, 2004.

[15] 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.

[16] V. Emiya, B. David, and R. Badeau, "A parametric method for pitch estimation of piano tones," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 2007, pp. 249–252.

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

[18] 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.

[19] 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, no. 5, pp. 1635–1644, Jul. 2007.

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

[21] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[22] M. G. Christensen, J. L. Højvang, A. Jakobsson, and S. H. Jensen, “Joint fundamental frequency and order estimation using optimal filtering,” *EURASIP J. on Applied Signal Processing*, vol. 2011, no. 1, pp. 1–18, Jun. 2011.

[23] J. Benesty and J. Chen, *Optimal Time-Domain Noise Reduction Filters – A Theoretical Study*, 1st ed. Springer, 2011, no. VII.

[24] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Education, Inc., 2005.

[25] J. Capon, “High-resolution frequency-wavenumber spectrum analysis,” *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[26] ——, *Nonlinear Methods of Spectral Analysis*. Springer-Verlag, 1983, ch. Maximum-Likelihood Spectral Estimation.

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

[28] 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.

[29] F. van der Heijden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax, *Classification, Parameter Estimation and State Estimation - An Engineering Approach using MATLAB®*. John Wiley & Sons Ltd, 2004.

[30] ITU-T, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” no. P.862, pp. 1–30, Feb. 2001.

[31] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, 1979, pp. 208–211.

[32] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[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] F. Plante, G. F. Meyer, and W. A. Ainsworth, “A pitch extraction reference database,” in *Proc. Eurospeech*, Sep. 1995, pp. 837–840.

Paper H

Joint Filtering Scheme for Nonstationary Noise Reduction

Jesper Rindom Jensen, Jacob Benesty, Mads Græsbøll Christensen and Søren Holdt Jensen

The paper has been accepted for publication in the
Proc. European Signal Processing Conference, 2012.

The layout has been revised.

Abstract

In many state-of-the-art filtering methods for speech enhancement, an estimate of the noise statistics is required. However, the noise statistics are difficult to estimate when speech is present and, consequently, nonstationary noise has a detrimental impact on the performance of most noise reduction filters. We propose a joint filtering scheme for speech enhancement which supports the estimation of the noise statistics even during voice activity. First, we use a pitch driven linearly constrained minimum variance (LCMV) filter to estimate the noise statistics. A Wiener filter is then designed based on the estimated noise statistics, and it is applied for the noise reduction of the speech. In experiments involving real signals, we show that the proposed filtering scheme outperforms other existing speech enhancement methods in terms of perceptual evaluation of speech quality (PESQ) scores in different nonstationary noise scenarios.

1 Introduction

Speech is frequently encountered in numerous signal processing applications such as telecommunications, teleconferencing, hearing-aids, and human-machine interfaces. The speech picked up by a microphone can be very noisy. Unfortunately, the noise will degrade the speech quality and intelligibility which, eventually, has a detrimental impact on speech applications. It is therefore highly relevant to develop methods for reducing the noise. In this paper, we consider filtering methods for noise reduction of single-channel speech recordings. Several such methods have been developed in the past decades. For an overview of such methods, we refer to [1] and the references therein. Many existing noise reduction filtering methods assume that the noise signal is directly available since they rely on the noise statistics. This is, of course, not the case in practice, so the noise statistics could, for example, be estimated when there is no voice activity. Some alternative methods based on, e.g., harmonic tunneling [2] and minimum statistics [3] have been proposed for estimating the noise statistics during speech presence.

In this paper, we propose a novel joint filtering scheme for nonstationary noise reduction of noisy quasi-periodic signals such as voiced speech. It is well-known that speech can be both voiced and unvoiced, so the proposed filtering scheme has to be combined with voiced/unvoiced speech detection (see, e.g., [4, 5]) when applied to speech enhancement. In the proposed scheme, we utilize two recently proposed filters, namely the orthogonal decomposition based Wiener (ODW) filter and the harmonic decomposition based linearly constrained minimum variance (HDLCMV) filter [6, 7]. Followingly, the proposed filtering scheme is described. First, we use the HDLCMV filter to obtain a rough estimate of the desired signal. In the HDLCMV filter, it is assumed that the desired signal is quasi-periodic and thereby has a harmonic structure which is a reasonable assumption for the voiced parts of speech signals. Therefore, the

HDLCMV filter is designed using the pitch, the number of harmonics, and the statistics of the observed signal, i.e., this filter does not rely on noise statistics. Pitch and model order estimation is not considered in this paper, but there exists a multitude of methods for this (see, e.g., [7] and the references therein). From the rough estimate of the desired signal, we obtain an estimate of the noise signal. That is, using this approach, we can easily estimate the noise statistics even when speech is present. The estimated noise statistics are used to design the ODW filter which, finally, performs the noise reduction of the observed speech signal. Besides proposing the joint filtering scheme, we also provide a few important closed-form performance measure expressions for the filters under the assumption that the desired signal is quasi-periodic.

The remainder of the paper is organized as follows. In Section 2, we introduce the signal model used in the paper, the problem of designing noise reduction filters, and the orthogonal and harmonic decompositions. Based on this, we derive the optimal ODW and HDLCMV filters in Section 3. We then propose a joint filtering scheme for noise reduction and evaluate its performance in Section 4. Finally, in Section 5, we conclude on the paper.

2 Signal Model

In this paper, we consider nonstationary noise reduction of single-channel speech recordings using filtering. The noise reduction problem is to extract a zero-mean desired signal, $x(n)$, from a mixture signal

$$y(n) = x(n) + v(n) \ , \tag{H.1}$$

where $v(n)$ is a zero-mean noise source, and n is the discrete time index. The noise source is assumed to be uncorrelated with the desired signal. Moreover, in some parts of the paper, we assume that the desired signal is quasi-periodic which is indeed a reasonable assumption for, e.g., voiced speech. When the desired signal is quasi-periodic, we can rewrite the signal model in (H.1) as

$$y(n) = \sum_{l=1}^{L} \left(a_l e^{jl\omega_0 n} + a_l^* e^{-jl\omega_0 n} \right) + v(n) \ , \tag{H.2}$$

where ω_0 is the fundamental frequency (aka the pitch), L is the model order, $a_l = \frac{A_l}{2} e^{j\phi_l}$ is the complex amplitude of the lth harmonic, A_l is the real amplitude of the lth harmonic, ϕ_l is the random phase of the lth harmonic, and $(\cdot)^*$ denotes the complex conjugation. Many real-life signals, however, have some degree of inharmonicity. The problem of inharmonicity is not considered in this paper, yet several methods dealing with it exist (see, e.g., [7] and the references therein).

