Pre-processing of Speech Signals for Robust Parameter Estimation

Esquivel Jaramillo, Alfredo

DOI (link to publication from Publisher):
10.54337/aau456472165

Publication date:
2021

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

PRE-PROCESSING OF SPEECH SIGNALS FOR ROBUST PARAMETER ESTIMATION

BY
ALFREDO ESQUIVEL JARAMILLO

DISSERTATION SUBMITTED 2021

AALBORG UNIVERSITY
DENMARK

Pre-processing of Speech Signals for Robust Parameter Estimation

Ph.D. Dissertation

Alfredo Esquivel Jaramillo

Aalborg University
Audio Analysis Lab
CREATE
Rendsburggade 14
DK-9000 Aalborg

Dissertation submitted: August, 2021

PhD supervisor: Professor Mads Græsbøll Christensen
Aalborg University

Assistant PhD supervisor: Associate Professor Jesper Kjær Nielsen
Siemens Gamesa

PhD committee: Associate Professor Markus Löchtefeld (chair)
Aalborg University

Professor Maria Sandsten
Lund University

Associate Professor Christian Fischer Pedersen
Aarhus University

PhD Series: Technical Faculty of IT and Design, Aalborg University

Department: Department of Architecture,
Design and Media Technology

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-984-8

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2021

Curriculum Vitae

Alfredo Esquivel Jaramillo

Alfredo was born in Mexico City, Mexico. He received a M.Sc degree in Electrical Engineering, with focus on digital signal processing, from the National Autonomous University of Mexico (UNAM), in 2015. From 2014 to 2016 he was a part-time lecturer of electrical engineering courses in the Autonomous University of Mexico City (UACM). In 2016, he joined the Audio Analysis Lab at Aalborg University, Denmark, as a Ph.D. student. In 2019, he visited the Centre of Mathematical Sciences at Lund University, Sweden. His research interests are speech and audio processing and speech analysis. Alfredo has served as a reviewer for IEEE Access and IEEE Internet of Things Journal.

Abstract

The topic of this thesis is methods of pre-processing speech signals for robust estimation of model parameters in models of these signals. Here, there is a special focus on the situation where the desired signal is contaminated by colored noise. In order to estimate the speech signal, or its voiced and unvoiced components, from a noisy observation, it is important to have robust estimators that can handle colored and non-stationary noise.

Two important aspects are investigated. The first one is a robust estimation of the speech signal parameters, such as the fundamental frequency, which is required in many contexts. For this purpose, fast estimation methods based on a simple white Gaussian noise (WGN) assumption are often used. To keep using those methods, the noisy signal can be pre-processed using a filter. If the colored noise is modelled as an autoregressive (AR) process, whose parameters are estimated from the noisy signal, it is possible to render the noise component closer to white with a simple pre-processing filter (pre-whitener). This makes it possible to estimate the fundamental frequency using the aforementioned assumption of white Gaussian noise. In non-stationary noise scenarios, it is possible to obtain better estimates of the noise spectral envelope as well as a higher degree of spectral flatness by using an adaptive pre-whitening filter based on supervised noise statistics estimates, than one based on unsupervised noise statistics. A pre-whitening filter also improves the accuracy of a source localization method. The problem of joint estimation of the parameters of the voiced speech and the stochastic signal parts (i.e., unvoiced and additive noise) is solved first by the cascade of a pre-whitening filter and the nonlinear least squares (NLS) fundamental frequency estimator, followed by an iterative estimation of the pre-whitening filter, based on the modelled residual, and a re-estimation of the fundamental frequency. This will further reduce the number of gross errors of fundamental frequnecy estimates and the voicing detection errors.

The second aspect is as follows: after a more accurate estimation of the parameters is obtained, the extraction of individual speech components (i.e., voiced and unvoiced speech) from a noisy speech signal, is investigated

through linear filtering based on the statistics of the individual components. A Wiener filtering approach allows for a better recovery of both components when compared to the state-of-the-art decomposition methods, which assume that the additive noise is small and insignificant. Instead of using a fixed segment length for the extraction, we also propose to use time-varying segment lengths that are adapted to the signal. The optimal segmentation is obtained once the parameter estimates of a hybrid speech model have been found for all possible candidate models and segment lengths.

Resumé

Emnet for denne afhandling er metoder til forbehandling af talesignaler til robust estimering af model parametre i modeller af disse signaler. Her er der især fokus på situationen hvor det ønskede signal forurenes af farvet støj. For at estimere et talesignal eller dets stemte og ustemte lyd fra støjfulde signaler, er det vigtigt at have robuste estimatorer, der kan håndtere farvet og ikke-stationær støj.

To vigtige aspekter undersøges. Det første aspeckt er en robust estimering af talesignalers parametre, såsom grundfrekvensen, er nødvending i mange sammenhænge. Til dette formål bruges ofte hurtige estimeringsmetoder baseret på en simpel antagelse om hvid, Gaussisk støj, hvor støjsignalerne forbehandles med et filter. Hvis den farvede støj modelleres med en autoregressiv (AR) proces, hvis parametre estimeres fra det støjfulde signal, er det muligt gøre støjen mere hvid med simpel forbehandlingsfilter. Herved er det muligt at estimere grundfrekvensen vha. metoder baseret på den førnævnte antagelse om hvid, Gaussisk støj. Tilsvarende er det i ikke-stationære støjscenarier muligt at få bedre estimater af støjspektre samt en højere grad af spektral fladhed ved brugen af et adaptivt forbehandlingsfilter baseret på trænede estimater af støjspektre end ved ikke-trænede estimater. Forbehandlingsfilteret forbedrer også nøjagtigheden af en metode til kildelokalisering. Problemet samtidigt at estimere parametrene af stemte lyde og stokastiske signaler (dvs. ustemte lyde og additiv støj) løses ved først at bruge en kaskadekobling af et forbehandlingsfilter og den ulineære mindste kvadraters grundfrekvens-estimator, efterfulgt af iterativ estimering af henholdsvis forbehandlingsfilteret, baseret på det modellerede støjsignal, og et nyt estimat af grundfrekvensen, hvorved store fejl i grundfrekvensen og detektionen minimeres.

Det andet aspekt er følgende: efter en mere præcis estimering af parametre er opnået, undersøges estimering af individuelle talekomponenter (dvs. stemte og ustemte lyde) fra støjfulde signaler vha. lineær filtering baseret på de individuelle komponenters statistikker. En tilgang baseret på Wienerfiltre fører til en bedre ekstrahering af begge komponenter i forhold til state-of-the-art dekomponeringsmetoder, der antager at den additive støj er lille

og ubetydelig. I stedet for at bruge en fast segmentlængde til ekstraktion, foreslår vi også at bruge tidsvarierende segmentlængder, der er tilpasset signalet. Den optimal segmentering opnås efter at parameterestimaterne for en hybrid talemodeler er fundet for alle mulige kandidatmodeller og segmentlængder.

Contents

List of publications

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

[A] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "A study on how Pre-whitening Influences Fundamental Frequency Estimation", *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[B] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "Adaptive Pre-whitening Based on Parametric NMF", *Proc. 27th European Signal Processing Conference (EUSIPCO)*, La Coruña, Spain, 2019.

[C] A. E. Jaramillo, A. Jakobsson, J. K. Nielsen, and M. G. Christensen, "Robust Fundamental Frequency Estimation in Coloured Noise", *Proc. 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020.

[D] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "An Adaptive Autoregressive Pre-whitener for Speech and Acoustic Signals Based on Parametric NMF", submitted to *Applied Acoustics*.

[E] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "On Optimal Filtering for Speech Decomposition", *Proc. 26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, 2018.

[F] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "Speech Decomposition based on a Hybrid Speech Model and Optimal Segmentation" *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, accepted, Brno, Czech Republic, 2021.

Preface

This thesis is submitted to the Technical Faculty of IT and Design at Aalborg University in a partial fulfilment of the requirements for the degree of Doctor of Philosophy. The thesis is concerned with pre-processing methods of speech signals for robust estimation of the parameters of models of speech signals. The thesis is divided into two parts. The first part gives an overview on models, speech signal and noise parameters, and processing techniques, followed by a summary of the contributions of the Ph.D. project. The main body of the thesis is the second part. This part consists of a number of papers which have been published in or submitted to peer-reviewed conferences or journals.

The work was carried out in the period of October 2016 to March 2021 in the Audio Analysis Lab at CREATE department at Aalborg University. I am grateful to the Mexican National Council of Science and Technology (CONACYT) for the granted scholarship during my Ph.D. studies. I am also thankful to the Secretary of Public Education, which provided me a complementary support during the first two years. I want to express my grattitude to Mads Græsbøll Christensen who gave me the opportunity to work on this research topic, and for orienting me in the direction to focus this work. I am also very grateful to my cosupervisor Jesper Kjær Nielsen for all the beneficial and fruitful discussions, and for thoroughly commenting on my papers and responding my emails, despite his busy schedule. I would also like to thank Andreas Jakobsson for supervising my work during three months, in my research stay at Lund University, Sweden.

I want also to thank my colleagues at the Audio Analysis Lab for their support and social gatherings we had during my stay at the Lab. I would also like to thank to my friends and family in Mexico, including grandparents, my numerous cousins, aunts and uncles for their support. I want to thank to my sister Diana Paulina for visiting me several times in Denmark, and for being very close to me in the last year of my stay at Aalborg. Finally, I would like to thank my brother Daniel, and my parents, Sandra and Alfredo, for their infinite love and support.

Alfredo Esquivel Jaramillo
Aalborg University, August 24, 2021

Part I

Introduction

Introduction

This thesis is concerned with the pre-processing of noisy speech signals for the purpose of a robust estimation of the model parameters of models of these signals. Particularly, we are interested in obtaining accurate estimates when the signals are degraded by additive colored noise. Traditionally, the undesirable presence of noise is addressed by trying to reduce its levels as much as possible, through the application of speech enhancement methods. Often, the goal is a robust extraction of the speech signal features, such as the fundamental frequency of voiced speech. In this case, it might be more useful to use the embedded noise characteristics in the noisy signal, instead of achieving more noise reduction. A possibility could be to jointly estimate the noise statistics and the parameters of the speech signals, once a noise model has been elicited. This approach, however, might be computationally costly and not very flexible, since new estimators would need to be derived when the model assumptions are not exactly fulfilled. Certainly, there is a plethora of methods in the speech and audio processing literature which were formulated under the simple, but possibly unrealistic, white Gaussian noise (WGN) assumption.

Although the WGN assumption allows to facilitate the mathematical formulation and the development of fast algorithms, the obtained estimators may not be very robust under some conditions, as in the case where the noise spectral content is contained in specific frequency bands (i.e., it is colored). Those estimation methods based on a WGN assumption, however, could still be reliably used in typical noise scenarios, if the signal is pre-processed in such a way that the model assumptions are better fulfilled. Thus, to keep using methods based on a simple WGN assumption, the signal should be pre-processed in such a way that the noise is rendered closer to white. However, due to the non-stationary nature of speech and noise signals, we hypothesize that a particular pre-whitening scheme will provide more robustness to the WGN-based estimation methods. A comparison of different ways of pre-whitening the signal is therefore of vital importance.

The first four contributions in this thesis consider how the noise spectral features need to be exploited to design a time-varying pre-whitener. There-

fore, methods which are mathematically formulated under the assumption that the noise is white and Gaussian (WGN) can be reliably applied even if the noise is not necessarily white. To explain how this problem can be addressed, it is necessary to define the speech models of the signal of interest. The signal models may include the undesirable additive noise component, which is very often of colored nature, and in many cases it can be modelled as an autoregressive (AR) process. The different models of speech and noise signals are explained in Section 1. In order to pre-process the noisy signals, information of the noise statistics is required. Section 2 is concerned on how the noise power spectral density (PSD) can be estimated in unsupervised and supervised ways, and also it gives an insight in the problem of estimating of the fundamental frequency, which is an important parameter of the speech signals. The performance measures for assessing the fundamental frequency estimation performance are briefly mentioned. In some cases, the processing is applied in the time domain, and we also address on how the signal and noise statistics can be estimated in that case. To have a robust estimation of the fundamental frequency, it is necessary to do some optimal pre-processing of the noisy signals, and this pre-processing is dependent of the noise statistics. Pre-processors such as optimal filters for cleaning up the noisy signal (i.e. speech enhancement) and for rendering the noise component closer to white (i.e., whitening) are introduced in Section 3. Another important processing is related to the decomposition of speech signals into its voiced and unvoiced components. In the noisy case, this problem can be solved using linear filtering which relies on the estimated statistics of each individual component. The last two contributions address this problem in the presence of additive noise. An insight of the speech decomposition problem and the state-of-the-art methods which have addressed this problem are then described. Next, we detail some performance measures to quantify how close a reference signal is to its estimate. Finally, Section 4 provides an overview and summary of the main results of the papers which form the main contribution of this thesis. An overview of directions for future research is also included.

1 Speech and Noise Modelling

Before addressing how an undesired noise signal affects a clean speech signal of interest, we describe how the speech can be modelled. The speech signal waveform can be split into short intervals, known as segments, where the signal is considered as stationary. According to how the different speech sounds are produced and to their spectral content, a speech segment can be categorized as silent, voiced, unvoiced or as a combination of silence, voiced and unvoiced [31]. The voiced sounds present a quasi-periodic behaviour,

while the unvoiced sounds reflect a stochastic nature. When neither a voiced or unvoiced excitation is present, the analyzed segment represents silence or a pause [10]. In a spectral representation, e.g., in a spectrogram, the individual spectral harmonics of voiced speech sounds have a regular spacing between them and are seen as horizontal lines. Instead of these horizontal striations, the unvoiced speech sounds are identified as rectangular patches [31]. The unvoiced parts of the speech can have different spectral characteristics. For example, the aspiration noise is of broadband nature [113] while the frication noise appears mainly in the high frequency region [29]. The conventional speech production model assumes that a single mode of excitation is possible (i.e., either voiced or unvoiced) [104]. In reality, in several speech segments, there may be a coexistence of noise-like and quasi-periodic energy [166], and this fact has been exploited in important applications of speech coders [53], diagnosing of illnesses [137] and speech synthesis [30, 100, 159, 170].

1.1 Harmonic model

In this model, a sum of harmonically related sinusoids (a.k.a. harmonics) can be used to describe sampled voiced speech segments [22, 78], i.e.,

$$s(n) = \sum_{l=1}^{L} A_l \cos(2\pi f_0 l n + \psi_l), \tag{1}$$

where $n = 0, 1, ..., N-1$ and N is the corresponding segment length. The lowest frequency corresponding to the first sinusoid in the summation is f_0 and is known as the fundamental frequency [138]. The other sinusoidal terms have a frequency which is an integer multiple of f_0. The total number of harmonics in the signal is L and is known as the model order [22]. A_l and ψ_l correspond, respectively, to the real amplitude and the phase of the l^{th} harmonic. Estimating f_0 is an important problem in many speech and audio processing applications [28, 57, 152]. Estimators which are based on parametric models, such as the harmonic model, need to jointly estimate the model order L and f_0 [22, 23] so that the possibility of erroneous estimates, including sub-harmonic errors, is reduced. The sub-harmonic errors occur when an integer multiple or divisor of the true f_0 is mistakenly estimated.

Another representation that has been found to be convenient to lower the computational complexity is obtained by employing the Hilbert transform [54], yielding the complex model

$$s(n) = \sum_{l=1}^{L} \alpha_l e^{j2\pi f_0 l n}, \tag{2}$$

where $\alpha_l = A_l e^{j\psi_l}$ and $j = \sqrt{-1}$. Despite its simplicity, this model ignores the interaction between positive and negative harmonics, and has appeared

to result in f_0 estimators which have sub-optimal performance [21]. The interaction between positive and negative harmonics is seen by rewriting (1) as

$$s(n) = \sum_{l=1}^{L} \left[\alpha_l e^{j2\pi f_0 ln} + \alpha_l^* e^{-j2\pi f_0 ln} \right], \tag{3}$$

where $\alpha_l = \frac{A_l}{2} e^{j\psi l}$ and $*$ denotes the complex conjugate. By using the real signal model (3) instead of the complex model (2), it is possible that f_0 estimators based on the maximum likelihood (ML) principle reach the Cramér-Rao Lower Bound (CRLB) [91] even under adverse conditions [20, 21, 125], such as small segment lengths or in the case that the f_0 is low relative to the number of samples in the analyzed segment. In order to have a robust estimation of f_0 under adverse conditions, here the real signal model is considered. We will return back to the f_0 estimation problem in Section 2.

It is common to assume that a stationarity condition is fulfilled for segments of small duration (∼20-40 ms) [131]. To keep the simplicity, speech is commonly analyzed using segments of fixed length [163], where it is assumed that the harmonic model perfectly holds. However, in some cases it might be convenient to take the non-stationary nature of speech into account [24, 82], as the speech signal characteristics could be changing across a fixed segment length. Therefore, it could be feasible to use adaptive segments of different length across the signal [62, 129, 135], which can better accommodate to the local characteristics.

1.2 Hybrid speech model

In the previous model, it is assumed that voiced speech segments are completely described by the harmonic model. If a voiced speech segment is reconstructed from its harmonic model parameters, a component of random nature will be observed [151, 158]. Therefore, in another model referred as the hybrid speech model, it is assumed that voiced speech segments contain a stochastic residual signal which is added to the sum of harmonically related sinusoids [46, 100]. This stochastic component accounts, among different possible sources, for glottal airflow turbulences [174], frication noise, formant transitions or jitter [29]. This component is considered as unvoiced speech, implying that unvoiced parts of speech may be also present in voiced speech segments [132], although purely unvoiced segments also exist.

In the simplified speech production model, either a voiced or unvoiced excitation is only present, but this is just a simplifying assumption [104]. In practice, an hybrid mode of excitation is more realistic of how speech is produced, and it results in speech which sounds more natural [174]. Both

voiced and unvoiced components coexist in the hybrid speech model [132], i.e.,

$$s(n) = v(n) + u(n) = \sum_{l=1}^{L} A_l \cos(2\pi f_0 l n + \psi_l) + u(n), \tag{4}$$

where $v(n)$ and $u(n)$ corresponds to samples of voiced and unvoiced speech, respectively. Not-voiced segments are described when there is absence of harmonic components, i.e., if $L = 0$. This includes the possibility of a pure unvoiced speech segment, if $u(n) \neq 0$, or pauses in speech, otherwise.