When designing optimal filters for noise reduction, we need several consecutive samples of the observed signal, $y(n)$. Therefore, we use the vector signal model given by

$$\mathbf{y}(n) = \mathbf{x}(n) + \mathbf{v}(n) \ , \tag{H.3}$$

where

$$\mathbf{y}(n) = \begin{bmatrix} y(n) & y(n-1) & \cdots & y(n-M+1) \end{bmatrix}^T \ , \tag{H.4}$$

with $(\cdot)^T$ denoting the transpose of a vector or matrix, M is the number of samples, and the definitions of $\mathbf{x}(n)$ and $\mathbf{v}(n)$ follow that of $\mathbf{y}(n)$. We know by assumption that the desired signal and the noise are uncorrelated. Therefore, we can obtain the following simple expression for the covariance matrix,

$$\mathbf{R}_\mathbf{y} = \mathrm{E}[\mathbf{y}(n)\mathbf{y}^T(n)] = \mathbf{R}_\mathbf{x} + \mathbf{R}_\mathbf{v} \ , \tag{H.5}$$

of the observed signal where $\mathrm{E}[\cdot]$ is the mathematical expectation operator, $\mathbf{R}_\mathbf{x} = \mathrm{E}[\mathbf{x}(n)\mathbf{x}^T(n)]$ is the covariance matrix of $\mathbf{x}(n)$, and $\mathbf{R}_\mathbf{v} = \mathrm{E}[\mathbf{v}(n)\mathbf{v}^T(n)]$ is the covariance matrix of $\mathbf{v}(n)$. When the desired signal is quasi-periodic, we can model the covariance matrix of $\mathbf{x}(n)$ as

$$\mathbf{R}_\mathbf{x} \approx \mathbf{Z}\mathbf{P}\mathbf{Z}^H \ , \tag{H.6}$$

where $(\cdot)^H$ denotes the complex conjugate transpose of a matrix or vector, and

$$\mathbf{P} = \mathrm{diag}\left(\begin{bmatrix} |a_1|^2 & |a_1^*|^2 & \cdots & |a_L|^2 & |a_L^*|^2 \end{bmatrix}\right) \ , \tag{H.7}$$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}(\omega_0) & \mathbf{z}^*(\omega_0) & \cdots & \mathbf{z}(L\omega_0) & \mathbf{z}^*(L\omega_0) \end{bmatrix}, \tag{H.8}$$

$$\mathbf{z}(l\omega_0) = \begin{bmatrix} 1 & e^{-jl\omega_0} & \cdots & e^{-jl\omega_0(M-1)} \end{bmatrix}^T \ , \tag{H.9}$$

with $\mathrm{diag}(\cdot)$ denoting the construction of a diagonal matrix from a vector.

The goal in noise reduction filtering methods is to design a filter which extracts one or more samples of the desired signal $x(n)$ from $\mathbf{y}(n)$. That is, the filter should attenuate the noise $v(n)$ as much as possible while not distorting the desired signal too much. In this paper, we focus on optimal filtering methods for extraction of a single sample of $x(n)$. A filtering operation which estimates $x(n)$ from $\mathbf{y}(n)$ can be written as

$$\hat{x}(n) = \sum_{m=0}^{M-1} h_m y(n-m) = \mathbf{h}^T\mathbf{y}(n) \ , \tag{H.10}$$

where $\mathbf{h} = \begin{bmatrix} h_0 & h_1 & \cdots & h_{M-1} \end{bmatrix}^T$, and $\hat{x}(n)$ is an estimate of $x(n)$. The main difference between optimal filtering methods for noise reduction is how the desired signal is decomposed. In this paper, we consider the orthogonal and harmonic decompositions [6, 7].

In the orthogonal decomposition, the signal vector $\mathbf{x}(n)$ is decomposed into two parts being proportional and orthogonal to $x(n)$, respectively. That is, using this decomposition, $\mathbf{x}(n)$ can also be written as

$$\mathbf{x}(n) = x(n)\boldsymbol{\rho}_{\mathbf{x}x} + \mathbf{x}_{\text{i}}(n) = \mathbf{x}_{\text{d}}(n) + \mathbf{x}_{\text{i}}(n) \ , \tag{H.11}$$

where

$$\boldsymbol{\rho}_{\mathbf{x}x} = \frac{\text{E}[\mathbf{x}(n)x(n)]}{\text{E}[x^2(n)]} \tag{H.12}$$

is the normalized correlation vector between $\mathbf{x}(n)$ and $x(n)$. If we insert (H.11) into (H.10), we get

$$\begin{aligned}\hat{x}_{\text{OD}}(n) &= \mathbf{h}^T[\mathbf{x}_{\text{d}}(n) + \mathbf{x}_{\text{i}}(n) + \mathbf{v}(n)] \\ &= x_{\text{fd}}(n) + x_{\text{ri}}(n) + v_{\text{rn}}(n) \ ,\end{aligned} \tag{H.13}$$

where $x_{\text{fd}}(n) = \mathbf{h}^T\mathbf{x}_{\text{d}}(n)$ is the filtered desired signal, $x_{\text{ri}}(n) = \mathbf{h}^T\mathbf{x}_{\text{i}}(n)$ is the residual interference, and $v_{\text{rn}}(n) = \mathbf{h}^T\mathbf{v}(n)$ is the residual noise. Using the orthogonal decomposition, we can define the following error signal

$$e_{\text{OD}}(n) = x(n) - [x_{\text{fd}}(n) + x_{\text{ri}}(n) + v_{\text{rn}}(n)] \ . \tag{H.14}$$

Optimal noise reduction filters based on the orthogonal decomposition can then, for example, be derived by minimizing $e(n)$ or parts of $e(n)$ subject to some constraints. Most commonly, the error is minimized in the mean-square error (MSE) sense. Clearly, this design procedure ensures that the aforementioned design goals are fulfilled.

In the harmonic decomposition approach, it is assumed that the desired signal is quasi-periodic which makes it useful for signals produced by voiced speech and musical instruments [7, 8]. Due to this assumption, the signal vector, $\mathbf{x}(n)$, can be written as

$$\mathbf{x}(n) = \mathbf{Z}\mathbf{a}(n) = \mathbf{x}'_{\text{d}}(n) \ , \tag{H.15}$$

where

$$\begin{aligned}\mathbf{a}(n) = \big[a_1 e^{j\omega_0 n} \ \ a_1^* e^{-j\omega_0 n} \ \ \cdots \\ a_L e^{jL\omega_0 n} \ \ a_L^* e^{-jL\omega_0 n}\big]^T \ .\end{aligned} \tag{H.16}$$

From the above expression, we can see that there is no interference in this decomposition as opposed to in the orthogonal decomposition. This is because all information in $\mathbf{x}(n)$ can be used to describe the desired signal when we know the signal model. We can obtain an estimate of $x(n)$ using a harmonic decomposition filter by inserting (H.15) into (H.10). This yields

$$\hat{x}_{\text{HD}}(n) = \mathbf{h}^T[\mathbf{x}'_{\text{d}}(n) + \mathbf{v}(n)] \ . \tag{H.17}$$

We define the following error function for the harmonic decomposition approach to filter design

$$e_{\text{HD}}(n) = x(n) - [x'_{\text{fd}}(n) + v_{\text{rn}}(n)] \ , \tag{H.18}$$

where $x'_{\text{fd}}(n) = \mathbf{h}^T\mathbf{x}'_{\text{d}}(n)$. We can then design a harmonic decomposition based filter for noise reduction by minimizing the effects of $e_{\text{HD}}(n)$ or parts of $e_{\text{HD}}(n)$ perhaps subject to some constraints (e.g., to avoid undesired distortion).