The unvoiced speech can be adequately modelled as an autoregressive (AR) process [100], i.e.,

$$u(n) = -\sum_{i=1}^{P} b_u(i) u(n-i) + g(n), \tag{5}$$

where $\{b_u(i)\}_{i=1}^{P}$ are the P AR coefficients of the unvoiced component and $g(n)$ is a driving WGN process with variance σ_g^2. The model is referred as hybrid since the harmonic model part (voiced speech) has a discrete spectrum while the unvoiced part has a continuous spectrum, i.e., they are of different nature [87, 88, 100]. An insight in how to estimate the voiced and unvoiced components of a speech signal is described in Section 3.

1.3 Noise Model

The noise is an unavoidable and undesired signal from the environment, which conveys no useful information and has, therefore, a detrimental effect on the speech signal of interest [104]. The noise effects can be perceived in the quality and intelligibility of those speech signals [68, 105], and its presence makes more difficult the task of estimating the speech signals parameters [6, 22, 119]. The additive noise can appear from different sources with various spectral shapes [98, 104, 149]. It can also be of stationary or non-stationary nature across the dimension of time [169]. For example, babble noise, i.e., the noise resulting from various speakers in the cocktail party situation [19], has levels that are constantly varying in the time [95], resulting in very different spectral shapes across the segments, and it can be, therefore, highly non-stationary. Moreover, this noise has very similar spectral content to the speech signal of interest, making the problem of enhancing the desired signal or estimating parameters [52, 79], as the fundamental frequency, more difficult. Another example is car noise [104], which is more stationary than babble noise, and it is concentrated mainly in the low spectrum. In most cases, the spectral content is predominant in specific frequency bands, which means that real noise types are typically colored [177, 179].

Another type of degradation [55] occurs when different attenuated and delayed versions of the signal of interest are received at a point of interest. This phenomenon is known as reverberation and is introduced by reflections of enclosed spaces and can have a detrimental effect on distinguishing between different speech sources [97, 143]. In this case, the distortion is of convolutive nature, and therefore, the methods which were derived for the additive noise assumption cannot suppress the resulting interference. Many studies have addressed the dereverberation problem [80, 120], to recover the original signal. We will here focus only on the additive noise case.

To facilitate the mathematical tractability and the possibility of having fast implementation algorithms in some estimation problems, it is common to assume that the noise is white and Gaussian (WGN) [21, 22, 56, 125, 179], i.e., that the noise has an uniform spectral content over all the frequencies. Of course, this assumption is often too simple, as it is violated in most of the acoustic scenarios. This WGN condition, however, should be merely regarded as a representation of the assumptions instead of an attempt of modelling the physics behind the problem [14]. In fact, the white Gaussian distribution results in a higher CRLB compared to more general noise models [32, 157], being the worst-case scenario [92, 153]. A better estimation performance could be obtained by making the correct assumption about the noise model, by e.g., modelling the correlation in the noise samples [14]. To keep working with the simple and fast methods derived under a WGN assumption, and retaining simplified models, a possibility is to apply a pre-whitening filter [22, 139, 177] so that the colored noise is rendered closer to white. For example, the colored noise can be well modelled as an autoregressive (AR) process [88, 150, 177], i.e.,

$$c(n) = -\sum_{i=1}^{P} a_c(i)c(n-i) + e(n), \tag{6}$$

where $\{a_c(i)\}_{i=1}^{P}$ are the P AR coefficients of the noise signal and $e(n)$ is a driving WGN process with variance σ_e^2. In a mixture signal, where the additive colored noise is combined with the speech signal of interest, the all-zero pre-whitening filter with frequency response $1 + \sum_{i=1}^{P} a_c(i)e^{-j\omega i}$ can be applied so that the noise component is rendered closer to white [69, 90, 177]. Other possibilities for pre-whitening are described later.

As described above, the additive noise term $c(n)$ is added to the speech signal of interest, so the observed signal is

$$x(n) = s(n) + c(n). \tag{7}$$

Either the harmonic model (1) or the hybrid speech model (4) are considered as the $s(n)$ part in this model.

Many efforts have been devoted to reduce the background noise levels [25, 60, 79, 104] of noisy speech recordings, i.e., to recover the signal $s(n)$ from the noisy observation $x(n)$. Speech enhancement algorithms serve for that purpose, and they are typically used as pre-processors in applications such as automatic speech recognition (ASR) [69], speaker diarization [3] and speech coding [178] systems. Very often, it is believed that enhancing the noisy signal is the best way to combat the noise presence so that the parameters of the signal of interest $s(n)$ can be better estimated and the capacity of detecting the speech presence is improved [70]. Although using speech enhancement as a pre-processing method, may provide a benefit to parameter estimation methods, such pre-processor may not be always the best choice to deal with the problem of noise presence. At low signal-to-noise ratio (SNR) values (e.g., negative SNRs), the recovery of the signal is more difficult since the noise term dominates in the observation, and it is very likely to incur in a high distortion [73, 160], i.e., the original speech characteristics may be highly modified. In those cases, the AR parameters describing the noise spectral shape could be estimated, and then applied as a pre-whitening filter. Such pre-processing may be more desirable, and could lead to obtaining better estimates of the speech signal parameters. In paper D, we investigated which pre-processing method (i.e., speech enhancement or pre-whitening) should be applied on noisy speech recordings in order to estimate with a better accuracy the parameters of the clean speech signal.

To derive time-varying filters for either cleaning up the noisy signal (i.e., speech enhancement), for rendering the noise closer to white (i.e., pre-whitening), or for decomposing noisy speech into is voiced and unvoiced components, a knowledge of the noise statistics is required. The problem of estimating those noise statistics is particularly difficult when the noise is non-stationary (e.g., babble noise) [104]. The noise statistics can include the noise covariance matrix [16, 79] and the noise power spectral density (PSD) [85, 124]. A vast amount of methods for obtaining these statistics have appeared in the recent decades [26, 47, 106, 124], and they can be mainly dichotomised as supervised and unsupervised methods. The main difference between them resides in that the supervised methods make use of information on noise and speech spectral features, as e.g., AR coefficients, obtained from a training step [115, 150]. Followingly, some of the unsupervised and supervised approaches are described. We also give an overview of the principles of non-negative matrix factorization (NMF) [42, 101] algorithms, which are often required in some of the supervised methods.

2 Noise Statistics and Speech Signal Parameter Estimation

2.1 Unsupervised noise PSD estimation

In applications of speech and audio processing, the noise PSD is normally time-varying, i.e., it is changing on a segment-by-segment basis [61]. A certain segment of N samples, $\mathbf{x} = [x(0)\ x(n)\ \ldots\ x(N-1)]^T$, i.e.,

$$\mathbf{x} = \mathbf{s} + \mathbf{c}, \tag{8}$$

is formed from the mixture of the clean speech signal vector $\mathbf{s}$ and the additive noise vector $\mathbf{c}$, where $\mathbf{s}$ and $\mathbf{c}$ are defined in the same way as $\mathbf{x}$, and $(\cdot)^T$ denotes the transposition operator. A typical way of defining the noise PSD, for an individual segment, is [124, 157]

$$\Phi_c(\omega) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[\left| \sum_{n=0}^{N-1} c(n) \exp\left(-j\omega n\right) \right|^2 |\mathbf{x} \right], \tag{9}$$

where $\mathbb{E}$ denotes the expectation operator.

Early unsupervised approaches [43, 49, 84] relied on the assumption that the noise behaves in a more stationary way compared to speech [61]. Thus, in those methods, a noise PSD quantity which was estimated from some previous segments could still be valid at a current one. Another assumption often made in unsupervised noise PSD estimators is that the speech and noise DFT coefficients are modelled as Gaussian random variables [39]. Next, we describe the most well-known approaches, starting first with simple voice activity detector (VAD) methods [147] and then describing some of the well-known algorithms which update the noise PSD in a continuous way across all the segments of the noisy speech signal, i.e., even in periods of speech activity [26, 47, 106, 124]. These last approaches are more suited not only for noise reduction, but also for the task of applying a pre-processing scheme which counteracts the noise spectral shape to render the noise closer to white. This is highly desirable, e.g., when the fundamental frequency of speech signals is continuously tracked [144], if it is desired to keep using parametric methods designed on the assumption that the noise is WGN [22].

VAD methods

The simplest approach for noise PSD estimation is through a voice activity detector (VAD) [147]. From simply using a VAD, the noise statistics are only estimated in segments which have absence of speech (i.e., silent segments). That is, they will not be updated in periods of speech presence

and the same values which were obtained when speech was absent will still be used during voice activity periods. To decide if speech is present or not, and based on the Gaussian assumption of the DFT coefficients, a logarithmic average of a likelihood ratio at all frequency bins is compared to a dynamically adjusted threshold. Despite its simplicity, this approach may be problematic in low SNR conditions [104], or when there is a long duration of uninterrumpted speech. A recursive smoothing can then be used to update the noise PSD estimate, during the periods of speech absence. However, under this approach, rapid variations of the noise PSD level may not be detected, specially when there is a sustained continuity of speech presence [173].

Minimum level tracking methods

The approach based on minimum statistics [106] uses a collection of consecutive smoothed periodogram values at the different frequency bins to continuously track the minimum value, which would then correspond to the noise spectral power. This relies on the assumption that over a sufficiently large time-span which comprises several segments (e.g. 1 s), there is at least some small speech absence interval which only includes noise [108]. However, a bias compensation is necessary since the minimum value is below the true mean [169]. Additionally, a proper smoothing parameter needs to be selected, according to the absence or presence of speech. An insufficient smoothing could lead to high variance estimates, while the minimum could be smoothed out to the wrong estimate when too much smoothing is applied [106]. The minimum statistics (MS) approach is accurate in intervals when the noise is slowly changing (i.e., when the noise has a stationary behavior), but some quite long delays will typically occur when the noise power suffers rapid and abrupt changes [61, 124].

Methods based on speech presence probability (SPP)

As opposed to binary decisions made in the VAD, a noise PSD estimator can be made dependent on speech presence probabilities (SPP) [26, 47, 51, 173]. A first effort based on soft decisions was introduced in [148], which also relied on a frequency dependent and adaptive smoothing parameter. An improvement was suggested by Cohen in [26], where the smoothing weight is time-frequency dependent and determined by the estimated SPP. Additionally, the speech absence probability is adjusted by minimum tracking, in a similar way to the MS approach [106]. This requires two iterations, in which at first a rough VAD per bin is applied and then the strong speech components are excluded. This approach is referred as IMCRA [26], but this algorithm still had similar problems of following rapid variations of the noise level since it is based on the same MS principle.

An improved method based on SPP was introduced in [47], where the *a posteriori* SPP was obtained from a fixed *a priori* SNR. Due to this, it is possible to get values of *a posteriori* SPP close to zero during speech absence periods, having a better tracking of the noise PSD, and not requiring to include a bias compensation step. Another advantage of fixed priors is that the noise level tracking is independent of subsequent steps of pre-processing based on the estimated noise levels (e.g., either speech enhancement or pre-whitening).

MMSE-based methods

Based on the assumed Gaussian distribution of the noise and speech DFT coefficients, [47, 64] introduced a minimum mean-square error (MMSE) estimate of the noise PSD. This estimate depends on both the *a posteriori* SNR $\gamma(k)$ and *a priori* SNR $\xi(k) = \frac{\lambda_s^2(k)}{\lambda_c^2(k)}$, and its solution is a weighted combination of the the prior noise PSD $\lambda_c^2(k)$ and the current periodogram element of the noisy signal $\Phi_x(k)$, i.e.,

$$\Phi_c(k) = \left(\frac{1}{1+\hat{\xi}(k)}\right)\Phi_x(k) + \left(\frac{\hat{\xi}(k)}{1+\hat{\xi}(k)}\right)\lambda_c^2(k), \tag{10}$$

However the *a priori* SNR $\xi(k)$ is obtained from a limited ML estimate $\hat{\xi}(k) = \max(0, \hat{\gamma}(k) - 1)$, leading to a binary instead of a soft decision between the prior PSD estimate $\lambda_c^2(k)$ and the current periodogram $\Phi_x(k)$. Final steps of bias compensation and a safety-net were required to overcome to the fact that the noise PSD is only updated during speech absence periods. A final recursive averaging is required to get an estimate of the noise PSD [48]. An unbiased method which was described in the previous paragraph, and also proposed in [47], was instead based on soft SPP to overcome the limitations and computational complexity of the original MMSE-based noise PSD. In papers B and D, an MMSE approach for noise PSD estimation based on (10) is proposed in order to derive a pre-whitening filter which adapts its coefficients on a segment-by-segment basis. However, the computation of $\xi(k)$ and $\lambda_c^2(k)$ will depend on values obtained from a supervised nonnegative matrix factorization (NMF), which requires some training data parametrized using autoregressive (AR) coefficients. Details on supervised noise PSD estimation approaches will be discussed in the next subsection.

Histogram-based Methods

Finally, there are methods which are based on choosing the maximum value of an histogram [1, 65], i.e., the most frequent frequency bin, which would

then correspond to the noise power level. However, these methods may only work well if there are many speech pauses [169], and the performance is often degraded under low SNR scenarios and under non-stationary noise conditions.

2.2 Supervised noise PSD estimation

The unsupervised methods are not able to separate a mixed signal into its components. They often need large buffers to store previous data [26, 106] and they rely on heuristics for trying to circumvent the difficulties of time-varying noise (e.g., bias compensation or SPP) [47, 106] instead of assuming an explicit model for the problem at hand. The conventional noise trackers may, therefore, suffer from limited performance in non-stationary environments [150]. For that reason, during recent decades, supervised approaches based on *a priori* spectral information have been shown to allow for a better spectral estimation accuracy of the noise PSD and also for a higher capacity of adapting to rapid changes of the noise levels [115, 124]. Approaches based on Hidden Markov Models (HMM) [142] or on pre-trained codebooks [124, 150] have addressed the limitations of unsupervised approaches. Including prior information allows to understand better the shortcomings of a model-based estimator on an individual segment basis. In the codebook-based approaches, models parametrized by AR coefficients can be derived from some training data, which can include a specific scenario where the noise environment or the speaker of interest are known in advance, or from more general data retrieved from a database which includes different speakers or noise conditions. The performance of various noise PSD estimation methods was evaluated in [85], including a model-based approach proposed in [124] which uses pre-trained spectral shapes stored in codebooks. This recent method allowed for a rapid tracking speed, better spectral estimation accuracy and more noise reduction, specially in non-stationary scenarios, when compared to the use of unsupervised approaches [26, 47, 106].

In supervised approaches, the speech and noise PSD can be respectively represented parametrically as

$$\Phi_s(\omega) = \frac{\sigma_s^2}{\left|1 + \sum_{i=1}^{P'} a_s(i) e^{-j\omega i}\right|^2}, \quad \Phi_c(\omega) = \frac{\sigma_c^2}{\left|1 + \sum_{i=1}^{Q'} a_c(i) e^{-j\omega i}\right|^2}, \tag{11}$$

where σ_s^2 and $\{a_s(i)\}_{i=1}^{P'}$ are the spectral gain (i.e., excitation variance) and AR coefficients of speech, respectively, and σ_c^2 and $\{a_c(i)\}_{i=1}^{Q'}$ are the noise spectral gain and noise AR coefficients, respectively. During the training stage, the goal is to obtain a finite set of representative AR parameters. These are obtained from the AR parameters of various segments of speech

and noise recordings which are converted to line spectral frequency (LSF) coefficients [72], and are then given as an input to a vector quantizer, such as the Linde-Buzo-Gray (LBG) algorithm [103]. The conversion to LSF coefficients is necessary to ensure that the obtained representative AR parameters correspond to stable processes. The vector quantizer will give as an output a finite number of codebook centers which correspond to the representative AR parameters of speech and noise signals. The obtained LSF codebook centers can be converted back to AR codebook centers.

When doing the processing on testing data, and given the already pre-trained codebooks, the speech and noise parameters are usually estimated by selecting the individual entry of the codebooks which minimizes the spectral distortion between the observed and the modelled spectrum [149]. The modelled spectrum is the one which is parametrized by the individual selected codebook entries. Although this approach was initially perceived as a maximum likelihood one, a Taylor expansion for the Itakura-Saito (IS) distortion is made to allow for an approximate log-spectral distortion (LSD), which lead to an inaccurate spectral gain estimation. That method, however, can give a zero excitation variance and assign most of the power in the noise model, when silent segments are processed. In paper F, we found that such appproach has a good performance when distinguishing between codebook entries of unvoiced speech and additive noise, when the observation is a stochastic residual. A later approach based on a Bayesian MMSE formulation [150] involved a soft decision between all the codebook entries, allowing for more noise reduction in a speech enhancement framework. Although a quick tracking of rapid changing levels was possible, an inaccurate spectral gain estimate due to a poor approximation lead to considerable distortion and remaining noise. Later, He [60] proposed a solution based on a multiplicative update (MU) rule to allow for improved estimates of the excitation variances (i.e., get better spectral gain estimates). This permitted for an improved quality in the speech enhancement framework. MU approaches were motivated by their use in non-negative matrix factorization (NMF) problems. It is possible to combine the ideas from a parametrized NMF based on prior spectral shapes obtained from pre-trained codebooks [86]. Next, we describe important concepts regarding NMF, which allow to understand our contribution of a time-varying pre-whitener which is based on noise statistics estimated from pre-trained spectral shapes.

Non-negative matrix factorization (NMF)

In various applications such as music transcription [146], source separation [145] and face recognition [141], a well-known data decomposition technique, namely NMF, allows to factorize an observation matrix $\mathbf{V} \in \mathbb{R}_{\geq 0}^{F \times R}$,

containing non-negative data, into two non-negative matrices as

$$\mathbf{V} \approx \mathbf{W}\mathbf{H}, \tag{12}$$

where $\mathbf{W} \in \mathbb{R}_{\geq 0}^{F \times K}$ and $\mathbf{H} \in \mathbb{R}_{\geq 0}^{K \times R}$ are, respectively, referred as the basis matrix and the activation matrix. F, K and R contain the dimensions. The dimension reduction is possible from choosing $FK + KR << FR$. The different observation vectors are contained as columns of $\mathbf{V}$, allowing to represent a single column element as $\mathbf{v} = \mathbf{W}\mathbf{h}$, where $\mathbf{h}$ is the corresponding individual column of $\mathbf{H}$ which describes the amount that each basis vector contained in $\mathbf{W}$ is required for such representation. The total number of basis vectors corresponds to K, which is the common dimension of $\mathbf{H}$ and $\mathbf{W}$. In speech and audio applications, it is common to choose F as the number of frequency points representing the spectrum [171], and $\mathbf{V}$ will contain in each of its columns, the periodogram or power spectrum of an individual noisy segment. The dimension R usually corresponds to the number of processed segments. We will herein focus in the case where the matrix $\mathbf{W}$ is trained, so only the individual activation coefficients contained in each column of $\mathbf{H}$ need to be computed. This approach is referred as supervised NMF [42], and the approximation in (12) is done under a β-divergence criterion, such as the Kullback-Leibler (KL) divergence, Itakura-Saito (IS) divergence or under the Euclidean distance.

In [41], it is shown that when the observation is modelled as a superposition of Gaussian components, the ML estimation of $\mathbf{H}$ corresponds to NMF under the IS divergence criteria. The involved mathematical problem can be solved using numerical optimization techniques based on the multiplicative gradient descent or on the expectation-maximization (EM) algorithm. Using MU rules, given the observed matrix $\mathbf{V}$ and the pre-trained basis matrix $\mathbf{W}$, the matrix $\mathbf{H}$ can be iteratively computed as (after it has been initialized with random nonnegative entries)

$$\hat{\mathbf{H}}^{(i+1)} \leftarrow \hat{\mathbf{H}}^{(i)} \odot \frac{\mathbf{W}^T \left(\mathbf{W}\hat{\mathbf{H}}^{(i)}\right)^{[\beta-2]} \odot \mathbf{V}}{\mathbf{W}^T \left(\mathbf{W}\hat{\mathbf{H}}^{(i)}\right)^{[\beta-1]}}, \tag{13}$$

where i corresponds to the index of the iteration and $\odot$ denotes the element-wise multiplication. Both division and exponentiation are also entry-wise operations. The case with $\beta = 0$ corresponds to the IS divergence. In papers B and D, we introduced a time-varying pre-whitening filter based on a noise PSD estimate obtained from the supervised NMF of the matrix $\mathbf{V}$. Such matrix is the one containing the noisy periodogram of all the segments, in each one of its columns. The spectral basis matrix $\mathbf{W}$ contains in each of its columns, a gain-normalized spectral shape parametrized by AR coefficients, defined as in (11) but with unitary excitation variances.

This is possible, as it is shown that if the noisy signal is formed by a superposition of autoregressive processes, the maximisation of the data likelihood corresponds to doing an NMF of the observed periodogram under the IS divergence as the criterion of optimization. After estimating the activation coefficients matrix from (13), the noise PSD can be obtained from the MMSE estimate in (10). In this case, the *a priori* SNR $\xi(k)$ (i.e., per individual frequency bin) can be obtained from individual elements of both $\mathbf{W}$ and $\mathbf{H}$. It is important to remark that as opposed to [47, 64], the way that the noise PSD is computed does not require bias compensation or a safety-net, and allows for a faster tracking speed.

2.3 Speech signal parameters

We now turn to the case of the speech signal of interest. For the case of voiced speech, we are interested in the fundamental frequency, the model order, the amplitudes and phases. We may also be interested in deciding if a segment is voiced or not, which will depend if the estimated harmonic model order is different from 0. For unvoiced speech, the AR parameters from an all-pole model may also be needed [100], as we could be interested in extracting the components produced by different excitation sources [77].

The knowledge of the fundamental frequency (f_0) is required in several applications, such as detection of speech disorders [137], automatic speech recognition [50], noise reduction [59, 79, 94, 151] and speaker diarization [66]. f_0 corresponds to a physical feature which describes how fast the vocal folds vibrate. Very often, the term pitch is used interchangeably with the fundamental frequency, however the term pitch is associated to perception. A wide range of f_0 estimation methods have appeared in the recent decades. The more traditional ones are non-parametric methods. Many of them are based on the similarity between the original signal and a delayed version of it. Some of the used similarity measures are the cumulative mean normalized mean difference function, cross-correlation and autocorrelation, which are found in the classic algorithms YIN [18], RAPT [164] and PRAAT [12], respectively. A voicing decision step and a final smoothing are often used (e.g., dynamic programming [121] in RAPT) to refine the estimated values. Although these time-domain approaches present a low computational complexity, they are not robust to the noise, have poor resolution and are prone to sub-harmonic errors [123, 125]. Other approaches exploit the presence of harmonic peaks in the spectral representation, in either the cepstral [126] or the frequency domain [15, 52]. For example, SHRP computes a ratio between sub-harmonics and harmonics [161], because a perceptual experiment has shown that f_0 is perceived differently when that ratio is above 0.4. SWIPE attempts to match the entire signal spectrum to a modified cosine kernel that ignores all non-prime harmonics except the first

one [15], employing a different window size for each f_0 candidate to reduce the number of sub-harmonic errors. The recently introduced PEFAC [52] convolves the log-frequency power spectrum with a comb filter that sums the energy of the harmonics and ignores broadband noise with a smooth spectrum.

Most of the aforementioned methods do not take into account a mathematical model for the speech signal segments. In that sense, those methods can be typically considered as non-parametric. On the other side, parametric methods [21, 22, 125] take into account a mathematical model for the signal and for the noise. Methods based on optimal filtering [22] and partitioning of signal and noise subspaces [23] have been proposed, although they may present time-frequency resolution problems and they are not robust at low SNRs [20, 21]. The best resolution and robustness to the noise can be achieved with estimators based on the maximum likelihood (ML) principle, which offer better performance than the subspace-based or optimal filtering methods, under adverse conditions [20, 21]. To simplify the mathematical problem of the ML f_0 estimator, the noise affecting the signal is often assumed to be WGN. Assuming an harmonic model for the voiced speech parts, the estimator is derived from the observed signal likelihood function which is parametrized by a vector containing the L amplitudes and phases of the different harmonics, and f_0. Under the simple WGN assumption, maximizing the likelihood with respect to the unknown parameters is equivalent to minimizing the 2-norm between the observed vector and the signal model. As the resulting cost function is a nonlinear function of f_0, the formulated problem is equivalent to a nonlinear least-squares (NLS) estimation problem [157]. To solve this problem, a grid search for all the possible candidate f_0 and model orders L is needed, and solving this problem for all candidate model orders seems to be computationally demanding at a first insight. Fortunately, [123, 125] recently introduced a fast order-recursive algorithm which achieves a similar computation time to the well-known harmonic summation [127]. After estimating the best f_0 for each candidate model order, the model order L which best explains the data can then be estimated by information theoretic criteria, as the minimum description length (MDL) or the Bayesian Information Criteria (BIC) [122, 156]. From this model selection, it is possible to choose the not-voiced model (speech pauses or pure unvoiced speech segment), i.e., $L = 0$. For the real signal model, the NLS estimator is able to reach the CRLB even for low values of f_0 or low segment lengths [20, 21]. This does not happen for non-parametric estimators such as YIN, as their performance exhibit a gap from the CRLB [125].

The WGN assumption in the NLS f_0 estimation problem seems very attractive from both a computational perspective and its good statistical performance, observed from experiments with synthetic signals. Unfortunately,

the WGN assumption is violated in most of the real acoustic scenarios. In general, the noise and including the stochastic parts of speech, is spectrally colored, and can therefore be better modelled as an AR process [100, 149]. In this case, the ML f_0 estimates do not longer correspond to the estimates obtained from the NLS f_0 estimator. In the colored noise scenario, the optimal approach is to jointly estimate the sinusoidal (f_0, L, $\{\alpha_l\}_{l=1}^{L}$) and noise ($\{a_c(i)\}_{i=1}^{P}$) parameters [74, 87, 89]. This is a problem of mixed-spectrum estimation, which has been addressed for the case of independent sinusoids. Below, we give an insight of the problem related to independent harmonic components (i.e., not harmonically related), and we shed light in what needs to be considered for the case of harmonically related sinusoids.

The f_0 estimation performance is typically assessed using two error metrics [172]: the Voicing Decision Error (VDE) and the Gross Pitch Error (GPE). Under some conditions, the estimators may present a low GPE but a high VDE. Therefore, [172] introduced a composite metric referred as the FFE (f_0 Frame Error), which takes into account all possible types of errors. In papers A and B, the performance of f_0 estimators is only evaluated in term of the GPE, while in Papers C and D, the three metrics are evaluated.

Mixed Spectrum Estimation

In the case of L independent complex sinusoids $s(n) = \sum_{i=1}^{L} \alpha_l e^{j2\pi f_i n}$ corrupted by AR noise $c(n) = -\sum_{i=1}^{P} a_c(i)c(n-i) + e(n)$, where $e(n) \sim \mathcal{N}(0, \sigma^2)$, the likelihood function of the observed complex vector $\mathbf{x} = \mathbf{s} + \mathbf{c}$ is given by [88, 89]

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{(\pi\sigma^2)^{N'}} \exp\left\{ -\frac{1}{\sigma^2} \sum_{n=p}^{N-1} \left| \sum_{k=0}^{P} a_c[k](x[n-k] - s[n-k]) \right|^2 \right\}, \quad (14)$$

where $a_c[0] = 1$ and $N' = N - P$. The unknown parameters are contained in the vector $\boldsymbol{\theta} = [\boldsymbol{\alpha}^T \ \mathbf{f}^T \ \mathbf{a}_c^T \ \sigma^2]$, where $\boldsymbol{\alpha} = [\alpha_1 \ \cdots \ \alpha_L]^T$, $\mathbf{f} = [f_1 \ \cdots \ f_L]^T$ and $\mathbf{a}_c = [a_c[1] \ \cdots \ a[P]]^T$. The unknown frequencies can be found by the minimization of [88]

$$J(\mathbf{f}) = \mathbf{x}_P^H \left[\boldsymbol{\Pi}_E^{\perp} - \boldsymbol{\Pi}_E^{\perp} \mathbf{H} (\mathbf{H}^H \boldsymbol{\Pi}_E^{\perp} \mathbf{H})^{-1} \mathbf{H}^H \boldsymbol{\Pi}_E^{\perp} \right] \mathbf{x}_P, \quad (15)$$

with $\boldsymbol{\Pi}_E^{\perp} = \mathbf{I} - \mathbf{E}(\mathbf{E}^H\mathbf{E})^{-1}\mathbf{E}^H$, where $\mathbf{x}_P = [x(P) \ x(P+1) \ \cdots \ x(N-1)]^T$ and

$$\mathbf{H} = \begin{bmatrix} x(P-1) & x(P-2) & \dots & x(0) \\ x(P) & x(P-1) & \dots & x(1) \\ \vdots & \vdots & \ddots & \vdots \\ x(N-2) & x(N-3) & \cdots & x(N-1-P) \end{bmatrix},$$

$$\mathbf{E} = \begin{bmatrix} e^{j2\pi f_1 P} & e^{j2\pi f_2 P} & \dots & e^{j2\pi f_L P} \\ e^{j2\pi f_1 (P+1)} & e^{j2\pi f_2 (P+1)} & \dots & e^{j2\pi f_L (P+1)} \\ \vdots & \vdots & \ddots & \vdots \\ e^{j2\pi f_1 (N-1)} & e^{j2\pi f_2 (N-1)} & \dots & e^{j2\pi f_L (N-1)} \end{bmatrix}.$$

Unfortunately, the above problem would require a multidimensional search for all candidate model orders $\{1, \cdots, L\}$ and AR orders $\{1, \cdots, P\}$. To address this, [74] proposed an iterative algorithm based on relaxation techniques, in which simple FFT operations were used to estimate the different sinusoids at individual steps (i.e., one single sinusoid per step). This solution was guaranteed to reach the global minimum of the cost function, even from ignoring the estimated AR noise parameters (i.e., the noise spectral shape) at the second step of the decoupled parameter estimation (DPE). That is, a reiteration between the estimation of noise AR parameters and sinusoidal parameters is not required, since the performance of the obtained frequency estimates reach asymptotically the CRLB even if the noise is colored [154].

As opposed to the case of independent sinusoids [74, 89], in this work we are interested in the case of L harmonically related sinusoids (i.e., $f_i = if_0, i = 1, \cdots, L$), being of evident interest the frequency with the lowest value, i.e., the fundamental frequency f_0. In this case, ignoring the noise spectral shape when estimating f_0 will most probably lead to selecting a wrong peak in the cost function as an estimate, i.e., sub-harmonic errors are most susceptible to appear. In paper A, we verified that when the noise spectral shape is considered, the number of gross errors of f_0 estimates (which also include the sub-harmonic errors) from the NLS estimation problem is considerably reduced. The details on how the noise spectral shape needs to be taken into account will be described in the next section. Using the NLS f_0 estimator is convenient from a computational perspective, since a fast implementation for all the candidate model orders is available [125]. Additionally, as opposed to [74], in the problem of f_0 estimation in colored noise, the reiteration between estimates of f_0 and the AR parameters of the noise may be necessary to achieve the ML solution. Through the iterative refined parameter estimation, the likelihood of the observed data never decreases and this may guarantee a convergent solution which resembles the ML one. This issue was investigated in papers C and D, and by applying an iterative re-estimation scheme, the number of gross errors and voicing detection errors was reduced compared to the case of a single cascade estimation of noise AR parameters and f_0 estimates.

As an example, in the Figure 1 we plot the ground truth of the f_0 values of a female excerpt from the Keele database [134] and the resulting estimates when the excerpt is corrupted by factory noise at an SNR of 5 dB. The unvoiced and silent segments do not have a fundamental frequency,

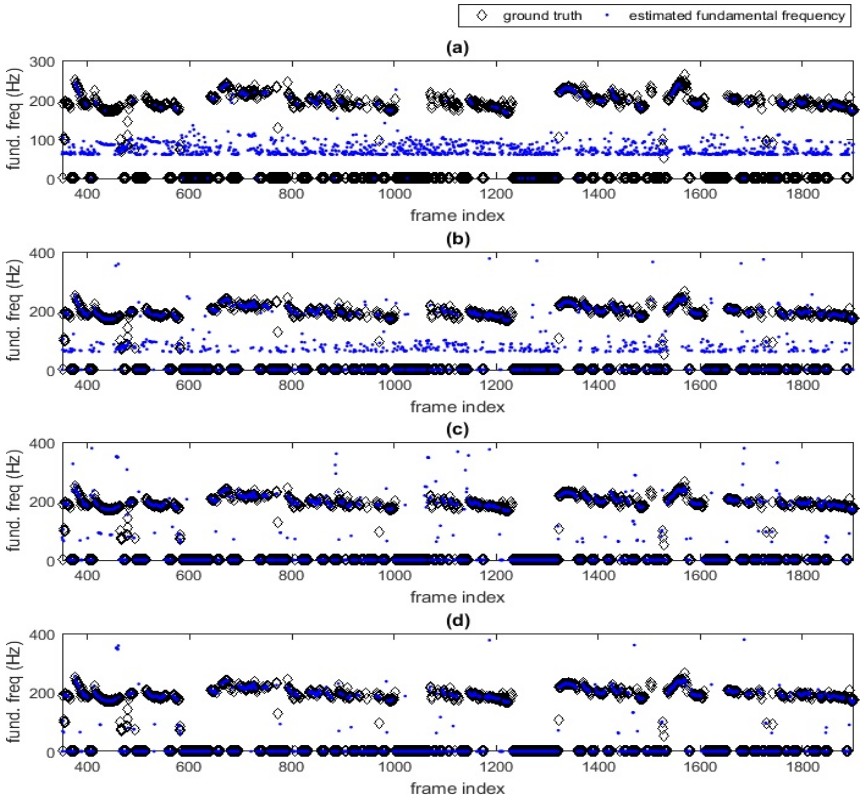

Fig. 1: Fundamental frequency ground truth from a female excerpt of the Keele database [134] and f_0 estimates when the excerpt is corrupted by factory noise [168] at 5 dB. (a) NLS f_0 estimates ignoring the noise colour, i.e., from a wrong WGN assumption, (b) Re-estimates of f_0 from the initial estimates of plot (a), (c) estimates from a single cascade step of noise AR estimates followed by f_0 NLS pitch estimates and (d) estimates based on approximate ML approach (i.e., applying re-iteration).

but to facilitate visualization here they are represented with a value of 0. The plots (a) and (b) correspond to the case when first the f_0 is estimated (without and with reiteration), while the plots (c) and (d) correspond to the case when first the noise AR parameters are estimated (without and with re-iteration). As opposed to the case of independent sinusoids [74], it is evident that the best accuracy is achieved from first estimating the noise AR parameters followed by the NLS f_0 estimation (plot (c)) and doing the re-estimation (plot (d)). More details are explained in the different papers.

In many cases, the f_0 values and the voicing state may form a correlated sequence, and it is possible to impose temporal continuity constraints on them. This has been addressed in some Bayesian tracking algorithms which were designed on a WGN assumption [144, 162]. To apply these methods

under real noise conditions, some form of pre-processing (e.g. pre-whitening) needs to be applied beforehand.

2.4 Speech signal and noise statistics in the time domain

Very often, it is desired to recover the desired signal of interest or the undesired one (e.g. estimate the noise signal) from the noisy observation. For this purpose, estimates of the statistics of the noise and speech signal are required. If the processing is done in the frequency domain, the PSD is utilized. However, in many cases the processing is applied in the time domain, and in such case a covariance matrix estimate is instead required.