3 Optimal Filters

In this section, we derive the ODW filter and the HDLCMV filter. Furthermore, we provide expressions for the filters and some of their performance measures; the performance measure expressions are closed-form when the desired signal is periodic.

3.1 Orthogonal Decomposition Wiener

The ODW filter is found by minimizing $\text{E}\{|e_{\text{OD}}(n)|^2\}$ with respect to the unknown filter response. This yields

$$\mathbf{h}_{\text{W}} = \sigma_x^2\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\rho}_{\mathbf{x}x} \ , \tag{H.19}$$

where σ_x^2 is the variance of $x(n)$. When the desired signal is periodic, we can also write the normalized correlation vector, $\boldsymbol{\rho}_{\mathbf{x}x}$, as

$$\boldsymbol{\rho}_{\mathbf{x}x} = \frac{\mathbf{R}_{\mathbf{x}}\mathbf{i}}{\mathbf{i}^T\mathbf{R}_{\mathbf{x}}\mathbf{i}} = \frac{\mathbf{Z}\mathbf{P}\mathbf{1}}{\sigma_x^2} \ , \tag{H.20}$$

where $\mathbf{1} = \begin{bmatrix}1 & \cdots & 1\end{bmatrix}^T$ and $\mathbf{i}$ is the first column of the $M \times M$ identity matrix. That is, for periodic signals, the OD Wiener filter is given by

$$\mathbf{h}_{\text{W}} = \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}\mathbf{P}\mathbf{1} \ . \tag{H.21}$$

The output signal-to-noise ratio (oSNR) of an orthogonal decomposition based filter is defined as the ratio between the variance of the filtered desired signal and the sum of the variances of the residual interference and noise [6]. It can be shown that the ODW filter achieves the maximum output SNR [6]. The output SNR of the ODW filter for periodic signals therefore equals

$$\text{oSNR}^{\text{OD}}(\mathbf{h}_{\text{W}}) = \frac{\mathbf{1}^T\mathbf{P}\mathbf{Z}^H\mathbf{R}_{\text{in}}^{-1}\mathbf{Z}\mathbf{P}\mathbf{1}}{\sigma_x^2} \ , \tag{H.22}$$

where $\mathbf{R}_{\text{in}} = \mathbf{R}_{\mathbf{x}_{\text{i}}} + \mathbf{R}_{\mathbf{v}}$ and $\mathbf{R}_{\mathbf{x}_{\text{i}}}$ is the covariance matrix of $\mathbf{x}_{\text{i}}(n)$. The harmonic distortion measure is useful when the desired signal is periodic. This measure is defined as the sum of the absolute differences between the harmonics before and after filtering. The harmonic distortion of the ODW filter can be show to be

$$\xi_{\text{hd}}(\mathbf{h}_{\text{W}}) = 2\sum_{l=1}^{L} P_l \left|1 - |\mathbf{1}^T\mathbf{P}\mathbf{Z}^H\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{z}(l\omega_0)|^2\right| . \tag{H.23}$$

3.2 Harmonic Decomposition LCMV

This filter is designed for noise reduction of periodic signals. The HDLCMV filter is designed such that the variance of the residual noise is minimized under the constraint that the harmonics of the desired signal are not distorted. This design can also be written as the following optimization problem

$$\min_{\mathbf{h}} \mathbf{h}^T\mathbf{R}_{\mathbf{v}}\mathbf{h} \quad \text{s.t.} \quad \mathbf{Z}^H\mathbf{h} = \mathbf{1} . \tag{H.24}$$

The solution to the quadratic optimization problem above is well-known and given by

$$\mathbf{h}_{\text{HDLCMV}} = \mathbf{R}_{\mathbf{v}}^{-1}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{R}_{\mathbf{v}}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1} \tag{H.25}$$

$$= \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1} . \tag{H.26}$$

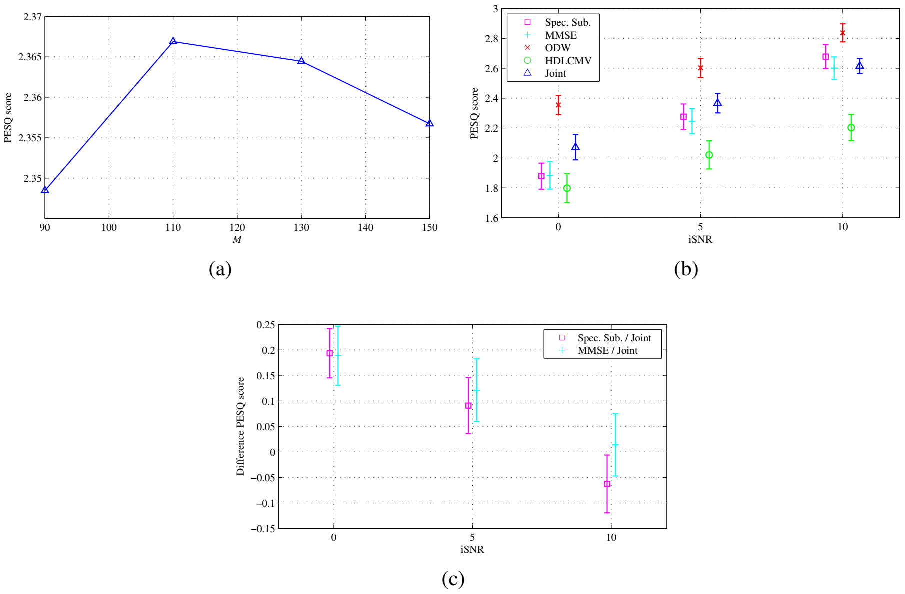
The step from (H.25) to (H.26) can be shown by using the matrix inversion lemma. From this expression, we can see that if we know the pitch ω_0 and the number of harmonics L then we only need the statistics of the observed signal $\mathbf{R}_{\mathbf{y}}$ to design the HDLCMV filter. Note that these parameters can be estimated using the very same HDLCMV filtering method [7]. In the ODW filter, we need to know either the statistics of the desired signal $\boldsymbol{\rho}_{\mathbf{x}x}$ or of the noise $\boldsymbol{\rho}_{\mathbf{v}v}$. When the filter order M becomes large and the desired signal is indeed periodic, it can be shown that the ODW and HDLCMV filters become identical. In the harmonic decomposition, there is is no interference term. The output SNR of a harmonic decomposition based filter is therefore simply defined as the ratio between the variances of the filtered desired signal and the residual noise. Therefore, the output SNR of the HDLCMV filter is given by

$$\text{oSNR}^{\text{HD}}(\mathbf{h}_{\text{HDLCMV}}) = \frac{\sigma_x^2}{\mathbf{1}^T(\mathbf{Z}^H\mathbf{R}_{\mathbf{v}}^{-1}\mathbf{Z})^{-1}\mathbf{1}} , \tag{H.27}$$

where $\mathbf{B} = \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{Z}$ and $\mathbf{C} = \mathbf{Z}^H\mathbf{B}$. The harmonic distortion in (H.23) of the HDLCMV filter is always 0 due to its constraints.