From a sub-vector of M samples (where $M < N$), the $M \times M$ noisy signal covariance matrix $\mathbf{R}_{\mathbf{x}} = E\left[\mathbf{x}(n)\mathbf{x}^T(n)\right]$ is commonly estimated [22, 155] from approximating the mathematical expectation definition with a sample average defined as

$$\bar{\mathbf{R}}_x(n) = \frac{1}{N-M+1}\sum_{n=0}^{N-M}\mathbf{x}(n)\mathbf{x}^H(n), \tag{16}$$

where $M < N/2+1$ to guarantee that this matrix can be inverted. As speech is non-stationary, this sample average should be done on a short temporal basis to follow the speech signal variations [17]. A recursive approach can also be used, in which a forgetting factor $0 < \alpha_x < 1$ controls how the previous samples will influence in the current covariance matrix at a particular time instant. When the processing is done on a segment-by-segment basis, the local time average and the recursive smoothing can be combined [17] to yield a segment covariance matrix estimate:

$$\hat{\mathbf{R}}_x(n) = \alpha_y\hat{\mathbf{R}}_x(n-1) + (1-\alpha_x)\bar{\mathbf{R}}_x(n). \tag{17}$$

From the assumption that the desired speech signal $s(n)$ and the additive noise $c(n)$ are uncorrelated, it follows that their covariance matrices are additive, i.e.,

$$\mathbf{R}_x(n) = \mathbf{R}_s(n) + \mathbf{R}_c(n), \tag{18}$$

in which $\mathbf{R}_c(n)$ can be estimated from the noise PSD, if it was estimated in the frequency domain by using supervised or unsupervised approaches. This is possible from the inverse Fourier transform between the noise PSD and the noise autocorrelation [157], and the different entries of $\mathbf{R}_c(n)$ would correspond to a particular time lag [128]. The inverse of the noise covariance matrix can also be parametrized with the AR parameters by using the Gohberg-Semencul (GS) formula [157]. From the statistical independence assumption, the speech signal covariance matrix can be estimated as

$$\hat{\mathbf{R}}_s(n) = \hat{\mathbf{R}}_x(n) - \mathbf{R}_c(n). \tag{19}$$

In the case of a quasiperiodic signal, the covariance matrix can be modeled as $\mathbf{R}_v(n) = \mathbf{Z}(\omega_0)\mathbf{P}\mathbf{Z}^H(\omega_0)$ [157], where $\mathbf{P}$ is the covariance matrix of the amplitudes, defined as $\mathbf{P} = E\{\mathbf{a}\mathbf{a}^H\} = \frac{1}{4}\,\mathrm{diag}([A_1^2\, A_1^2 \ldots\, A_L^2\, A_L^2])$ [21] and $\mathbf{Z}$ is a matrix of Fourier vectors which contains the different L harmonic frequencies. As voiced speech may not be perfectly periodic across a fixed length segment, a similar recursive smoothing to (17) can be applied to this short-term based estimate to take the non-stationarity character of speech into account. In paper E, we used the difference between $\hat{\mathbf{R}}_s(n)$ and $\mathbf{R}_v(n)$ to estimate the covariance matrix of the stochastic speech parts (i.e., unvoiced speech).

3 Processing based on signal parameters and noise PSD estimates

Given the observed signal statistics, the noise statistics (PSD or covariance matrix), and the parameters of the speech signal, it is possible to either apply speech enhancement, apply a pre-whitening scheme or decompose the speech into its voiced and unvoiced components. Each type of processing is described separately.

3.1 Speech Enhancement

From the estimated noise PSD or noise covariance matrix, the noisy signal can be pre-processed for some useful purpose. The most typical way of pre-processing the signal has been to apply speech enhancement. From the noisy observation in (7), the objective of speech enhancement is to recover the desired signal $s(n)$, trying to attenuate the noise component $c(n)$ as much as possible while having the less possible amount of distortion in the desired component [16]. The enhancement can be done in the time [16] or in the frequency domain [75]. The initial efforts were based on simple spectral subtraction of the noise spectrum from the noisy spectrum [13], but shortcomings such as musical noise were observed. Additionally, a loss of intelligibility is observed since unvoiced parts of speech and high frequency formants are de-emphasized [59].

Rather than using heuristic principles, $s(n)$ can be estimated in an optimal way where a linear filter which minimizes the mean squared error (MSE) between the desired and estimated signal is applied. This is the classical Wiener filter $\mathbf{h}_W$ [16], and applying it to a vector observation in (8) will lead to the estimate $\hat{\mathbf{s}} = \mathbf{h}_W^T x$. The estimation error is $\mathbf{e} = \mathbf{h}_W^T\mathbf{x} - \mathbf{s}$, so the seeked MSE to be minimized is $E[\left(\mathbf{h}_W^T\mathbf{x} - \mathbf{s}\right)^2]$. Although the solution of this problem is optimal in the MSE sense, it may introduce considerable levels of distortion in the desired signal. This can be overcome by using

filters which can trade off between signal distortion and noise reduction by introducing an additional tuning parameter [8]. The Wiener filter can also be formulated in the frequency domain, and in that case it is expressed as [102]

$$H(\omega_k) = \frac{\Phi_{ss}(\omega_k)}{\Phi_{ss}(\omega_k) + \Phi_{cc}(\omega_k)}, \tag{20}$$

so that the speech spectrum can be estimated as $\hat{S}(\omega_k) = H(\omega_k)X(\omega_k)$, yielding a linear relationship between the observed data and the estimates.

As opposed to the Wiener filter, nonlinear estimators [104] have been derived by modeling the probability density function (pdf) of the speech DFT coefficients, being possible Gaussian [25, 39] and supergaussian [107] distributions. Instead of estimating the complex spectrum, the magnitudes of the DFT coefficients have been estimated based on ML [109], MMSE [39] and log-MMSE [38] criteria, which differ in the optimization criteria they rely on, and on how they handle the parameters of interest. For example, the MMSE estimator [39] assumed that the magnitudes of interest were random variables, as opposed to the ML [109], which assumed that they were deterministic. Compared to the ML and the typical MMSE estimator, and motivated by the human perception, an MMSE criteria of the log-magnitude spectra (log-MMSE) [38] allowed for more noise attenuation without distorting speech too much. The previous nonlinear estimators do not take into account the fact that speech can have several pauses, even if there is speech activity (e.g., stop closures before the stop consonants bursts), and this caused the appearance of estimators which instead of a hard decision gain, used a soft one which rely on the correlation of the probability of speech presence between segments [27]. A well-known approach in this sense is the optimally modified log-spectral amplitude estimator (OMLSA) [25], whose gain function at a bin k and segment l is specified by $G(k,l) = \{G_{LSA}(k,l)\}^{p(k,l)} \cdot G_{min}^{1-p(k,l)}$, where G_{LSA} is the same gain of the log-MMSE estimator, but this time is powered to the SPP $p(k,l) \in [0,1]$. An additional parameter with small value G_{min} defined by subjective criteria is used to take into account the possibility of speech absence. Compared to the previous nonlinear estimators, OMLSA yielded a higher segmental SNR improvement [27], therefore distorting less the signal (without attenuating weak speech components) and achieving more noise reduction, particularly at low SNR levels [25, 104]. In paper B, we combine a pre-whitening step with OMLSA-based speech enhancement based on the WGN assumption, to show an application of the presented pre-whitener. In paper D, we also estimate the f_0 after applying a pre-processing step based on OMLSA-based speech enhancement to the noisy observation, to determine if for the f_0 estimation purpose this kind of pre-processing could be more convenient than pre-whitening. In paper F, we applied

OMLSA-based enhancement as a pre-processor for speech decomposition methods [37, 175] which ignore the presence of additive noise.

Recently, speech enhancement based on linear filtering was combined with principles of subspace-based methods which make use of the eigenvectors of the joint diagonalization of the noise and speech signal covariance matrices [9]. In that case, by choosing a different number of eigenvectors, it is possible to better trade off between the speech signal distortion and the noise reduction under a more general framework than simply using a sub-optimal Wiener filter [8, 16]. Such framework is known as the variable span linear filtering framework. In paper F, we used all the eigenvectors and eigenvalues from the joint diagonalization of the voiced part covariance matrix and the stochastic part (superposition of unvoiced and noise components) covariance matrix represented by AR parameters, in order to derive a Wiener filter for extracting the voiced speech part. This is, however, more relevant for the speech decomposition problem, which will be described in subsection 3.3.

Other enhancement approaches [79, 128], formulated as linear filtering problems, are based on models of the signal of interest, as e.g., the harmonic model which describes voiced speech. Optimal filters reminiscent of the Capon beamformer and the linearly constrained minimum variance (LCMV) filter [45] used in multi-channel setups, can be used for that purpose, however as they rely on f_0 and the number of harmonics L, they would require accurate estimates of these parameters. The filter for obtaining a single sample of voiced speech is obtained from the solution of an optimization problem and results in the solution:

$$\mathbf{h}_{\text{LCMV}} = \mathbf{R}_w^{-1}\mathbf{Z}(\omega_0)(\mathbf{Z}(\omega_0)^H\mathbf{R}_w^{-1}\mathbf{Z}(\omega_0))^{-1}\mathbf{1}, \tag{21}$$

where $\mathbf{1}$ is a vector of ones, and $\mathbf{R}_w$ is the covariance matrix of the stochastic parts of noisy speech. In paper C, we noted that using an harmonic decomposition LCMV filter [79] could also be required for re-estimating iteratively the f_0 and obtaining similar estimates to the approximate ML method. This implies that before trying to extract a desired periodic signal, these filters could be instead firstly used to allow for a more robust f_0 estimation in colored noise scenarios. Another set of speech enhancement methods are those which consider a combined deterministic-stochastic speech model [63, 94, 110, 151].

3.2 Pre-whitening

In many cases, it is desired to uncorrelate the samples of a desired or undesired signal prior to another processing step. Therefore, another way of pre-processing the signal known as pre-whitening can be applied, and it also demands a knowledge of the noise statistics. The problem of pre-whitening

arises in direction of arrival (DoA) methods [44], multiuser communications [35], sonar [90], biomedical engineering [11] and speech enhancement using subspace methods [58, 76]. A linear whitening transformation $\mathbf{W}$ can be applied to a vector $\mathbf{c} \in \mathbb{R}^N$ so that $\mathbf{b} = \mathbf{W}\mathbf{c}$ has a covariance matrix $\mathbf{R}_b = \sigma^2 \mathbf{I}_N$. Different $\mathbf{W}$ are possible, for example, based on the eigenvectors problem or on the Cholesky decomposition of $\mathbf{R}_c$ [36]. For example, some subspace speech enhancement methods which require that the noise is WGN, can apply the Cholesky factorization $\mathbf{R}_c = \mathbf{Q}_c^T \mathbf{Q}_c$ under the case of more general noise conditions [58, 76], where $\mathbf{Q}_c$ is a full rank upper triangular matrix. After transforming the noisy signal as $\mathbf{Q}_c^{-T}\mathbf{x}$ and applying enhancement methods based on singular value decompositions (SVD), it is necessary to undo the pre-whitening effect (i.e., de-whitening) by applying the pre-whitening inverse matrix $\mathbf{Q}_c^T$. In the generalized eigenvalue decomposition (EVD), an integrated pre-whitening step is applied when the noise and speech covariance matrices are jointly diagonalized [76].

Although being effective in whitening the noise component, applying general linear whitening transformations can have an important effect on the desired speech signal [76, 129]. This may not be convenient if one desires to use model-based estimators derived under a WGN condition, since they may contain nonlinear parameters of interest (e.g., f_0 or the time of arrival) which may be modified by, e.g., the Cholesky factorization. In such cases, pre-whitening based on linear filtering (e.g., autoregressive pre-whitening) can be more suitable, since applying this pre-processing will not distort the frequency components of the desired signal [129]. In fact, as we outlined in the f_0 estimation problem, an iterative joint estimator of f_0 and the noise AR parameters will likely result in the ML solution if first the noise AR parameters are estimated and then the estimated parameters are used as the coefficients of a pre-whitening filter. The notion of a linear filtering approach for pre-whitening the signal, comes from the fact that if a sequence $\mathbf{c}(n)$ with PSD $\Phi_{cc}(\omega)$ is passed through a stable filter $\mathbf{h}_W(n)$ (or $H_W(\omega)$ as the frequency response), then the output becomes $\mathbf{e}(n) = \mathbf{c}(n) * \mathbf{h}_W(n)$ with its corresponding PSD $\Phi_{ee}(\omega) = \Phi_{cc}(n)\,|H_W(\omega)|^2$ [165]. Here $*$ denotes the convolution operator. If $c(n)$ is a correlated sequence (i.e., spectrally colored) and it is desired to whiten it at the output, i.e., $\Phi_{ee}(\omega) = \sigma^2$ (i.e., spectrally flat), the filter $H_W(\omega)$ will be a whitening filter only if $|H_W(\omega)|^2 = \frac{1}{\Phi_{cc}(\omega)}$ [99], so that the frequency response satisfies

$$|H_W(\omega)| = \frac{1}{\sqrt{\Phi_{cc}(\omega)}}. \tag{22}$$

This can hold for any phase [99], but if it is desired to minimize the MSE between the input $c(n)$ and the output $e(n)$, the whitening filter needs to be the zero-phase filter [36]. A problem with the FIR pre-whitening filter

of (22) occurs when there are small or inaccurate values of $\Phi_{cc}(\omega)$, and it may be instead desirable to use a smoother spectrum. This is possible when the colored noise $c(n)$ is modeled as an AR process, i.e., it has a PSD of the form $\frac{1}{|A_c(\omega)|^2}$, where $A_c(\omega) = 1 + \sum_{i=1}^{P} a_c(i)e^{-j\omega i}$. Substituting this parametrized PSD into (22), will reveal that the pre-whitening filter in this case corresponds to the so-called autoregressive pre-whitening filter $A_c(\omega)$, sometimes also called the pre-whitener based on linear prediction or LPC pre-whitener.

It might be convenient to measure the amount of stochasticity in a signal after it was pre-whitened, i.e., how much will the resulting signal resemble white noise. In this sense, a measure known as the spectral flatness measure (SFM) [167] can be used to determine how uncorrelated will become the samples of the signal after such pre-processing. In an information theory context, in [34] it was shown that for the case of Gaussian signals, this measure corresponds to the multi-information growth rate of every new observed signal sample over the time. The SFM will tend to zero if a salient peak is shown along the spectral distribution, while it will be nearly one if the spectral distribution across the frequency bands is flat [130].

To apply either pre-whitening based on general transformations or based on a linear pre-filter, the noise statistics described in subsections 2.1 and 2.2 are needed. The pre-whitener which is obtained directly from the noise signal is referred to as the oracle pre-whitener, and it serves as a benchmark to bound the best possible performance of a pre-whitener. In paper A, we investigated the performance of general FIR and AR (a.k.a. LPC) pre-whiteners based on different unsupervised noise PSD estimates (e.g., MS, IMCRA and MMSE) in terms of the SFM and on how they can help in reducing the number of gross errors of the NLS f_0 estimator which was formulated under a WGN condition. In papers B and D, we investigated if pre-whitening based on supervised noise statistics, particularly based on parametric NMF, allowed to achieve a higher noise SFM and a better accuracy of the NLS f_0 estimator, when compared to the use of a pre-whitener based on unsupervised noise PSD estimates. Additionally in paper D, pre-whitening based on pre-trained spectral shapes was applied in a context of time-of-arrival (TOA) estimation [40, 81], making promising the benefits of considering prior spectral information in several applications. For more details of these contributions, we refer to Section 4.

3.3 Speech Decomposition

In some applications, instead of estimating the speech signal of interest, it could be desirable to extract separately the voiced and unvoiced components which constitute this speech signal. This decomposition has appeared to be of particular relevance in diagnosing voice pathologies [2, 116, 137], speech

synthesis [4, 33, 159], speech coding [53, 93], modification [100, 113] and speech enhancement [59, 83, 176].

In the classical speech production model, either a quasiperiodic train of glottal pulses (for voiced speech) or white Gaussian noise (for unvoiced speech) is used as a modelled source of excitation of a filter which models the vocal tract [104]. However, in real speech, a single mode of excitation does not describe accurately the generation of some sounds [4, 111]. In general, the voiced speech segments can be obtained from a mixed quasiperiodic excitation combined with a random noise component in the excitation [4, 174]. This is evident for the case of voiced fricatives (e.g., /z/ and /v/) [174] and voiced plosives (e.g., /b/) [140], which have energy in both high and low-frequency regions of the spectrum [31]. In voiced fricatives, although there is vibration of the vocal cords, the air flow is turbulent in the neighborhood of the vocal tract constriction [140] and this reveals that their spectrum display both a very low frequency formant [31] and considerable energy in the high spectrum. The coexistence of voiced and unvoiced speech also holds for some vowel sounds, in the case when there is some broadband noise aspiration component resulting from an incomplete glottal closure [29], and also for the cases of breathy voice [113], hoarse [118] and whisphered speech [77], and other forms of vocal dysphonia [113].

We remark that several speech analysis techniques sometimes refer to the unvoiced speech component as the noise speech part [37, 132], but we here use the term unvoiced speech, since one of our contributions deals with the problem of extracting the voiced and unvoiced parts of speech in the presence of additive noise. This additive noise is very often ignored in the speech analysis literature [37, 77, 174], or assumed to be insignificant, as it is assumed that clean signals are always available. Both voiced and unvoiced speech are associated to the deterministic and stochastic components of speech [110, 158, 174]. This is in accordance to the Wold decomposition theorem [136, 165], which states that a signal can be separated in something that can be fully predicted and something that cannot. The harmonic model fulfills the representation of voiced speech as the speech deterministic component since it has unknown but deterministic parameters, which if known, permit to completely predict the signal without any modelling errors. On the other hand, the stochastic parts of speech cannot be completely predicted but it is possible to model them based on their general characteristics. The relative level of the power of the deterministic to the stochastic component of speech is known as the harmonic-to-noise ratio (HNR) [96], and this characteristic has been used as an important indicator for speech pathologies [7, 117]. Other studies have used the voicing quotient (VQ) [37], which quantifies the proportion of energy in the voiced part. An accurate extraction of both components seems to be important for an accurate diagnose of illnesses [114], and for being able to synthesize

speech with a more natural sounding [67, 159].

There have been some efforts to extract the voiced and unvoiced speech parts (sometimes also referred as periodic and aperiodic parts of speech as they do not distinguish between unvoiced speech and additive noise). Some of the methods [53, 100], such as the multiband excitation (MBE) vocoder, made for each speech segment a binary frequency decision in which the frequency bins correspond to either voiced speech or unvoiced speech, but not shared simultaneously. Below a maximum voicing frequency (e.g. 3 or 4 kHz), only voiced speech could be present, while everything that remains above that frequency corresponds only to the stochastic parts. This simplification is inconsistent with the mechanism of producing speech, since the roughness characteristic is mainly present in the low-frequency region [96, 174]. The spectrum was also separated by a maximum voiced frequency in the harmonic plus noise model of [158, 159], which required a large numbers of parameters to be estimated since it assumes that the harmonic amplitudes are increasing along the segment. The method for estimating such parameters was robust in white Gaussian noise conditions.

On the other hand, there are methods which perform the decomposition, given the fact that each frequency bin has a contribution from both stochastic and deterministic speech components [77, 174]. The method of [29, 174] operates directly on the residual signal, directly obtained from linear prediction techniques (i.e., an all-zero whitening filter). The argument to do this is that as both signal components are generated at the source level, the decomposition should be done directly into it instead of the speech signal. Next, the residual is converted to the cepstral domain and the periodic energy is liftered out in the quefrency domain [96]. This gives initial estimates of the aperiodic and periodic components of the excitation source, in which the aperiodic part initially contains gaps at the harmonic frequencies. Therefore, the estimate of the unvoiced (i.e., aperiodic) excitation across the full spectrum (including the harmonic regions) is further refined by using an iterative algorithm, and after convergence, the periodic excitation is obtained by subtracting this estimate from the initial residual signal. To obtain each individual speech component, these excitation source estimates are passed through an all-pole filter whose coefficients come from the first linear prediction analysis step. This method did not assess the quality of the recovered signals, as it only evaluated the relative level between them, i.e., the harmonics-to-noise ratio (HNR).

Three years later in [77], it was suggested a spectral technique which estimates the voiced part by placing a comb filter directly on the speech signal. Based on the property that the harmonics of f_0 fall at certain frequency bins, an analysis window of a length which consists of an integer number of f_0 periods is used, so this requires that f_0 is estimated at each time before the comb filtering is applied. The method is known as the pitch-

scaled harmonic filter (PSHF), as the comb filter will only pass harmonic frequencies of f_0. In a similar way to [174], to reconstruct the unvoiced part spectrum in the harmonic bins, a spectral interpolation is applied prior to extracting the signal in the time domain by assuming that the unvoiced spectral envelope is smooth. A final spectral subtraction is applied to obtain the spectrum of the unvoiced part. It was also shown in [77] that the iterative periodic-aperiodic decomposition algorithm of [174] converged to the original mixed excitation signal, and that using the residual signal did not improve the decomposition robustness. The PSHF method is able to preserve the stochastic part modulation characteristics, and it had desirable properties in terms of the segmental SNR, but still presented problems with shimmer, jitter and transients.

Recently, [37] introduced a method that estimates f_0 by taking into account the colored nature of the unvoiced speech component. The f_0 is estimated from a pre-whitened cumulative periodogram, which was obtained from an estimate of the noise PSD, which was the output of a median filter [139] applied to the noisy periodogram. The candidate f_0 values were post-processed to reduce the number of sub-harmonic errors. That method is able to have a good separation of the speech parts for voiced fricatives, in cases where the level of the stochastic component is comparable or even higher than the deterministic one. The authors, however, hypothesize that a better performance can be achieved by taking into account the estimation of the number of harmonics and by including adaptive windows to consider an enough number of periods, specially for the case of low f_0.

The described methods are not optimal in the sense that they rely on various heuristics. For example, they apply an interpolation to reconstruct the unknown spectra in some of the frequency bins, or they do not address the problem of estimation of the number of harmonics of the voiced speech component. They also ignore the possible presence of additive noise, which should be considered e.g., in telemedicine applications [116]. In paper E, we propose a decomposition method based on optimal filtering which extracts separately the voiced and unvoiced speech parts by using two different Wiener filters. As opposed to the state-of-the-art methods, the proposed one takes into account the presence of additive noise in the mathematical model. This method was motivated by the fact that linear filtering has been widely used for speech enhancement, and in a similar way, we applied optimal filtering based on estimated statistics to the speech decomposition problem. In paper F, we used supervised statistics for both the unvoiced and noise components, and we propose to do the extraction based on adaptive segments which depend on the signal characteristics, and which represent a better fit of the mathematical model within that particular optimal segment length.

Performance measures for speech decomposition

When evaluating how accurate is a particular method in extracting the speech components, it is convenient to have some measures to assess the performance. The following are objective measures which quantify how closer is an estimated signal to the ground truth component.

To compare how close is an estimated signal $\hat{s}(n)$ to the original one $s(n)$ in a MSE sense, the segmental SNR is computed in the following way if the processed and clean signals are alligned in time [104]:

$$\text{segSNR} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Nm}^{Nm+N-1}s^2(n)}{\sum_{n=Nm}^{Nm+N-1}(s(n)-\hat{s}(n))^2}, \tag{23}$$

where M denotes the number of segments and N is the segment length. This measure is not very perceptually relevant, and therefore frequency domain measures may be preferred. The log-Spectrum distortion can be used to compare the spectrum of the original and estimated signal, and is defined as [112]

$$\text{LSD} = \frac{1}{\pi}\int_0^{\pi}(\log_{10}(S(\omega)) - \log_{10}(\hat{S}(\omega)))^2 d\omega. \tag{24}$$

Another frequency domain measure, which is more correlated to the human perception than the LSD is the Itakura-Saito distortion (ISD) [31, 71], expressed as

$$\text{ISD} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\frac{S(\omega)}{\hat{S}(\omega)} - \ln\frac{S(\omega)}{\hat{S}(\omega)} - 1\right] d\omega. \tag{25}$$

Both frequency measures are computed in individual overlapped segments. These measures were computed in papers E and F, when evaluating the different speech decomposition methods, where $s(n)$ represents either voiced or unvoiced speech.

4 Contributions

This section gives an overview of the papers A through F which form the main body of this thesis. We have focused on how the colored noise needs to be addressed in order to have a robust estimation of the speech signals parameters. The focus of papers A-D was on showing how the performance of methods that were derived under a WGN condition can be considerably improved by using an autoregressive pre-whitening filter. The applications of interest were for f_0 estimation and time of arrival estimation. Papers E and F dealt with the problem of estimating individual parts of speech (i.e., voiced and unvoiced) in common background scenarios.

Paper A The first paper in this thesis evaluates the estimation accuracy of the NLS f_0 estimator (i.e., the ML estimator in WGN condition) when its input signal is the pre-whitened signal, obtained from a general FIR or an AR pre-whitening filter based on well-known noise PSD trackers (e.g., MS, IMCRA and MMSE based on SPP). Different noise types at different SNRs were considered. The results were also evaluated separately for female and male speech. The closest performance to the AR oracle pre-whitener was achieved from AR pre-whitening reliant on MMSE based on SPP. In that case, a lower number of gross f_0 errors is observed and a higher SFM is achieved, i.e., the noise gets whiter compared to pre-whitening based on other noise trackers (IMCRA and MS). Despite this, using a higher AR order P to fit the estimated noise PSD to an AR spectrum, did not necessarily increase the performance of the NLS f_0 estimator, and a gap to the oracle performance was observed for non-stationary noise types such as babble noise. Additionally, we observed that a general FIR pre-whitening filter did not improve the performance as much as an AR pre-whitening filter did.

Paper B As seen in paper A, it might be possible to achieve a performance closer to the AR oracle pre-whitener. Therefore, in this paper we investigated if pre-whitening based on supervised noise statistics, which requires the training of some data, was able to increase the noise flatness and improve the estimation accuracy of the NLS f_0 estimator. The noise PSD was obtained from an AR mixture modeling of the speech and noise components of the noisy observation. Maximising the likelihood of the data will result in a supervised NMF of the noisy periodogram into a spectral basis matrix and an activation coefficient matrix where the optimization criteria is the IS divergence, when it is required obtain the activation coefficients of pre-trained spectral shapes. The spectral shapes are described by AR parameters, and the activation coefficients are obtained using a MU rule. Additionally, pre-whitening was combined with OMLSA-based speech enhancement to show an application of the proposed scheme.

Paper C From paper A, it could be wrongly concluded that in some cases (e.g., male speech excerpts) at higher SNRs, it was better to assume that the noise is WGN, instead of pre-processing the signal with a pre-whitening filter. Motivated by this, in this paper we re-estimated iteratively the f_0 and the noise AR parameters. The motivation also came from the fact that in the case of independent sinusoids, a single cascade estimation of sinusoidal and noise AR parameters is enough to have the ML solution, even if the frequencies are estimated before the noise AR parameters. We here investigated if for the harmonically related sinusoids (i.e., harmonic model), the reiteration between estimation of AR parameters and harmonic components, was needed to achieve a similar solution to the ML one. Two iterative schemes were proposed: an approximate ML solution and one based on the LCMV filter to extract a periodic signal, and both had similar

performance. In the iterative process, an updated pre-whitening filter, based on the estimated residual, is applied. Through extensive evaluations, it was seen that the NLS f_0 estimator derived on the simple WGN assumption is still useful in colored noise situations, as long as the reiterations are applied. These reiterations were needed to outperform classical f_0 estimators, such as YIN, RAPT and a cepstrum-based, and the reiterations showed more improvement at non-negative SNRs. The performance metrics also included the voicing detection errors and the full frame errors.

Paper D Here we extended the work based on the previous papers. We also addressed the common question of why the signals should be pre-whitened, instead of enhanced, when it is desired to have a more accurate f_0 estimation. Particularly, the comparison was done to an OMLSA-based speech enhancement pre-processor. The conditions for a higher noise SFM, a better capture of the noise spectral envelope, and therefore a better f_0 estimation accuracy were also evaluated. It was found that using few speech spectral shapes and a a considerable number of noise spectral shapes in the dictionaries, lead to a higher SFM of the pre-whitened noise and a a more accurate capturing of the noise spectral envelope, particularly at non-stationary noise environments, as babble noise and restaurant noise. Such conditions allowed for an improved f_0 estimation accuracy, and a further improvement was obtained from a reiterative estimation between f_0 and the AR parameters. When the input signal of non-parametric f_0 estimators (e.g., SWIPE and RAPT) was pre-processed, an improvement was also observed. However, specially in non-stationary noise conditions and at low SNR levels, the best f_0 estimation accuracy was obtained from our proposed framework, and only when the pre-whitener based on a parametric NMF method is initially used as a pre-processor. Therefore, we verified that using a pre-whitener based on supervised noise statistics, provides more robustness when estimating f_0. The proposed pre-whitener based on parametric NMF also improved the accuracy of a ML TOA estimator in a colored noise scenario.

Paper E This paper tackles the problem of decomposing a noisy speech signal into its voiced and unvoiced components, using a method based on the estimation of the statistics of the different components (voiced, unvoiced and noise) and on optimal linear filtering based on a Wiener filter formulation. The noise statistics were estimated using the minimum statistics noise PSD tracker. The performance of the proposed method was compared to that of state-of-the-art speech decomposition methods which do not distinguish between unvoiced speech and additive noise. Better performance than the iterative periodic-aperiodic decomposition algorithm was observed, and similar one to the PSHF method in terms of segSNR and LSD for voiced speech, but the proposed had lower ISD between the true and the estimated voiced part at low SNRs. Informal listening tests also reveal that the

proposed method allowed to clearly perceive the different unvoiced stop sounds.

Paper F This paper also addresses the problem of decomposing a noisy speech signal, but it uses supervised statistics estimates of the noise and unvoiced speech component. Parameter estimates of the hybrid speech model are obtained for different possible candidate segment lengths, in order to find the markers of the optimal segmentation of the signal. After the markers are found, linear filtering based on the estimated statistics of individual components is applied to extract the voiced and unvoiced parts. The segmentation is obtained from a maximum a posteriori (MAP) criterion for voiced speech and a log-likelihood criterion for unvoiced speech. A Wiener filter which relies on prior spectral information about unvoiced speech and noise AR parameters is used to obtain the unvoiced speech component given the modelled stochastic sequence. The periodic or deterministic speech parts are better modelled from the use of an optimal segmentation in comparison to a fixed segmentation. In noisy conditions, a higher segSNR and a lower distortion is possible from considering optimal segments instead of fixed ones. Better performance in segSNR for both components and LSD for the voiced part is also seen when compared to state-of-the-art methods when their input signal was an enhanced signal.

5 Conclusion and Future Research Directions

The main focus of this thesis has been on how the colored noise can be handled in order that more accurate estimates of the speech signal parameters (e.g., the fundamental frequency) are obtained. Very often, it is believed that trying to reduce the noise levels as much as possible (i.e., applying speech enhancement) is the kind of pre-processing which needs to be applied to combat the undesirable noise presence, but the papers show that a pre-processing scheme which renders the colored noise closer to white should be instead considered to obtain better parameter estimates. In other words, this can be interpreted as saying that the noise still needs to be present, but in such a way that the model assumptions regarding a particular estimation method (in this case, the methods assuming that the noise is WGN) are better fulfilled.

Different issues have been investigated, regarding how applying a pre-whitening filter based on an AR spectral envelope fitted to either the most cited noise PSD tracking algorithms (e.g., MS, MMSE based on SPP) or on pre-trained spectral information (e.g., supervised NMF approach) allows for a better improvement of WGN-based methods under different noise conditions (e.g., stationary or non-stationary). Clearly, taking into account prior spectral information offers a more robust performance to methods

based on a simple WGN assumption, specially in non-stationary situations. Additionally, even if the accuracy from a single cascaded estimation (of noise AR parameters and f_0 estimates) can be significantly improved compared to that of a wrong WGN assumption, a reiteration between the noise AR parameters and the estimated f_0 may be necessary in order to approach the ML solution and obtain better f_0 estimates. In fact, starting from the second iteration of the iterative procedure, the estimated AR noise parameters would also include the contribution of the stochastic parts of speech, as in high SNRs the additive noise level does not dominate too much, and therefore, the estimated AR parameters of the stochastic speech parts are those which contribute more in obtaining refined f_0 estimates. Although not reported in the results, this was experimentally verified for the case when no additive noise was imposed on the clean signals of the Keele database [134]. Very low value of gross error rates and voicing detection errors were observed. Also, a good pitch estimation accuracy in clean speech was also observed in the whole speech material of other speech databases [5, 133], which have an annotated ground truth of the pitch. In those databases, a similar performance to methods as SWIPE and SHRP was seen in the clean speech case. The benefit of iteratively (jointly) estimating the f_0 and AR parameters is that it allows to extract the inherent parts of the speech signal (i.e., voiced and unvoiced), and have a parametric representation of them.

It is important to mention that in the additive noise case, the noise AR parameters could also be correlated across different segments. In a similar way to a recently introduced Bayesian pitch tracker [144], it would be useful to integrate the evolution of the AR model of the noise along with that of the f_0, the number of harmonics L and the voicing state, as an alternative to the cascading of a general purpose pre-whitener to the tracking method which assumes whiteness on the noise. It could also be interesting to evaluate how the proposed framework of f_0 estimation works for the case of voiced fricatives, since the typical spectral envelopes of pre-trained noise codebooks have most of their power in low frequency parts of the spectrum. A possibility for more robustness might be to include codebook entries which represent the stochastic parts of such voiced fricatives.

In the problem of decomposing noisy speech into stochastic and deterministic parts, we have seen that the f_0 should be estimated along with the AR parameters describing the aperiodic or stochastic parts of speech. By doing that, it is seen that a better f_0 estimation accuracy is possible at low SNR conditions. For future work, it is also desirable to further improve the extraction of unvoiced speech. It is then necessary to investigate if by using other filters in the variable span filtering framework (for trading off the distortion and residual noise), or choosing a different number of eigenvectors, could lead in reducing the distortion of this component. It

might also be convenient to include perceptual criteria in the filter designs, as we believe that the unvoiced component may be masked by the noise part at very low SNRs, but this deserves further investigation. Once a better estimate of the unvoiced component can be obtained, it could be interesting to assess how the introduced decomposition methods are useful in applications such as remote diagnosis of illnesses (e.g., Parkinson's disease) which is typically assessed using sustained vowel phonations. As mentioned before, important features for detecting speech pathologies (e.g., HNR) obtained from an accurate decomposition in severe noise conditions, might be relevant in the remote diagnosis. Some types of unvoiced sounds may be better described with ARMA models, and other future direction could also be to investigate how the methods based on codebooks should be modified in order to integrate the information on them.

Although the additive noise is one of the most common types of degradations which affects the speech signals, other types such as clipping and reverberation may be present. It would be, therefore, interesting to investigate on how the pre-whitening methods need to be modified in order to take into account the reverberation phenomenon, as in this case the harmonic components might be completely blurred in some cases.

References

[1] B. Ahmed and P. Holmes, "A voice activity detector using the chi-square test," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–625.

[2] Z. Ali, I. Elamvazuthi, M. Alsulaiman, and G. Muhammad, "Automatic voice pathology detection with running speech by using estimation of auditory spectrum and cepstral coefficients based on the all-pole model," *Journal of voice*, vol. 30, pp. 757.e7–757.e19, 2016.

[3] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "Speaker indexing and speech enhancement in real meetings / conversations," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 93–96.

[4] B. Atal and J. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *ICASSP. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7, 1982, pp. 614–617.

[5] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching." 1993.

[6] L. Barbier and G. Chollet, "Robust speech parameters extraction for word recognition in noise using neural networks," in *ICASSP: International Conference on Acoustics, Speech, and Signal Processing*, 1991, pp. 145–148 vol.1.

[7] E. Belalcazar-Bolanos, J. Orozco-Arroyave, J. Arias-Londono, J. Vargas-Bonilla, and E. Nöth, "Automatic detection of Parkinson's disease using

noise measures of speech," in *Symposium of Signals, Images and Artificial Vision-2013: STSIVA-2013*. IEEE, 2013, pp. 1–5.

[8] J. Benesty, J. Chen, Y. A. Huang, and S. Doclo, "Study of the Wiener filter for noise reduction," in *Speech Enhancement*. Springer, 2005, pp. 9–41.

[9] J. Benesty, M. G. Christensen, and J. R. Jensen, *Signal enhancement with variable span linear filters*. Springer, 2016, vol. 7.

[10] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer, 2007.

[11] G. E. Birch, P. D. Lawrence, J. C. Lind, and R. D. Hare, "Application of prewhitening to AR spectral estimation of EEG," *IEEE Transactions on Biomedical Engineering*, vol. 35, no. 8, pp. 640–645, Aug 1988.

[12] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *IFA Proceedings 17*, 1993, pp. 97–110.