4 Joint ODW and HDLCMV Filtering

In this section, we propose to use the ODW and HDLCMV filters jointly for noise reduction in voiced speech segments. The joint use of the filters is relevant since they have



(a)

(b)

(c)

Fig. H.1: Average PESQ scores (a) for the joint filtering scheme as a function of M for an iSNR of 5 dB, and (b) for several enhancement methods as a function of the iSNR for $M = 110$ with 95% confidence intervals. In (c), the average differences in PESQ scores between the joint filtering scheme and the spectral subtraction and MMSE-based methods, respectively, are plotted with 95% confidence intervals.

complementary advantages and disadvantages. The ODW filter is in practice reliant on the noise statistics. The noise signal is, however, not available directly in practice, so the noise statistics are relatively difficult to estimate. That is, nonstationary noise has a detrimental impact on the performance of the ODW filter. The HDLCMV filter, on the other hand, is driven by the pitch, and the observed signal statistics. It should therefore be more robust against nonstationary noise since the noise statistics are not needed directly in the filter design. The HDLCMV filter, however, assumes that the desired signal is quasi-periodic which is not exactly true for all parts of speech. As a result of that, distortion will be introduced by the HDLCMV filter due to model mismatch. Therefore, it should be beneficial to use the filters jointly. In the joint filtering scheme, the HDLCMV filter is used to obtain a rough estimate of the desired signal. This estimate is then subtracted from the observed signal to obtain an estimate of the noise. The estimated noise is used to find the noise statistics which, eventually, are applied in the design of the ODW filter. Finally, the ODW filter is utilized for estimating the desired signal.

The proposed joint filtering scheme was evaluated by measuring “Perceptual Evaluation of Speech Quality” (PESQ) scores [9]. The PESQ score is an objective measure that reflects the subjective quality of a speech signal, and the score can be measured relative to an original speech signal or not. That is, by evaluating the proposed scheme using PESQ scores, we evaluate the perceptual performance of the scheme. We compared the PESQ scores of the signals enhanced using the joint filtering scheme with those enhanced using the ODW filter only, the HDLCMV filter only, a spectral subtraction based method [10], and a method using MMSE spectral amplitudes [11]. In the design of the ODW filter, the noise signal is assumed available, so the performance of this method can be thought of as an upper bound on the performance of the proposed method. Followingly, we describe how the enhancement methods were set up for the evaluation. The statistics needed for the filter designs were replaced by the respective sample covariance matrices calculated from the past 400 samples. The filters in the joint filtering scheme were reguralized using [12]

$$\hat{\mathbf{R}}_{\text{reg}} = (1-\gamma)\hat{\mathbf{R}} + \gamma \text{Tr}\left\{\hat{\mathbf{R}}\right\} M^{-1}\mathbf{I}\,, \tag{H.28}$$

where Tr$\{\cdot\}$ is the trace operator and γ is the regularization factor. Regularization was necessary due to estimation error on the signal statistics and model mismatch. We chose $\gamma = 0.7$ which gave consistently good results in terms of PESQ scores. At each time instance, the model order was set to $L = \min\{[15, \lfloor \pi/\omega_0 \rfloor - 1]\}$. The speech signals used for the evaluation contains both voiced and unvoiced parts, however, the HDLCMV filter in the proposed filtering scheme is suited for voiced speech enhancement only. Therefore, in the simulations, we updated the HDLCMV filter as follows; for voiced speech segments the HDLCMV filter was designed using (H.26) while, for unvoiced speech segments, it was updated as

$$\mathbf{h}(n) = (1-\lambda)\mathbf{0} + \lambda\mathbf{h}(n-1)\ , \tag{H.29}$$

when $\|\mathbf{h}(n-1)\|_2 > 0.1$ with $\lambda = 0.95$ and $\mathbf{0}$ is the zero vector. The spectral subtraction and MMSE based methods are available in the VOICEBOX toolbox[1] for MATLAB in which they are implemented using noise power spectral density estimates calculated using optimal smoothing and minimum statistics [3]. We used the defaults settings in the toolbox for these enhancement methods.

We conducted a number of experiments where we used the joint filtering scheme for nonstationary noise reduction. For these experiments, we used two female and two male speech excerpts of length 4-6 seconds taken from the Keele database [13]. In this paper, we treat the pitch and the harmonic model order as known parameters to evaluate the maximum achievable performance of the proposed method. Therefore, we used the pitch information from the Keele database to design the HDLCMV filter. Moreover, we do not consider voiced/unvoiced speech detection in this paper. The pitch

[1] `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`

track from the Keele database contains zeros when the speech signal is unvoiced or no speech is present, so this information was used to circumvent the detection problem. We then generated observed signals by adding different noise types to the different speech excerpts; the added noise types were white Gaussian noise, car noise, babble noise, exhibition hall noise, and street noise. All noise sources except the white noise were taken from the AURORA database [14]. First, we enhanced the noisy signals at an iSNR of 5 dB at different filter lengths, and the PESQ scores were measured and average across the different excerpts. The resulting PESQ scores are shown in Fig. H.1a. It can be seen that the perceptual performance is highest around $M = 110$. We then enhanced the noisy signals for different iSNRs when the filter length was $M = 110$. The average PESQ scores with 95 % confidence intervals are depicted in H.1b. These results indicate that the proposed scheme outperforms the spectral subtraction and MMSE-based methods for iSNRs of 0 and 5 dB on average in terms of perceptual confidence. To investigate this further, we measured the average difference in PESQ scores between the proposed scheme and the two other methods; the average differences are shown in Fig. H.1c. From these results, we can conclude that the proposed scheme outperforms the spectral subtraction and MMSE-based methods in terms of average PESQ scores with 95 % confidence for low SNRs.