[13] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[14] G. L. Bretthorst, "The near-irrelevance of sampling frequency distributions," in *Maximum Entropy and Bayesian Methods Garching, Germany 1998*. Springer, 1999, pp. 21–46.

[15] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

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

[17] J. Chen, J. Benesty, and Y. A. Huang, "Study of the noise-reduction problem in the Karhunen-Loeve expansion domain," *IEEE Transactions on Speech and Audio Processing*, vol. 17, pp. 787–802, 2009.

[18] A. D. Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[19] K. B. Christensen, M. G. Christensen, J. B. Boldt, and F. Gran, "Experimental study of generalized subspace filters for the cocktail party situation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 420–424.

[20] M. G. Christensen, "On the estimation of low fundamental frequencies," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011, pp. 169–172.

[21] ——, "Accurate estimation of low fundamental frequencies from real-valued measurements," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2042–2056, 2013.

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

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

[24] M. G. Christensen and J. R. Jensen, "Pitch estimation for non-stationary speech," in *48th Asilomar Conference on Signals, Systems and Computers*, 2014, pp. 1400–1404.

[25] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113–116, 2002.

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

[27] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[28] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10 – 49, 2015.

[29] C. d'Alessandro, V. Darsinos, and B. Yegnanarayana, "Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 12–23, Jan 1998.

[30] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, "Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis," *Speech Communication*, vol. 55, no. 2, pp. 278 – 294, 2013.

[31] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. Wiley-IEEE Press, 2000.

[32] J. . Delmas and H. Abeida, "Stochastic Cramer-Rao bound for noncircular signals with application to DOA estimation," *IEEE Transactions on Signal Processing*, vol. 52, no. 11, pp. 3192–3199, 2004.

[33] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 968–981, 2011.

[34] S. Dubnov, "Generalization of spectral flatness measure for non-gaussian linear processes," *IEEE Signal Processing Letters*, vol. 11, no. 8, pp. 698–701, 2004.

[35] Y. C. Eldar and A. M. Chan, "An optimal whitening approach to linear multiuser detection," *IEEE Transactions on Information Theory*, vol. 49, no. 9, pp. 2156–2171, 2003.

[36] Y. C. Eldar and A. V. Oppenheim, "MMSE whitening and subspace whitening," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1846–1851, 2003.

[37] B. Elie and G. Chardon, "Robust tonal and noise separation in presence of colored noise, and application to voiced fricatives," in *22nd International Congress on Acoustics (ICA)*, 2016.

[38] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[39] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[40] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 4, pp. 477–489, 1988.

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

[42] C. Févotte, E. Vincent, and A. Ozerov, "Single-channel audio source separation with NMF: divergences, constraints and algorithms," in *Audio Source Separation*. Springer, 2018, pp. 1–24.

[43] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd, "The voice activity detector for the pan-european digital cellular mobile telephone service," in *International Conference on Acoustics, Speech, and Signal Processing*, 1989, pp. 369–372 vol.1.

[44] E. M. Friel and K. M. Pasala, "Direction finding with compensation for a near field scatterer," in *IEEE Antennas and Propagation Society International Symposium. 1995 Digest*, vol. 1, pp. 106–109.

[45] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[46] T. Fujimoto, T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Speech synthesis using wavenet vocoder based on periodic/aperiodic decomposition," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2018, pp. 644–648.

[47] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based Noise Power Estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[48] ——, "Improved MMSE-based noise PSD tracking using temporal cepstrum smoothing," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 105–108.

[49] S. V. Gerven and F. Xie, "A comparative study of speech detection methods," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[50] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2014, pp. 2494–2498.

[51] Q. Gong, B. Champagne, and P. Kabal, "Noise power spectral density matrix estimation based on modified IMCRA," in *2014 48th Asilomar Conference on Signals, Systems and Computers*, 2014, pp. 1389–1395.

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

[53] D. W. Griffin and J. Lim, "Multiband excitation vocoder," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 8, pp. 1223–1235, aug 1988.

[54] S. L. Hahn, *Hilbert Transforms in Signal Processing*. Artech House, 1996.

[55] C. W. Han, S. J. Kang, and N. S. Kim, "Reverberation and noise robust feature compensation based on IMM," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1598–1611, 2013.

[56] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 795–805, April 1991.

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

[58] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP Journal on Advances in Signal Processing*, pp. 092 953–, 2007.

[59] J. Hardwick, C. D. Yoo, and J. S. Lim, "Speech enhancement using the dual excitation speech model," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, April 1993, pp. 367–370 vol.2.

[60] Q. He, F. Bao, and C. Bao, "Multiplicative update of auto-regressive gains for codebook-based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 457–468, March 2017.

[61] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*, 2013.

[62] R. C. Hendriks, R. Heusdens, and J. Jensen, "Adaptive time segmentation for improved speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2064–2074, 2006.

[63] ——, "An MMSE estimator for speech enhancement under a combined stochastic–deterministic speech model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 406–415, 2007.

[64] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4266–4269.

[65] H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1995, pp. 153–156 vol.1.

[66] A. O. T. Hogg, C. Evers, and P. A. Naylor, "Speaker change detection using fundamental frequency with application to multi-talker segmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5826–5830.

[67] Y. Hono, S. Takaki, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Periodnet: A non-autoregressive waveform generation model with a structure separating periodic and aperiodic components," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6049–6053.

[68] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[69] J. Huang and Y. Zhao, "An energy-constrained signal subspace method for speech enhancement and recognition in white and colored noises," *Speech Communication*, vol. 26, no. 3, pp. 165–181, 1998.

[70] Q. Huang, C. Bao, X. Wang, and Y. Xiang, "Speech enhancement method based on multi-band excitation model," *Applied Acoustics*, vol. 163, p. 107236, 2020.

[71] F. Itakura and I. F, "A statistical method for estimation of speech spectral density and formant frequency," *Trans. Inst. Electron. Commun. Eng. (Japan)*, vol. 53, pp. 36–43, 1970.

[72] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S35–S35, 1975.

[73] Y. Iwai and T. Shimamura, "Speech enhancement with zero replacement signal at low SNR, pages=143-147„" in *2015 International Symposium on Intelligent Signal Processing and Communication Systems*.

[74] J. Li and P. Stoica, "Efficient mixed-spectrum estimation with applications to target feature extraction," *IEEE Transactions on Signal Processing*, vol. 44, no. 2, pp. 281–295, Feb 1996.

[75] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[76] F. Jabloun and B. Champagne, "Signal subspace techniques for speech enhancement," in *Speech Enhancement*. Springer, 2005, pp. 135–159.

[77] P. J. B. Jackson and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 713–726, Oct 2001.

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

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

[80] J. R. Jensen, S. Karimian-Azari, M. Christensen, and J. Benesty, "Harmonic beamformers for speech enhancement and dereverberation in the time domain," *Speech Communication*, vol. 116, pp. 1 – 11, 2020.

[81] J. R. Jensen, U. Saqib, and S. Gannot, "An EM method for multichannel Toa and Doa estimation of acoustic echoes," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2019, pp. 120–124.

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

[83] S. Jo and C. D. Yoo, "Psychoacoustically constrained and distortion minimized speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2099–2110, 2010.

[84] J.-C. Junqua, B. Reaves, and B. Mak, "A study of endpoint detection algorithms in adverse conditions: incidence on a DTW and HMM recognizer," in *Second European conference on speech communication and technology*, 1991.

[85] M. S. Kavalekalam, J. K. Nielsen, M. G. Christensen, and J. B. Boldt, "A study of noise PSD estimators for single channel speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5464–5468.

[86] M. S. Kavalekalam, J. K. Nielsen, L. Shi, M. G. Christensen, and J. Boldt, "Online parametric NMF for speech enhancement," in *2018 26th European Signal Processing Conference*, 2018, pp. 2320–2324.

[87] S. Kay and V. Nagesha, "Extraction of periodic signals in colored noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, 1992, pp. 281–284 vol.5.

[88] S. M. Kay and V. Nagesha, "Spectral analysis based on the canonical autoregressive decomposition," in *1991 International Conference on Acoustics, Speech, and Signal Processing*, pp. 3137–3140.

[89] ——, "Maximum likelihood estimation of signals in autoregressive noise," *IEEE Transactions on Signal Processing*, vol. 42, no. 1, pp. 88–101, Jan 1994.

[90] S. Kay and J. Salisbury, "Improved active sonar detection using autoregressive prewhiteners," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1603–1611, 1990.

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

[92] K. Kim and G. Shevlyakov, "Why Gaussianity?" *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 102–113, 2008.

[93] W. B. Kleijn and J. Haagen, "Transformation and decomposition of the speech signal for coding," *IEEE Signal Processing Letters*, vol. 1, no. 9, pp. 136–138, 1994.

[94] M. Krawczyk-Becker and T. Gerkmann, "Fundamental frequency informed speech enhancement in a flexible statistical framework," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 940–951, 2016.

[95] N. Krishnamurthy and J. H. L. Hansen, "Babble noise: Modeling, analysis, and applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1394–1407, 2009.

[96] G. d. Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, pp. 254–266, 1993.

[97] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *European Signal Processing Conference*, 2014, pp. 61–65.

[98] M. Kuropatwinski and W. B. Kleijn, "Estimation of the excitation variances of speech and noise AR-models for enhanced speech coding," in *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, vol. 1, 2001, pp. 669–672 vol.1.

[99] A. Lapidoth, *A foundation in digital communication*. Cambridge University Press, 2017.

[100] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic+noise model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 1993, pp. 550–553 vol.2.

[101] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[102] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[103] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, January 1980.

[104] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, Inc., 2013.

[105] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 47–56, 2011.

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

[107] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, 2005.

[108] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. European Signal Processing Conference*, 1994, pp. 1182–1185.

[109] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[110] M. McCallum and B. Guillemin, "Stochastic-deterministic MMSE STFT speech enhancement with general a priori information," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 7, pp. 1445–1457, 2013.

[111] A. V. McCree and T. P. Barnwell, "A mixed excitation lpc vocoder model for low bit rate speech coding," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 242–250, 1995.

[112] I. McLoughlin, *Applied Speech and Audio Processing: With Matlab Examples*, 1st ed. New York, NY, USA: Cambridge University Press, 2009.

[113] D. Mehta and T. F. Quatieri, "Synthesis, analysis, and pitch modification of the breathy vowel," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.*, 2005, pp. 199–202.

[114] D. D. D. Mehta, "Aspiration noise during phonation: Synthesis, analysis, and pitch-scale modification," Ph.D. dissertation, Massachusetts Institute of Technology, 2006.

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

[116] R. J. Moran, R. B. Reilly, P. de Chazal, and P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 468–477, March 2006.

[117] P. J. Murphy, "Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2866–2881, 1999.

[118] H. Muta, T. Baer, K. Wagatsuma, T. Muraoka, and H. Fukuda, "A pitch-synchronous analysis of hoarseness in running speech," *The Journal of the Acoustical Society of America*, vol. 84, no. 4, pp. 1292–1301, 1988.

[119] T. Nakatani, S. Amano, T. Irino, K. Ishizuka, and T. Kondo, "A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments," *Speech Communication*, vol. 50, no. 3, pp. 203 – 214, 2008.

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

[121] H. Ney, "A dynamic programming algorithm for nonlinear smoothing," *Signal Processing*, vol. 5, no. 2, pp. 163–173, 1983.

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

[123] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast and statistically efficient fundamental frequency estimation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2016, pp. 86–90.

[124] J. K. Nielsen, M. Kavalekalam, M. Christensen, and J. Boldt, "Model-based noise PSD estimation from speech in non-stationary noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

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

[126] A. M. Noll, "Cepstrum pitch determination," *The journal of the acoustical society of America*, vol. 41, no. 2, pp. 293–309, 1967.

[127] ——, "Pitch determination of human speech by the harmonic product spectrum, the harmonic surn spectrum, and a maximum likelihood estimate," in *Symposium on Computer Processing in Communication, ed.*, vol. 19. University of Broodlyn Press, New York, 1970, pp. 779–797.

[128] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Enhancement and noise statistics estimation for non-stationary voiced speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 645–658, April 2016.

[129] ——, "Instantaneous fundamental frequency estimation with optimal segmentation for nonstationary voiced speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2354–2367, Dec 2016.

[130] N. Obin and M. Liuni, "On the generalization of Shannon entropy for speech recognition," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 97–102.

[131] K. K. Paliwal, J. G. Lyons, and K. K. Wójcicki, "Preference for 20-40 ms window duration in speech analysis," in *2010 4th International Conference on Signal Processing and Communication Systems*, 2010, pp. 1–4.

[132] Y. Pantazis and Y. Stylianou, "Improving the modeling of the noise part in the harmonic plus noise model of speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4609–4612.

[133] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[134] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *EUROSPEECH*, 1995.

[135] P. Prandoni, M. Goodwin, and M. Vetterli, "Optimal time segmentation for signal modeling and compression," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, 1997, pp. 2029–2032 vol.3.

[136] A. Prochazka, N. Kingsbury, P. Payner, and J. Uhlir, *Signal analysis and prediction*. Springer Science & Business Media, 2013.

[137] Y. Qi, R. E. Hillman, and C. Milstein, "The estimation of signal-to-noise ratio in continuous speech for disordered voices," *The Journal of the Acoustical Society of America 105*, vol. 105, pp. 2532–2535, 1999.

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

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

[140] L. Rabiner, "Fundamentals of speech recognition," *PTR Prentice Hall*, 1993.

[141] M. Rajapakse, J. Tan, and J. Rajapakse, "Color channel encoding with NMF for face recognition," in *International Conference on Image Processing, 2004*, vol. 3, pp. 2007–2010 Vol. 3.

[142] H. Sameti, H. Sheikhzadeh, Li Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 445–455, 1998.

[143] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 240–251, 2015.

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

[145] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[146] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 177–180.

[147] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan 1999.

[148] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1. IEEE, pp. 365–368.

[149] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 163–176, Jan 2006.

[150] ——, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 441–452, Feb 2007.

[151] J. Stahl and P. Mowlaee, "A pitch-synchronous simultaneous detection-estimation framework for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 436–450, 2018.

[152] J. Stahl and P. Mowlaee, "Exploiting temporal correlation in pitch-adaptive speech enhancement," *Speech Communication*, vol. 111, pp. 1 – 13, 2019.

[153] P. Stoica and P. Babu, "The Gaussian data assumption leads to the largest Cramér-Rao Bound [lecture notes]," *IEEE Signal Processing Magazine*, vol. 28, no. 3, pp. 132–133, 2011.

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

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

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

[157] P. Stoica and R. L. Moses, "Spectral analysis of signals," *Pearson*, 2005.

[158] Y. Stylianou, "Decomposition of speech signals into a deterministic and a stochastic part," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, vol. 2, 1996, pp. 1213–1216 vol.2.

[159] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, Jan 2001.

[160] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback–Leibler divergence," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 7, pp. 1233–1242, 2015.

[161] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–333–I–336.

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

[163] S. M. Tahir, A. Z. Shaameri, and S. H. S. Salleh, "Time-varying autoregressive modeling approach for speech segmentation," in *Proceedings of the Sixth International Symposium on Signal Processing and its Applications*, vol. 2, 2001, pp. 715–718 vol.2.

[164] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[165] C. W. Therrien, *Discrete random signals and statistical signal processing*. Prentice Hall PTR, 1992.

[166] M. Thomson, S. Boland, M. Wu, J. Epps, and M. Smithers, "Decomposition of speech into voiced and unvoiced components based on a state-space signal model," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2003, pp. I–I.

[167] P. P. Vaidyanathan, *The Theory of Linear Prediction*. Morgan & Claypool, 2007.

[168] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[169] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding And Error Concealment*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2006.

[170] D. Vincent, O. Rosec, and T. Chonavel, "A new method for speech synthesis and transformation based on an ARX-LF source-filter decomposition and HNM modeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, 2007, pp. IV–525–IV–528.

[171] T. Virtanen and T. Barker, "Separation of known sources using non-negative spectrogram factorisation," in *Audio Source Separation*. Springer, 2018, pp. 25–48.

[172] W. Chu and A. Alwan, "Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 3969–3972.

[173] R. Yao, Z. Zeng, and P. Zhu, "A priori SNR estimation and noise estimation for speech enhancement," *EURASIP journal on advances in signal processing*, vol. 2016, no. 1, p. 101, 2016.

[174] B. Yegnanarayana, C. D'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 1–11, 1998.

[175] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Transactions on Speech and Audio processing*, vol. 6, no. 1, pp. 1–11, 1998.

[176] C. D. Yoo and J. S. Lim, "Speech enhancement based on the generalized dual excitation model with adaptive analysis window," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1995, pp. 832–835.

[177] Zenton Goh, Kah-Chye Tan, and B. T. G. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 510–524, 1999.

[178] H. Zhao and X. Zou, "A speech enhancement preprocessor for low bit rate speech coding," in *Pacific-Asia Conference on Circuits, Communications and Systems*, 2009, pp. 443–445.

[179] L. Zão and R. Coelho, "Generation of coloured acoustic noise samples with non-gaussian distributions," *IET Signal Processing*, vol. 6, no. 7, pp. 684–688, 2012.

Part II

Papers

Paper A

A Study on how Pre-whitening Influences Fundamental Frequency Estimation

Alfredo Esquivel Jaramillo, Jesper Kjær Nielsen, Mads Græsbøll Christensen

The paper has been published in the
Proceedings IEEE Int. Conf. Acoust., Speech, Signal Process.
pp. 6495–6499, 2019.

The layout has been revised.

Abstract

This paper deals with the influence of pre-whitening for the task of fundamental frequency estimation in noisy conditions. Parametric fundamental frequency estimators commonly assume that the noise is white and Gaussian and, therefore, they are only statistically efficient under those conditions. The noise is coloured in many practical applications and this will often result in problems of misidentifying an integer divisor or multiple of the true fundamental frequency (i.e., octave errors). The purpose of this paper is to see if pre-whitening can reduce this problem, based on noise statistics obtained from existing noise PSD estimation algorithms. For this purpose, different noise types and prediction orders of LPC pre-whitening are considered. The results show that pre-whitening improves significantly the estimation accuracy of an NLS pitch estimator when the noise is fairly stationary. For nonstationary noise, the improvements are modest at best, but we hypothesize that this is due to the noise PSD estimation performance rather than the LPC pre-whitening principle.