5 Conclusions

In this paper, we proposed a joint filtering scheme for nonstationary noise reduction of quasi-periodic signals. The joint scheme consists of the ODW and HDLCMV filters. The ODW filter is driven only by the noise statistics and is therefore appropriate for enhancement of any desired signal. However, in practice the noise is not available directly, so the noise statistics are difficult to estimate. As a consequence of that, the performance of the ODW filter is deteriorated by nonstationary noise. The HDLCMV filter assumes that the desired signal is periodic and thereby has a harmonic structure. This is a good assumption for voiced parts of speech signals. Using this assumption, the HDLCMV filter is designed using the pitch and the model order of the desired harmonic signal, and the statistics of the observed signal, i.e., this filter is not dependent on the noise statistics. The HDLCMV filter is therefore more robust against nonstationary noise, but it will introduce some distortion in practice due to the periodicity assumption. The advantages and disadvantages of the ODW and HDLCMV filter are complementary, and we therefore proposed to use the filters jointly. In the joint scheme, the HDLCMV filter is used to estimate the noise statistics which are then used to design the ODW filter. The noise reduction is then performed by the ODW filter. We showed that the proposed joint filtering method outperforms existing speech enhancement methods in terms of average PESQ scores with 95 % confidence for relatively low iSNRs ($\leq$ 5 dB).

References

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

[2] D. Ealey, H. Kelleher, and D. Pearce, "Harmonic tunnelling: tracking non-stationary noises during speech," in *Proc. Eurospeech*, 2001.

[3] 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.

[4] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical v/uv classification algorithm," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 3, pp. 333–338, may 1999.

[5] 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, mar 2006.

[6] J. Benesty and J. Chen, *Optimal Time-Domain Noise Reduction Filters – A Theoretical Study*, 1st ed. Springer, 2011, no. VII.

[7] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[8] ——, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

[9] ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," no. P.862, pp. 1–30, Feb. 2001.

[10] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, 1979, pp. 208–211.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[12] F. van der Heijden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax, *Classification, Parameter Estimation and State Estimation - An Engineering Approach using MATLAB®*. John Wiley & Sons Ltd, 2004.

[13] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. Eurospeech*, Sep. 1995, pp. 837–840.

[14] 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.

Paper I

An Optimal Spatio-Temporal Filter for Extraction and Enhancement of Multi-Channel Periodic Signals

Jesper Rindom Jensen, Mads Græsbøll Christensen and
Søren Holdt Jensen

The paper has been published in the
Rec. Asilomar Conf. Signals, Systems, and Computers, 2010.

The layout has been revised.

Abstract

Filtering methods have been widely used for extraction of signals in both time and space. Recently, two multi-channel filters have been proposed which can be applied for extraction of multi-channel periodic signals. In these filters, the harmonic structure of periodic signals is exploited. The filters were based on the periodogram and the LCMV beamformer, respectively. The periodogram-based filter is unsuitable for multi-source scenarios whereas the LCMV-based filter has an erratic filter response behaviour at high SNRs. We propose an optimal filtering method which is useful in multi-source scenarios and which has a nicely behaving filter response for a larger range of SNRs compared to the LCMV-based filter. Our simulations show that our method solves the high SNR issue of the LCMV-based filter and that the proposed filter is applicable to real-life signals.

1 Introduction

In many applications, it is beneficial to separate or extract one or more desired signals from a mixture. A few examples of such applications are teleconferencing, surveillance systems and hearing-aids. Often, the signal of interest (SOI) in these applications is speech and/or musical instrument signals. These types of signals are known to be quasi-periodic. Thus, for short data segmentents, we can model such signals as

$$s(n_t) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_t n_t} \text{ , for } n_t = 0, \dots, N_t - 1 \text{ ,} \tag{I.1}$$

where N_t is the number of temporal samples, L is the model order and $\alpha_l = A_l e^{j\phi_l}$ with $A_l > 0$ and ϕ_l being the real amplitude and the phase of the lth harmonic, respectively. Previously, it has been investigated how a single-channel signal as in (I.1) can be extracted from a noisy mixture using, for example, algebraic separation [1] and comb filtering [2]. More recently, optimal filter designs for fundamental frequency estimation were proposed [3–5], and these filters can be seen as either generalizations of the MVDR beamformer [6] or special cases of the LCMV beamformer [7]. While the filtering methods have a good parameter estimation performance in settings with multiple interfering sources, they perform poorly for extraction purposes. This is particularly true for the high signal-to-noise ratio (SNR) settings. In these settings, the filter design problem becomes ill-conditioned, hence, the poor extraction performance. In [8, 9], a set of optimal filters for extractions and enchancement purposes were derived. As opposed to the MVDR/LCMV-like optimal filters, these filters have a good performance regarding extraction of synthetic as well as real-life periodic sources.

Sometimes, however, the signal is recorded by an array of microphones in the aforementioned applications. The mentioned single-channel methods are therefore inappropriate in such cases. In a multi microphone scenario, we can write the signal observed at the n_sth microphone as

$$x_{n_s}(n_t) = s_{n_s}(n_t) + w_{n_s}(n_t) \text{ , for } n_t = 0, \dots, N_t - 1 \text{ ,} \tag{I.2}$$

where $s_{n_s}(n_t) = s(n_t - \tau_{n_s})$ is the SOI, $w_{n_s}(n_t)$ is the noise on the n_sth sensor, N_s is the number of microphones and τ_{n_s} is the time delay of the sound wave from sensor n_s to a reference point. We assume a uniform linear array (ULA) structure so the time delay is given by $\tau_{n_s} = n_s \frac{d \sin \theta}{c}$ for $\theta \in [-90°; 90°]$ where d is the microphone spacing, θ is the DOA and c is the wave propagation velocity. Combining (I.1) and (I.2) leads to a multi-channel harmonic model

$$x_{n_s}(n_t) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_t n_t} e^{-jl\omega_s n_s} + w_{n_s}(n_t) \text{ ,} \tag{I.3}$$

where $\omega_s = \omega_t f_s d c^{-1} \sin \theta$. Due to the harmonic structure, the voiced speech and audio extraction problem can be considered as extraction of L narrowband sources while, traditionally, voiced speech and audio have been considered broadband in multi-channel extraction methods. The narrowband simplification enables us to derive much simpler extraction algorithms which is evident from the following sections.

Recently, two methods for joint DOA and fundamental frequency esimation were proposed [10]. One of them was signal independent since it was based on the periodogram while the other was based on the LCMV beamformer and therefore signal dependent. Although these filtering methods could also be used for extraction of multi-channel periodic sources, they suffer from the same issues as the corresponding single-channel methods. In this paper, we therefore derive a new joint spatio-temporal optimal filter for extraction of (quasi-)periodic sources from multi microphone recordings. We will term the filter design method as the filtering-based multi-channel periodic signal extraction (FIMPSIX) method. The filter is designed optimally from the observed signal and is therefore signal adaptive. Like the filters in [8, 9], the proposed filter is inspired by the well known amplitude and phase estimation (APES) method [11]. We expect that the proposed filter will outperform the filtering methods in [10] regarding extraction, since this is the case for the analogous single-channel filtering methods. The main application of the proposed method is extraction and enhancement, however, it can also be used for joint DOA and fundamental frequency estimation, model order selection and amplitude estimation of the individual harmonics.