1 Introduction

The lowest rate at which a periodic signal repeats itself is known as the fundamental frequency. Fundamental frequency estimation is of particular interest in speech applications such as speech enhancement [1], diagnosing illnesses [2], speech decomposition [3, 4] and automatic speech recognition [5]. For example, the speech recordings obtained for the purpose of pathological voice analysis may be corrupted by background noise, and this could affect a proper diagnosis [6]. Fundamental frequency estimators can be grouped as non-parametric and parametric. The non-parametric estimators (e.g. YIN [7]), although fast and conceptually simple, have poor time-frequency resolution and poor noise robustness [8]. A signal model which takes into account the noise presence can be used to derive a parametric estimator [9], based on statistical assumptions. Recently, a fast algorithm which considerably reduces the computational complexity of a nonlinear least squares (NLS) estimator has been proposed [8, 10]. This NLS fundamental frequency estimator is only statistically efficient under a white Gaussian noise (WGN) condition. However, in most real acoustic scenarios the noise is coloured such as car noise and street noise. Estimating the fundamental frequency with a WGN assumption sometimes results in misidentifying a multiple or divisor of the true value (i.e., octave errors). Therefore, a pre-whitening scheme should be applied to the noisy signals, which renders the coloured noise closer to WGN.

The pre-whitening of noisy speech can be done either via the Cholesky

factorization [9] or with a FIR filter, for example one based on linear prediction [11]. By applying the Cholesky factor, the signal model needs to be modified as in [12]. Therefore, since the structure of the problem is altered, the fast NLS method cannot be directly applied. A pre-whitening FIR filter which changes the coloured noise into white noise, can preserve the model as only the amplitudes and phases are altered [13]. We focus on this principle in this paper. Therefore, information on the noise spectrum, i.e., noise statistics, is needed. For example, in [11, 14, 15], the noise statistics and the AR parameters of the coloured noise are only estimated during speech-absence periods, assuming that the noise is stationary. Those can be obtained from a voice activity detector (VAD). However, some noise types such as babble and restaurant noise may be non-stationary, so their noise characteristics are time-varying. This issue has been addressed in some noise power spectral density (PSD) estimation algorithms, such as minimum statistics (MS) [16], improved minima controlled recursive averaging (IMCRA) [17], and minimum mean squared error (MMSE) based estimation [18]. This paper intends to extend the work in [13] on pre-whitening. In order to study the effectiveness of these noise PSD estimation algorithms when applying pre-whitening for the purpose of fundamental frequency estimation, the evaluation will be done for both male and female speech, as well as considering different types of real-life noise.

The rest of the paper is structured as follows. Section 2 details the signal model, the fundamental frequency estimator that assumes WGN and details on the pre-whitening schemes. Section 3 explains the experimental setup and the results in terms of spectrograms, gross error rates and spectral flatness measure. Finally, section 4 concludes the work.

2 Signal model and pre-whitening

We present the signal model, the fundamental frequency estimator, and the pre-whitening schemes in this section. For voiced speech segments, the signal $s(n)$ is modelled by L harmonic components whose frequencies are an integer multiple of the fundamental frequency ω_0, having real amplitude $A_l > 0$ and phase $\psi_l \in [0, 2\pi)$. The signal is buried in additive (white or coloured) Gaussian noise $e(n)$, which is uncorrelated with $s(n)$. For $n = 0, 1, ..., N-1$ (where the clean signal is considered being stationary), the signal model is given as

$$x(n) = s(n) + e(n) = \sum_{l=1}^{L} A_l \cos(n\omega_0 l + \psi_l) + e(n). \tag{A.1}$$

By using the Euler's identity, the model can be expressed as

$$x(n) = \sum_{l=1}^{L} \left(a_l z^l(n) + a_l^* z^{-l}(n) \right) + e(n), \tag{A.2}$$

where $a_l = \frac{A_l}{2} e^{j\psi_l}$, $z(n) = e^{j\omega_0 n}$, and * denotes complex conjugation. For a frame of length N, (A.2) can be written in vector form as

$$\mathbf{x} = \mathbf{Z}\mathbf{a} + \mathbf{e}, \tag{A.3}$$

where $\mathbf{x} = [x(n)\ x(n+1)\ ...\ x(n+N-1)]^T$ and $\mathbf{e}$ is defined in the same form, $\mathbf{Z} = [\mathbf{z}(1)\ \mathbf{z}(-1)\ ...\ \mathbf{z}(L)\ \mathbf{z}(-L)]$ with $\mathbf{z}(l) = \left[(z(1))^l\ ...\ (z(N))^l\right]^T$, $\mathbf{a} = \left[a_1\ a_1^*\ ...\ a_L\ a_L^*\right]$ and $(\cdot)^T$ denotes transpose. With the WGN assumption, $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$, σ^2 being the noise variance and $\mathbf{I}_N$ the $N \times N$ identity matrix, the maximum likelihood estimator of ω_0 is found by first replacing the amplitudes in (A.3) by their least-squares estimates, $\hat{\mathbf{a}} = (\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{x}$, and then by minimizing the residual power $\|\mathbf{x} - \mathbf{Z}\hat{\mathbf{a}}\|_2^2$, i.e.,

$$\hat{\omega}_0 = \arg\min_{\omega_0} \|\mathbf{x} - \mathbf{Z}\hat{\mathbf{a}}\|_2^2 = \arg\min_{\omega_0} \|\mathbf{x} - \mathbf{Z}(\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{x}\|_2^2. \tag{A.4}$$

Here $(\cdot)^H$ denotes hermitian-transposition. This nonlinear least squares (NLS) minimization problem can be solved in a fast way by exploiting the matrix structure (for further details, see [8]). However, this is only statistically efficient with the WGN assumption. In real scenarios, the noise is usually coloured, i.e., $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_e)$, where $\mathbf{Q}_e$ is the noise covariance matrix. A matrix $\mathbf{L}$ can be used to transform the observed signal as $\mathbf{L}^H\mathbf{x} = \mathbf{L}^H\mathbf{Z}\mathbf{a} + \mathbf{L}^H\mathbf{e}$ such that $\mathbf{v} = \mathbf{L}^H\mathbf{e}$ now is distributed as $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$, i.e., the noise is now WGN. The required matrix $\mathbf{L}$ must be the Cholesky factor of $\mathbf{Q}_e^{-1}$, i.e., $\mathbf{L}\mathbf{L}^H = \mathbf{Q}_e^{-1}$. However, the harmonic part is also affected and therefore, the structure of the matrices involved in the fast computation of the cost function of (D.33). Another approach to pre-whiten the noisy signal (i.e., that renders coloured noise white) is by applying a filter.

To apply a filter that pre-whitens the noisy signal, the coloured noise can be seen as the output of a filter $H(\omega)$ excited with WGN. When the coloured noise is the output of an all-pole (IIR) filter $H(\omega) = \frac{1}{B(\omega)}$, where $B(\omega) = 1 + \sum_{p=1}^{P} b_p e^{-j\omega p}$, the process is said to be autoregressive (AR). Here, P denotes the prediction order and $b_1, ..., b_P$ are the linear prediction coefficients (LPC). In this sense, the inverse FIR filter $B(\omega)$, can be used to recover the white Gaussian samples given the samples of the AR process and the LPC AR coefficients. Applying this filter (b_n in the time domain) to the noisy signal preserves the signal model for the harmonic model part in (A.2), since

$$b_n * s(n) = b_n * \sum_{l=-L, l\neq 0}^{L} a_l e^{jn\omega_0 l} = \sum_{l=-L, l\neq 0}^{L} \tilde{a}_l e^{jn\omega_0 l}, \tag{A.5}$$

where $\tilde{a}_l = a_l \sum_{p=0}^{P} b_p e^{-j\omega_0 p}$, $b_0 = 1$, so only the complex amplitudes are affected and the fundamental frequency remains unchanged. An estimate of b_p, $p = 1, ...P$ can be obtained from the Levinson-Durbin recursion of order P [19] after the noise statistics are estimated. Given $\mathbf{x}$, some noise tracking algorithms such as MS, IMCRA, and MMSE can be used to estimate the noise PSD, defined as [20]

$$\phi_e(\omega) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[|E(\omega)|^2 \, |\mathbf{x}\right] \tag{A.6}$$

where $E(\omega) = \mathbf{f}^H(\omega)\mathbf{e}$ is the DFT of the noise with $\mathbf{f}(\omega) = \left\{e^{jn\omega}\right\}_{n=0}^{N-1}$, and $\mathbb{E}$ denotes the statistical expectation operator. The inverse DTFT of the noise PSD allows us to recover the noise covariance sequence via [20]

$$r_e(n) = \int_{-\pi}^{\pi} \phi_e(\omega) e^{jn\omega} \frac{d\omega}{2\pi}. \tag{A.7}$$

From this estimated covariance, the LPC parameters can be found from the Levinson-Durbin recursion, which form the b_n pre-whitening FIR filter of order P. We refer to this as the LPC pre-whitener.

Another possibility [13] is to derive a FIR filter directly from the N frequency coefficients of the noise PSD $\phi_e(\omega)$. Since $\phi_e(\omega) = \sigma^2 \, |H(\omega)|^2 = \frac{\sigma^2}{|B(\omega)|^2}$, and assuming a white Gaussian unit variance $\sigma^2 = 1$, the frequency response of the pre-whitening filter is obtained as $B(\omega) = \frac{1}{\sqrt{\phi_e(\omega)}}$, for N frequency points. An FIR filter of order N is found via the inverse DTFT, i.e. $b_n = \int_{-\pi}^{\pi} B(\omega) e^{jn\omega} \frac{d\omega}{2\pi}$, $n = 0, 1, ...N-1$. We refer to this as the FIR pre-whitener.

3 Experimental evaluations

In this section, we evaluate the influence of the LPC and FIR pre-whitening filters on the fundamental frequency estimation performance, and how well they render the coloured noise closer to white.

We start by demonstrating how pre-whitening can lead to better fundamental frequency estimates. For this, we consider the voiced female speech sentence "Why were you away a year, Roy?", sampled at 8 kHz, with added babble noise from the AURORA database [21] at an SNR of 3 dB. The fundamental frequency is estimated using the NLS estimator every 25 ms from the interval [55 Hz, 370 Hz]: first from WGN assumption and then, after applying an LPC-prewhitener where the LPC coefficients are directly obtained from the noise signal using $P = 7$. The results are depicted in Fig. A.1. As observed, the fundamental frequency estimates

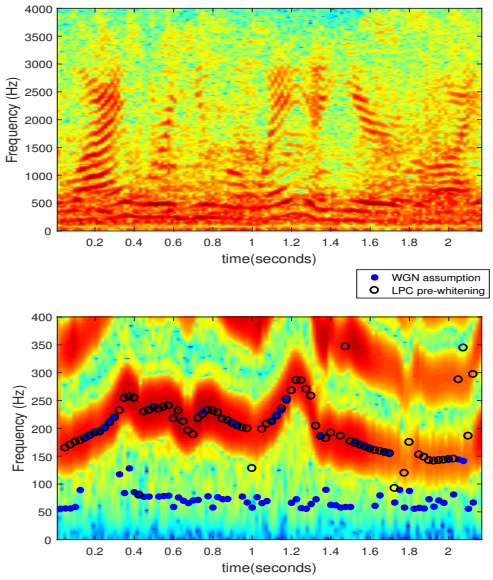

Fig. A.1: Spectrogram of a female speech signal contaminated by babble noise at SNR = 3dB (top), and estimated fundamental frequency estimates imposed on the clean signal spectrogram (bottom).

obtained after pre-whitening result in fewer errors compared to the case with no pre-whitening (WGN assumption).

We now consider the speech signals from the Keele reference database [22], which consists of five male and five female speech recordings, where the fundamental frequency is annotated from laryngograph measurements at a frame rate of 10 ms. The signals are resampled from 20 kHz to 8 kHz. The evaluation was done on the first 80,000 samples (10 s) of each speech file. It is important to notice that the annotated fundamental frequencies do not necessarily correspond to the ground truth, but they also correspond to an estimate which was obtained using an autocorrelation method [23].

For evaluating the fundamental frequency estimation accuracy, only the voiced speech frames with periodicity in both the laryngograph signal and on the speech data were considered (refer to [22] for further description). The assessment was done in terms of the gross error rate (GER), which is defined as the percent of voiced frames whose estimated fundamental frequency deviates more than a certain percentage from the ground truth [24]. We here use 10%. The segment length was set to be $N = 240$ (corresponding to 30 ms), and the fundamental frequency was searched using the NLS estimator in an interval [55 Hz, 370 Hz][1], with a maximum possible of $L = 15$

[1]The lowest fundamental frequency in an evaluated segment of the Keele database is 57

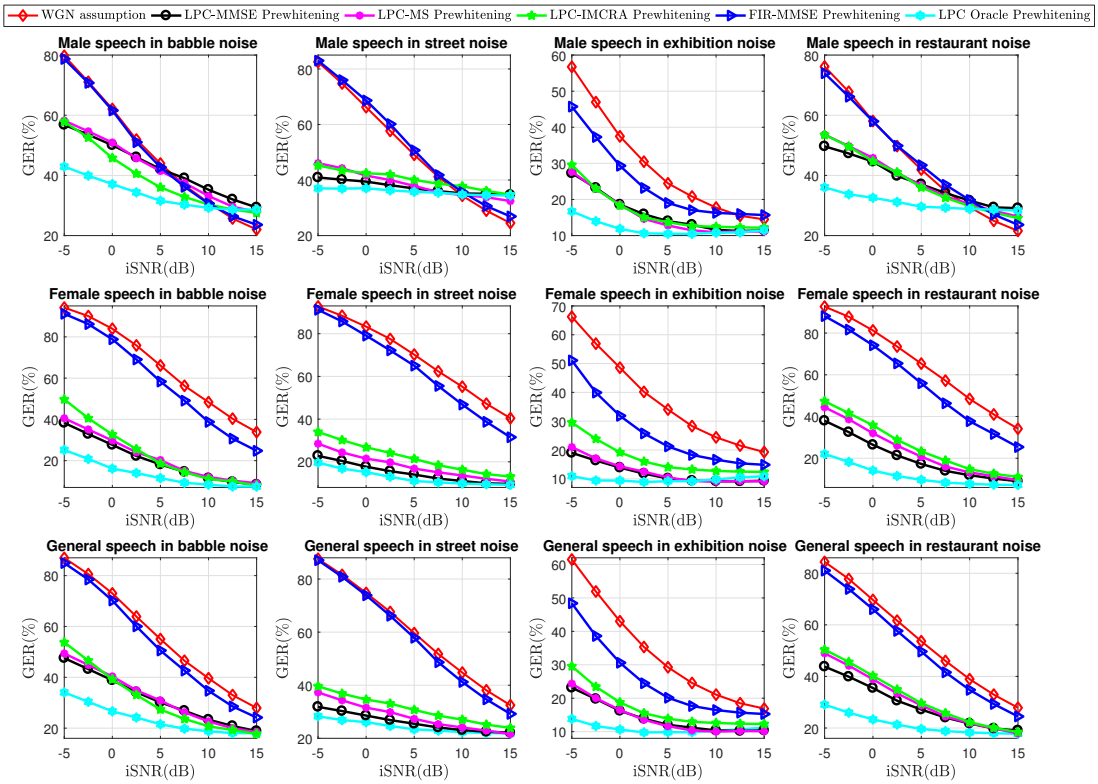

Fig. A.2: Gross error rate (GER) as a function of the iSNR for male, female and general speech on different types of real noise.

harmonics. In order to have the same frame rate as the ground truth, the shift between frames was set to $N = 80$ (i.e., 10 ms). The evaluation was done with four noise types: street, babble, exhibition and restaurant, which are obtained from the AURORA database [21]. The iSNR is varied from -5 to 15 dB. Three different LPC pre-whiteners were used, according to three noise PSD estimates: MMSE [18], MS [16], and IMCRA [17], so the comparison will allow us to determine which one of them helps better for the task of fundamental frequency estimation. For the FIR pre-whitener, only the MMSE noise PSD estimate is presented, since similar results were observed with respect to the other noise PSD estimators. In order to get an insight in to what is the best performance that can be achieved, the results also include the case where an LPC oracle pre-whitener is used, i.e., where the LPC parameters were computed directly from the noise signal. The order of the LPC pre-whiteners was set to $P = 7$, as this seemed to work well (see also the explanation for the next experiment). The results are

Hz.

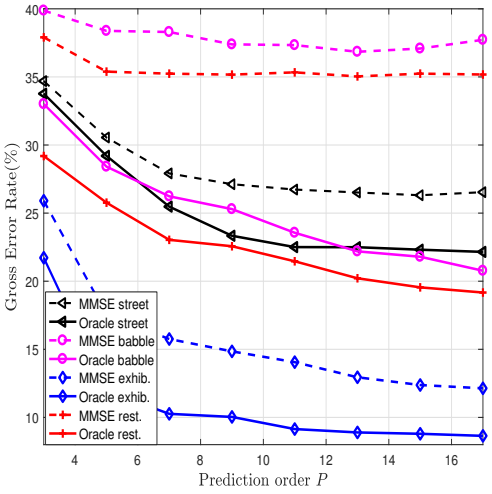

Fig. A.3: Gross error rate (GER) as a function of the prediction order P at iSNR = 0 dB, for general speech.

displayed in Fig. A.2, the results are shown separately for male and female speech, and also for general speech. In general, the GER from the LPC oracle pre-whitener is lower for female than for male speech, since most of the power of the coloured noise is in the lower frequencies which coincide with the range of fundamental frequencies of male speech.

The performance from the LPC pre-whitener based on MMSE noise PSD estimation is mostly the closest to the LPC oracle pre-whitener, followed by the one based on MS. For the case of male speech above an iSNR of 10 dB, it seems that it is better to assume WGN or to do FIR pre-whitening to estimate the fundamental frequency (except in the exhibition noise case). Otherwise, in most cases, the benefit of LPC pre-whitening is clear, as the GER resulting from WGN assumption and from FIR pre-whitening is higher. The performance of LPC pre-whitening from noise PSD MMSE estimates is very close to the oracle for the street noise case, while for the other noise types (babble, exhibition and restaurant) there is still room for improvement for attaining lower GERs (closer to the oracle performance).