The rest of the paper is organized as follows. In Section 2, we state the filter design problem and introduce the notation. We solve the filter design problem in Section 3. In Section 4, we describe the experimental evaluation of the proposed filter design, and, finally, we conclude on our work in Section 5.

2 Joint Spatio-Temporal Filter Design Problem

We consider the problem of designing a joint optimal spatio-temporal filter for extraction of periodic sources. Generally speaking, the output $y(n_t)$ of an FIR filter with the coefficients $h(n_s, m_t)$ from the input $x_{n_s}(n_t - m_t)$ can be written as

$$y(n_t) = \sum_{n_s=0}^{N_s-1} \sum_{m_t=0}^{M_t-1} h(n_s, m_t) x_{n_s}(n_t - m_t) \ , \tag{I.4}$$

for $n_{\{s,t\}} = 0, \ldots, N_{\{s,t\}-1}$. Our goal is to design the filter such that its output resembles a desired signal $\hat{y}(n_t)$ as much as possible in the mean squared error (MSE) sense. The MSE P is given by

$$P = \frac{1}{N_t - M_t + 1} \sum_{n_t=M_t-1}^{N_t-1} |y(n_t) - \hat{y}(n_t)|^2 \ . \tag{I.5}$$

In this filter design, the desired signal is defined as the noise-free signal given by the signal model in (I.1). If we insert (I.1) and (I.4) into (I.5) we get

$$P = \frac{1}{N_t - M_t + 1} \sum_{n_t=M_t-1}^{N_t-1} \tag{I.6}$$
$$\left| \sum_{n_s=0}^{N_s-1} \sum_{m_t=0}^{M_t-1} h(n_s, m_t) x_{n_s}(n_t - m_t) - \sum_{l=1}^{L} \alpha_l e^{jl\omega_t n_t} \right|^2 \ .$$

Whereas we initially assume that the fundamental frequency ω_0 and the model order L are known, it is shown later how the proposed filter can also estimate these parameters. The expression in (I.6) can be simplified by introducing matrix/vector notation. Consider, for example, the filter and signal matrices, $\mathbf{H}$ and $\mathbf{X}(n_t)$, defined as

$$\mathbf{H} = \begin{bmatrix} h^*(0,0) & \cdots & h^*(0, M_t - 1) \\ \vdots & \ddots & \vdots \\ h^*(N_s - 1, 0) & \cdots & h^*(N_s - 1, M_t - 1) \end{bmatrix} \tag{I.7}$$

$$\mathbf{X}(n_t) = \begin{bmatrix} x_0(n_t) & \cdots & x_0(n_t - M_t + 1) \\ \vdots & \ddots & \vdots \\ x_{N_s-1}(n_t) & \cdots & x_{N_s-1}(n_t - M_t + 1) \end{bmatrix} , \tag{I.8}$$

where $(\cdot)^*$ denotes the complex conjugate. We define two new vectors $\mathbf{h} = \text{vec}\{\mathbf{H}\}$ and $\mathbf{x}(n_t) = \text{vec}\{\mathbf{X}(n_t)\}$ with vec$\{\cdot\}$ denoting the column-wise matrix stacking operator. This enable us to obtain a much more convenient MSE expression as

$$P = \frac{1}{N_t - M_t + 1} \sum_{n_t=M_t-1}^{N_t-1} |\mathbf{h}^H \mathbf{x}(n_t) - \boldsymbol{\alpha}^H \mathbf{e}(n_t)|^2 \ , \tag{I.9}$$

where

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \cdots & \alpha_L \end{bmatrix}^H \tag{I.10}$$

$$\mathbf{e}(n_t) = \begin{bmatrix} e^{j\omega_t n_t} & \cdots & e^{jL\omega_t n_t} \end{bmatrix}^T . \tag{I.11}$$

It turns out that we can expand the MSE expression in (I.9) as

$$P = \mathbf{h}^H \hat{\mathbf{R}} \mathbf{h} - \boldsymbol{\alpha}^H \mathbf{G} \mathbf{h} - \mathbf{h}^H \mathbf{G}^H \boldsymbol{\alpha} + \boldsymbol{\alpha}^H \mathbf{E} \boldsymbol{\alpha} , \tag{I.12}$$

with

$$\hat{\mathbf{R}} = \frac{1}{N_t - M_t + 1} \sum_{n_t = M_t - 1}^{N_t - 1} \mathbf{x}(n_t) \mathbf{x}^H(n_t) \tag{I.13}$$

$$\mathbf{G} = \frac{1}{N_t - M_t + 1} \sum_{n_t = M_t - 1}^{N_t - 1} \mathbf{e}(n_t) \mathbf{x}^H(n_t) \tag{I.14}$$

$$\mathbf{E} = \frac{1}{N_t - M_t + 1} \sum_{n_t = M_t - 1}^{N_t - 1} \mathbf{e}(n_t) \mathbf{e}^H(n_t) . \tag{I.15}$$

We recognize that $\hat{\mathbf{R}}$ is the spatio-temporal sample covariance matrix [10].

3 Derivation of the Optimal Filter

Following, we derive the optimal spatio-temporal filter by solving the filter design problem introduced in Section 2. First, if we differentiate and solve with respect to $\boldsymbol{\alpha}$ in (I.12) we get that

$$\hat{\boldsymbol{\alpha}} = \mathbf{E}^{-H} \mathbf{G} \mathbf{h} . \tag{I.16}$$

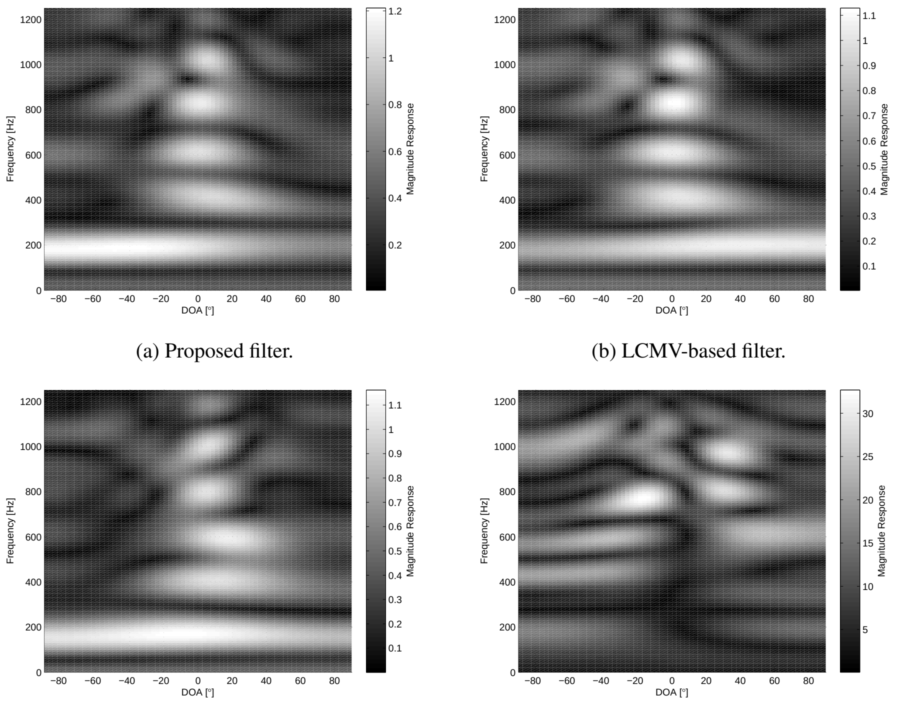
Inserting the amplitude estimate $\hat{\boldsymbol{\alpha}}$ into (I.12) yields

$$P = \mathbf{h}^H (\hat{\mathbf{R}} - \mathbf{G}^H \mathbf{E}^{-1} \mathbf{G}) \mathbf{h} \tag{I.17}$$

$$= \mathbf{h}^H \hat{\mathbf{Q}} \mathbf{h} , \tag{I.18}$$

where $\hat{\mathbf{Q}} = \hat{\mathbf{R}} - \mathbf{G}^H \mathbf{E}^{-1} \mathbf{G}$ can be interpreted as an estimate of the noise covariance matrix [12]. Note that asymptotically, the matrix $\mathbf{E}$ equals $\mathbf{I}$ which can be exploited to obtain a computationally simpler algorithm [8].

The optimal filter is derived from (I.18). However, solving directly for the unknown filter leads to the zero vector solution. We circumvent this by introducing some additional constraints. The constraints are formulated such that the filter has a unit gain



(a) Proposed filter.

(b) LCMV-based filter.

(c) Proposed filter.

(d) LCMV-based filter

Fig. I.1: Frequency responses of the filters at SNRs of (a),(b) -20 dB and (c),(d) 20 dB, respectively.

at all of the harmonic frequencies and DOA pairs of the SOI. This leaves us with the following constrained optimization problem

$$\min_{\mathbf{h}} \mathbf{h}^H \hat{\mathbf{Q}} \mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^H \mathbf{z}_{l\omega_t, l\omega_s} = 1 \ , \qquad \text{for} \quad l = 1, \dots, L \ , \tag{I.19}$$

where

$$\mathbf{z}_{l\omega_t, l\omega_s} = \mathbf{z}_{l\omega_t} \otimes \mathbf{z}_{l\omega_s} \tag{I.20}$$

$$\mathbf{z}_{l\omega_t} = \begin{bmatrix} 1 & e^{-jl\omega_t} & \cdots & e^{-jl\omega_t(M_t-1)} \end{bmatrix}^T \tag{I.21}$$

$$\mathbf{z}_{l\omega_s} = \begin{bmatrix} 1 & e^{-jl\omega_s} & \cdots & e^{-jl\omega_s(N_s-1)} \end{bmatrix}^T \ , \tag{I.22}$$

with $\otimes$ denoting the Kronecker product operator. Note that all constraints can be written as a singel matrix-vector product as

$$\mathbf{h}^H \mathbf{Z}_{\omega_t,\omega_s} = \mathbf{1} \ , \tag{I.23}$$

where

$$\mathbf{Z}_{\omega_t,\omega_s} = \begin{bmatrix} \mathbf{z}_{\omega_t,\omega_s} & \cdots & \mathbf{z}_{L\omega_t,L\omega_s} \end{bmatrix} . \tag{I.24}$$

We recognize that the problem in (I.19) is a quadratic optimization problem which is solvable using the Lagrange multiplier method. If we introduce the Lagrange multiplier vector $\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 & \cdots & \lambda_L \end{bmatrix}$, the Lagrangian dual function is given by

$$\mathcal{L}(\mathbf{h}, \boldsymbol{\lambda}) = \mathbf{h}^H \hat{\mathbf{Q}} \mathbf{h} - \left(\mathbf{h}^h \mathbf{Z}_{\omega_t,\omega_s} - \mathbf{1}^T \right) \boldsymbol{\lambda} \ . \tag{I.25}$$

By differentiating the Lagrange dual function with respect to the unknown Lagrange multiplier $\boldsymbol{\lambda}$ and the unknown filter $\mathbf{h}$, by equating with $\mathbf{0}$, and by inserting the so-obtained expressions into each other, we get that the optimal filter $\hat{\mathbf{h}}$ is given by

$$\hat{\mathbf{h}} = \hat{\mathbf{Q}}^{-1} \mathbf{Z}_{\omega_t,\omega_s} \left(\mathbf{Z}^H_{\omega_t,\omega_s} \hat{\mathbf{Q}}^{-1} \mathbf{Z}_{\omega_t,\omega_s} \right)^{-1} \mathbf{1} \ . \tag{I.26}$$

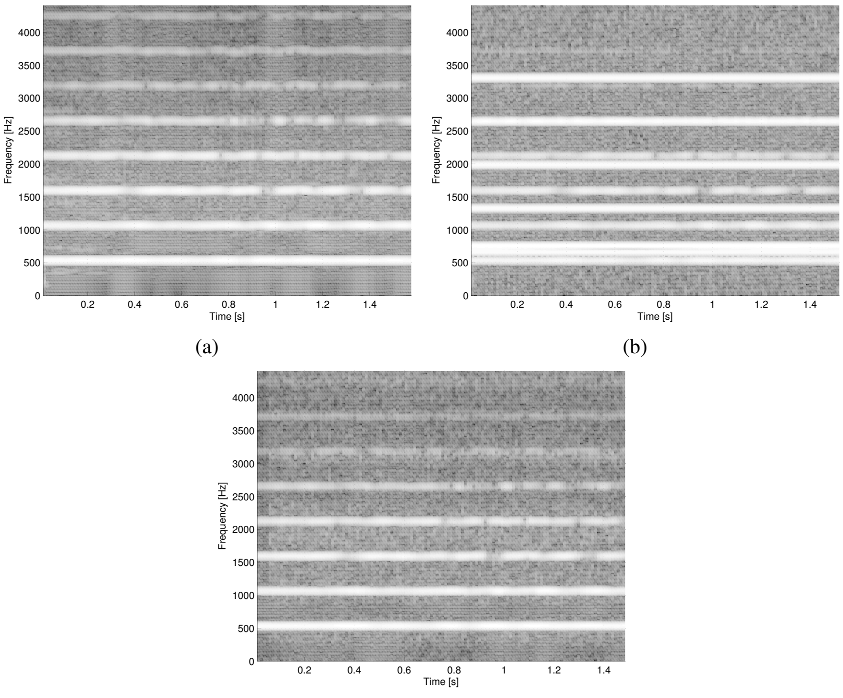
Note that the optimality criterion for the filter design is twofold: 1) the filter gain should be one at all harmonic frequencies and DOA pairs while the filter minimizes all other frequency/DOA components, and 2) the filter output should resemble a sum of sinusoids as much as possible under the given constraints. If we insert the optimal filter response in (I.26) into (I.16), we can obtain estimates of the amplitudes of the harmonics

$$\hat{\boldsymbol{\alpha}} = \mathbf{E}^{-1} \mathbf{G} \hat{\mathbf{Q}}^{-1} \mathbf{Z}_{\omega_t,\omega_s} (\mathbf{Z}^H_{\omega_t,\omega_s} \hat{\mathbf{Q}}^{-1} \mathbf{Z}_{\omega_t,\omega_s})^{-1} \mathbf{1} \ . \tag{I.27}$$

Introductory, we asummed that the fundamental frequency ω_0 was known. If this is not the case we could either estimate it using another method or using the just proposed optimal filter. To estimate it using the proposed filter, the optimal filter is applied on the input signal and the output power is then estimated. This procedure is repeated for a two-dimensional grid of candidate fundamental frequencies and DOAs. The fundamental frequency estimated is obtained by taking the argument of the maximizing fundamental frequency and DOA pair as

$$\{\hat{\omega}_t, \hat{\theta}\} = \arg \max_{(\omega_t,\theta) \in \Omega_t \times \Theta} \hat{\mathbf{h}}^H \hat{\mathbf{R}} \hat{\mathbf{h}} \ , \tag{I.28}$$

with Ω_t and Θ being sets of candidate fundamental frequencies and DOAs, respectively. Likewise, the optimal filtering method can be used for model order L estimation according to [5].



(a)

(b)

(c)

Fig. I.2: Spectrograms of (a) a trumpet signal, (b) a trumpet signal in noise consisting of interfering periodic sources and complex Gaussian noise at a 30 dB SNR and (c) a signal extracted using the optimal filter.

4 Experimental Results

Following, we describe the experimental evaluation of the proposed filter design. As mentioned previously, filtering methods based on the minimum variance principle suffers from a bad performance regarding signal extraction. This is well known, and the main reason is their unfortunate behavior at high SNRs. Several books and papers (e.g., [13, 14]) have dealt with this issue and a common fix is to, for instance, use diagonal loading techniques. The erratic high SNR behavior is also apparent from our first experiment. In this experiment, we investigate the frequency response of the proposed filter and the LCMV-based filter proposed in [10]. The filters were designed to extract a multi-channel periodic signal with $N_t = 250$, $f_s = 2,500$ Hz, $f_t = 200$ Hz, $L = 5$, $\theta = 6°$, and unit amplitudes of the harmonics. Moreover, the signal was corrupted by complex Gaussian noise, the array was specified by $N_s = 6$, $c = 343.2$ m/s and

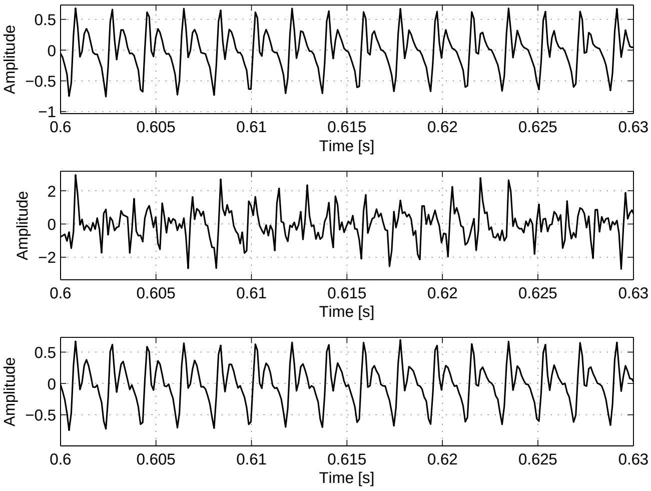


Fig. I.3: Segments of (top) the trumpet signal, (middle) the observed signal with complex Gaussian noise at a SNR of 30 dB and interfering periodic sources, and (bottom) the signal extracted using the optimal filter.

$d = c/f_s$, and the filter was of orders $M_t = 20$ and $M_s = 6$. We designed the filters for SNRs of -20 dB (depicted in Fig. I.1a and I.1b) and 20 dB (depicted in Fig. I.1c and I.1d), respectively. From the plots, we observe that both filter types behave nicely at low SNRs. At high SNRs, the proposed filter still seems to perform nicely in terms of emphasizing the harmonics while the LCMV-based filter has huge side lobes.

In the second experiment, we applied the proposed filter for extraction of a trumpet signal. The utilized trumpet signal was originally single-channel and sampled at $f_s = 8,820$ Hz. Therefore, we resynthesized it spatially as if it was impinging on a 4-element ULA with a DOA of $\theta = 17°$. We corrupted the trumpet signal with additional synthetic periodic sources and complex Gaussian noise at an SNR of 30 dB. The FIMPSIX filter was designed for 60 ms segments with a filter length of 40 and it was updated every 30 ms. For each input segment, we estimated the fundamental frequency of the trumpet signal using an MVDR-based method. The filter was designed for the estimated fundamental frequency and for a fixed model order of $L = 8$. The spectrograms of the original trumpet signal, the noisy signal, observed on the first sensor, and the extracted signal are depicted in Fig. I.2(a)-(c), respectively. Furthermore, short segments of the different signals are shown in Fig. I.3. It is clear from these figures that the proposed filter design are useful for extraction of periodic signals (or nearly periodic signals such as the trumpet signal).

5 Conclusion

In this paper, we proposed a novel optimal joint spatio-temporal filtering method for extraction of periodic signal recorded in time and space using a uniform linear microphone array. The proposed filter is based on a harmonic model which makes it suitable for all signals being (quasi-)periodic of nature such as audio and speech. Specifically, the proposed filter is inspired by the amplitude and phase estimation (APES) method. By using the APES principle rather than the minimum variance principle in the filter design, we obtain a filter with a less erratic filter response at high SNRs compared to minimum variance based filters. This is also evident from the experimental results. Due to the better filter response behaviour, the proposed filter is better suited for signal extraction. Our simulation results showed that the proposed filter is also applicable for extraction of real-life signals such as a trumpet signal. From the results, it is clear that the filter is useful for suppressing both random noise and interferering sources.

References

[1] Z. Mouyan, C. Zhenming, and R. Unbehauen, "Separation of periodic signals by using an algebraic method," in *Proc. IEEE Int. Symp. Circuits and Systems*, Jun. 1991, pp. 2427–2430.

[2] 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.

[3] 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.

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

[5] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[6] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

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

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

[9] ——, “Optimal filter designs for separating and enhancing periodic signals,” *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

[10] 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.

[11] 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.

[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, no. 2, pp. 338–352, Feb. 2000.

[13] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Education, Inc., 2005.

[14] L. Du, T. Yardibi, J. Li, and P. Stoica, “Review of user parameter-free robust adaptive beamforming algorithms,” *Digital Signal Processing*, vol. 19, no. 4, pp. 567–582, Jul. 2009.