In the next experiment, we investigate the influence of the prediction order P for LPC pre-whitening. We used the same setup from previous experiment. Since from it, lower GERs were seen from the MMSE noise PSD estimate, and due to the lack of space, we only show the curves corresponding to the pre-whitener from the MMSE noise PSD tracker and compare them to those obtained from LPC oracle pre-whitening. The results are shown in Fig. A.3. for an iSNR = 0 dB for the general speech case. The GERs corresponding to the WGN assumption and the FIR pre-whitening

can be seen for comparison purposes from Fig. F.2 at 0 dB. From the oracle pre-whitening curves, the best possible performance was obtained for the exhibition noise, followed by restaurant and with street and babble noise having the highest GER depending on which P is used. However, by increasing P the GER slightly reduced or kept nearly constant. By applying LPC pre-whitening based on the MMSE noise PSD estimate, the GER also slightly decreased or remained nearly constant as P increased. The lowest GER is also seen for the exhibition noise, but the next lower GER is for street and not for restaurant noise, as opposed to the oracle pre-whitener case. The differences between the GER from oracle and estimated LPC pre-whitener are larger for restaurant (between 8.5 and 16 %, increasing with P) and babble noise (between 6.5 and 17 %, increasing with P) than for street (between 1 and 4.5 %) and exhibition (between 3.5 and 5.5 %) noise types. We speculate that this is due to that street and exhibition are more stationary than restaurant and babble noise types, whose statistics may be more difficult to estimate. Larger differences occuring when P is high, for the babble and restaurant noise types, implies that even if a better noise PSD spectrum could be captured (since a lower GER could be achieved), the conventional noise PSD estimators do not react quickly to nonstationary noise conditions and, therefore, the estimated noise PSD spectrum does not correctly fit the true one. This suggests a future improvement of pre-whitening, for example based on codebook based approach [25, 26], which can better encompass the noise characteristics. Based on this, we did not select a very high value of P for the previous experiment.

A measure of the correlation structure of the noise, and therefore its color degree, is given by the spectral flatness measure (SFM). Therefore, the pre-whitening schemes can be compared in terms of this SFM, which is defined as

$$\text{SFM} = \frac{\exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln \phi(\omega) d\omega\right)}{\frac{1}{2\pi}\int_{-\pi}^{\pi} \phi(\omega) d\omega} \tag{A.8}$$

which is interpreted as the ratio between the geometric mean and the arithmetic mean of the power spectrum $\phi(\omega)$ [19]. The larger this value, the flatter the noise becomes. This quantity is bounded between 0 and 1, where SFM $\to 0$ means that the noise is more coloured and SFM $\to 1$ implies white noise.

The mean SFM was calculated at an iSNR = 0 dB for the different noise types, for two prediction orders $P = 7$ and $P = 14$. The SFM values after pre-whitening are similar to other iSNRs, as was also evaluated in [13], so only the results at 0 dB are shown in Table A.1. The SFM for each noise type before pre-whitening is shown in brackets. The table reports the SFM of the noise after pre-whitening the noisy signal with the FIR method using MMSE noise PSD estimate, and also with the LPC pre-whitening

Table A.1: Comparison of SFM at iSNR = 0 dB for general speech.

		SFM (Spectral Flatness Measure)				
		FIR	LPC1	LPC2	LPC3	LPCO
Street (0.04)	$P = 7$	0.13	0.45	0.44	0.34	**0.50**
	$P = 14$	0.13	0.46	0.45	0.35	**0.53**
Babble (0.07)	$P = 7$	0.17	0.40	0.39	0.37	**0.47**
	$P = 14$	0.17	0.41	0.39	0.36	**0.51**
Exhib. (0.29)	$P = 7$	0.43	0.45	0.45	0.43	**0.48**
	$P = 14$	0.43	0.48	0.47	0.43	**0.53**
Rest. (0.08)	$P = 7$	0.20	0.42	0.40	0.38	**0.49**
	$P = 14$	0.20	0.43	0.40	0.35	**0.52**

with the noise trackers MMSE, MS and IMCRA (LPC1, LPC2 and LPC3, respectively). The last column, LPCO, corresponds to the SFM obtained by using the LPC oracle pre-whitener, i.e., the highest possible SFM with a specific P. For MMSE and MS LPC pre-whiteners, the SFM increases as P increases, something that not always happens by using IMCRA. The closest SFM to the oracle SFM can be obtained from the LPC MMSE pre-whitener. The difference between them is larger for $P = 14$ than for $P = 7$. The SFM obtained from FIR pre-whitening is much lower compared to LPC pre-whitening in most cases, except for exhibition noise, in which the value is very near to the one attained from the LPC pre-whitening. Larger differences between the SFM from oracle and noise trackers are seen for more nonstationary noise types, i.e., restaurant and babble.

4 Conclusions

In this paper, we evaluated the influence of pre-whitening filters based on noise PSD estimation methods for fundamental frequency estimation. We also evaluated how well the LPC and FIR pre-whiteners can distribute the noise power across the entire frequency range in terms of the SFM measure. The LPC pre-whitening based on MMSE results in lower GER of the fundamental frequency estimates and highest SFM compared to the LPC pre-whitening based on the other noise PSD estimates. Moreover, a better improvement is still possible to be achieved, specially in the case of nonstationary noise types.

References

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

[2] R. J. Moran, R. B. Reilly, P. de Chazal, and P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 468–477, March 2006.

[3] P. J. B. Jackson and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 713–726, Oct 2001.

[4] A. Esquivel, J. Nielsen, and M. Christensen, "On optimal filtering for speech decomposition," in *26th European Signal Processing Conference (EUSIPCO)*, May 2018.

[5] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2014, pp. 2494–2498.

[6] A. H. Poorjam, M. A. Little, J. R. Jensen, and M. G. Christensen, "A supervised approach to global signal-to-noise ratio estimation for whispered and pathological voices," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2018, pp. 296–300.

[7] A. D. Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

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

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

[10] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast and statistically efficient fundamental frequency estimation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2016, pp. 86–90.

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

[12] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP Journal on Advances in Signal Processing*, pp. 092 953–, 2007.

[13] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Instantaneous fundamental frequency estimation with optimal segmentation for non-stationary voiced speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2354–2367, Dec 2016.

[14] J. Huang and Y. Zhao, "An energy-constrained signal subspace method for speech enhancement and recognition in white and colored noises," vol. 26, pp. 165–181, 1998.

[15] ——, "An energy-constrained signal subspace method for speech enhancement and recognition in colored noise," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, May 1998, pp. 377–380 vol.1.

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

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

[18] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based Noise Power Estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[19] P. P. Vaidyanathan, *The Theory of Linear Prediction*. Morgan & Claypool, 2007.

[20] P. Stoica, *Introduction to spectral analysis*. Prentice Hall, 1997.

[21] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[22] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *EUROSPEECH*, 1995.

[23] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[24] F. Flego and M. Omologo, "Robust f0 estimation based on a multi-microphone periodicity function for distant-talking speech," in *14th European Signal Processing Conference*, Sept 2006, pp. 1–4.

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

[26] J. K. Nielsen, M. Kavalekalam, M. Christensen, and J. Boldt, "Model-based noise PSD estimation from speech in non-stationary noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

Paper B

Adaptive Pre-whitening Based on Parametric NMF

Alfredo Esquivel Jaramillo, Jesper Kjær Nielsen, Mads Græsbøll Christensen

The paper has been published in the
Proceedings European Signal Processing Conf.. , 2019

The layout has been revised.

Abstract

Several speech processing methods assume that a clean signal is observed in white Gaussian noise (WGN). An argument against those methods is that the WGN assumption is not valid in many real acoustic scenarios. To take into account the coloured nature of the noise, a pre-whitening filter which renders the background noise closer to white can be applied. This paper introduces an adaptive pre-whitener based on a supervised non-negative matrix factorization (NMF), in which a pre-trained dictionary includes parametrized spectral information about the noise and speech sources in the form of autoregressive (AR) coefficients. Results show that the noise can get closer to white, in comparison to pre-whiteners based on conventional noise power spectral density (PSD) estimates such as minimum statistics and MMSE. A better pitch estimation accuracy can be achieved as well. Speech enhancement based on the WGN assumption shows a similar performance to the conventional enhancement which makes use of the background noise PSD estimate, which reveals that the proposed pre-whitener can preserve the signal of interest.

1 Introduction

The presence of additive noise is inevitable in many acoustic scenarios. Although the noise characteristics can be explicitly taken into account for estimating the parameters of a signal of interest (as in [1, 2]), many methods rely on a white Gaussian noise (WGN) condition (see, e.g. [3–5]), since this is convenient from a mathematical point of view. This WGN assumption can be quite unrealistic, as real noise types are typically coloured. Applying methods based on the WGN assumption in real noise scenarios can degrade their performance. One example is when sub-harmonic errors appear when estimating the fundamental frequency (a.k.a. pitch) of voiced speech segments [6, 7] from estimators which assume WGN. A pre-processor which renders the coloured noise closer to white, namely a pre-whitener, can alleviate this problem. Applying pre-whitening using a linear filter is advantageous compared to a general linear transformation with, e.g., the Cholesky factor [4], since the effect of linear filtering can be modeled by only changing the sinusoidal amplitudes and phases [6, 7]. Unlike general linear transformations, linear filtering thus enables us to use many existing model-based estimators based on a WGN assumption. A linear FIR filter with response

$$A(\omega) = 1 + \sum_{i=1}^{P} a_z(i) e^{-j\omega i} \tag{B.1}$$

can be used to whiten the noise if the coloured noise is modeled as an autoregressive process AR(P) resultant by passing white Gaussian excitation noise with variance σ_e^2 through an IIR filter with response $H(\omega) = 1/A(\omega)$. Here, P denotes the linear prediction order, and $\{a_z(i)\}_{i=1}^{P}$ are known as the prediction coefficients. The filter in (E.1) is referred as the LPC pre-whitening filter, and it corresponds to a FIR filter with coefficients $\{1, a_z(1), ..., a_z(P)\}$, which in practice are found from the estimated second-order noise statistics, namely the noise PSD (power spectral density). The influence of filtering-based pre-whitening schemes based on well-known noise PSD estimates, such as minimum statistics (MS) [8], improved minima controlled recursive average (IMCRA) [9], and minimum mean squared error (MMSE) [10]), on the pitch estimation performance, was studied in [6]. Although these schemes will help, for example, in reducing the subharmonic errors of the pitch estimates, it was found that the performance is far from that of the oracle pre-whitener. Consequently, we believe that performance improvements are possible if a more accurate noise PSD is estimated.

Including prior spectral information about typical speech and noise spectral shapes has been shown to be beneficial for the noise PSD estimation accuracy [11], specially under non-stationary noise conditions. In a similar way, we here investigate if an adaptive pre-whitener (i.e., an FIR filter whose parameters change every time frame) based on offline trained speech and noise spectral envelopes can render the noise closer to white, and thereby improve the estimation accuracy of a maximum likelihood (ML) pitch estimator [12, 13]. Specifically, a sum of AR processes model [14] is considered, which was motivated by the source/filter speech production model. In this model, the likelihood maximization corresponds to a parametric non-negative matrix factorization (NMF) [15] of the observed periodogram matrix into a dictionary matrix of pre-trained spectral envelopes, parametrized by AR coefficients, and a matrix of activation coefficients, with the Itakura-Saito (IS) divergence as the optimization criterion.

The rest of the paper is organized as follows. In Section II, the problem is formulated. In Section III, we detail how to estimate the noise PSD using a parametric NMF approach, and we give a summary of the pre-whitening process. Next, in section IV, we compare the noise flatness from the new pre-whitener to others based on conventional noise PSD estimators, and we also evaluate its influence on pitch estimation and on speech enhancement. Finally, section V concludes the presented work.

2 Problem formulation

In this work, we assume that coloured noise $z(n)$ is added to a clean speech signal of interest $s(n)$, i.e.,

$$x(n) = s(n) + z(n), \tag{B.2}$$

where $x(n)$ is the observed noisy signal. For the purpose of pre-whitening $z(n)$ with an LPC pre-whitener, i.e., rendering coloured noise white, the prediction coefficients $\{a_z(i)\}_{i=1}^{P}$ in (E.1) have to be estimated. Given the noise PSD $\phi_z(k)$, $k = 1, ..., K$, the noise autocovariance sequence is obtained from the Wiener-Khintchine theorem as [13]

$$r_z(n) = \frac{1}{K} \sum_{k=0}^{K-1} \phi_z(k) \exp\left(j \frac{2\pi}{K} nk\right), 0 \leq n \leq P, \tag{B.3}$$

where k denotes the frequency bin and K is the number of frequency bins. Then, the Levinson-Durbin recursion [16] is used to compute the WGN excitation variance σ_e^2 and the P noise prediction coefficients $\{a_z(i)\}_{i=1}^{P}$, which forms the LPC pre-whitening filter in (E.1).

In practice, the noise PSD $\phi_z(k)$ is estimated for every frame from the noisy signal periodogram $\phi(k)$. This can be done for example, with one of the well-known noise tracking methods, such as MS [8] or MMSE based on speech presence probabilities [10]. However, as was seen in [6], LPC pre-whitening performance based on these noise PSD estimates is still far from the oracle one in, e.g., non-stationary noise. An MMSE-based noise PSD estimate can be obtained as [10]

$$\phi_z(k) = \left(\frac{1}{1+\xi(k)}\right) \phi(k) + \left(\frac{\xi(k)}{1+\xi(k)}\right) \lambda_z^2(k), \tag{B.4}$$

where $\xi(k) = \lambda_S^2(k)/\lambda_Z^2(k)$ is known as the *a priori* SNR, with $\lambda_S^2(k)$ and $\lambda_Z^2(k)$ being the PSDs of $s(n)$ and $z(n)$ respectively, at frequency bin k. For the proposed pre-whitener, we still use (B.4). However, we obtain an estimate of $\xi(k)$ from a parametric NMF derived from the sum of AR processes model introduced in [14], and explained in the next section. Because of the Kolmogorov-Szego theorem [16], even if the sum of two or more AR processes is not theoretically AR, in order to apply an LPC pre-whitener, an AR approximation of the PSD is possible if a large prediction order P is used. [1]

[1] The value of P will be limited by the available data [16] (usually $P < K/3$), where a low P could result on a very smooth spectrum, while a P too large could result on spurious peaks.

3 Noise PSD estimate based on Parametric NMF

In [14], the sum of AR processes model was introduced in an NMF context. There, a noisy signal frame $\mathbf{x} = [x(0), ..., x(K-1)]^T$ is represented as a sum of $U = U_s + U_z$ AR processes $\mathbf{t}_u$, i.e.,

$$\mathbf{x} = \sum_{u=1}^{U} \mathbf{t}_u = \sum_{u=1}^{U_s} \mathbf{t}_u + \sum_{u=U_s+1}^{U} \mathbf{t}_u, \tag{B.5}$$

where U_s is the number of AR processes corresponding to the speech signal, U_z is the number of AR noise processes, $(\cdot)^T$ denotes transpose, and K is the segment length in samples, corresponding also to the number of frequency bins. Each one of these AR processes is expressed as a multivariate Gaussian $\mathbf{t}_u \sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \mathbf{Q}_u)$. Here, $\mathbf{Q}_u$ is the gain normalized covariance matrix, which can asymptotically be approximated as $\mathbf{Q}_u = K^{-1}\mathbf{F}\mathbf{D}_u\mathbf{F}^H$ [17], where $\mathbf{F} = \{\exp(j2\pi nk/K)\}, n, k = 0, 1, ..., K-1$ and

$$\mathbf{D}_u = \left(\mathbf{\Lambda}_u^H \mathbf{\Lambda}_u\right)^{-1}, \quad \mathbf{\Lambda}_u = \operatorname{diag}\left(\mathbf{F}^H \left[\mathbf{a}_u^T \; \mathbf{0}\right]^T\right), \tag{B.6}$$

where $\mathbf{a}_u$ is the AR coefficients vector of the u^{th} spectral basis. The different pre-trained basis, i.e., spectral envelopes, are contained in a dictionary matrix $\mathbf{D} \in \mathbb{R}_{\geq 0}^{K \times U}$. In order to maximize the likelihood as a function of U excitation variances and U AR spectral envelopes, the $U \times 1$ vector of activation coefficients $\boldsymbol{\sigma} = \left[\sigma_1^2 \; ... \; \sigma_U^2\right]^T$ is estimated online as

$$\hat{\boldsymbol{\sigma}} = \underset{\boldsymbol{\sigma} \geq 0}{\arg\max}\; p(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{D}) = \underset{\boldsymbol{\sigma} \geq 0}{\arg\max}\; \mathcal{N}\left(0, \sum_{u=1}^{U} \sigma_u^2 \mathbf{Q}_u\right). \tag{B.7}$$

This vector corresponds to the excitation variances of each one of the trained *a priori* AR processes. The log-likelihood can be computed and simplified as (see [14] for further details)

$$\ln p(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{D}) = -\frac{K}{2}\ln 2\pi - \frac{1}{2}\sum_{k=0}^{K-1}\left(\frac{\phi(k)}{\sum_{u=1}^{U}\hat{\phi}_u(k)} + \ln \sum_{u=1}^{U}\hat{\phi}_u(k)\right) \tag{B.8}$$

.The summation over U spectral basis in (B.8) is the parametrized representation of the PSD per frequency bin k, and is expressed as $\sum_{u=1}^{U}\hat{\phi}_u(k) = \mathbf{d}_k^T\boldsymbol{\sigma}$, where $\mathbf{d}_k = [d_1(k) \; ... \; d_U(k)]^T$ is the k^{th} row of $\mathbf{D}$. Therefore, the likelihood maximization is equivalent to the minimization of the IS divergence between the observed periodogram $\boldsymbol{\phi} = [\phi(1) \; ... \; \phi(K)]^T$ and the parametrized PSD $\mathbf{D}\boldsymbol{\sigma}$ where $\mathbf{D} = [\mathbf{d}_1 \; ... \; \mathbf{d}_K]^T$, under the constraint $\phi(k) > 0 \; \forall k$, i.e.,

$$\hat{\boldsymbol{\sigma}} = \underset{\boldsymbol{\sigma} \geq 0}{\arg\min}\; d_{IS}\left(\boldsymbol{\phi}|\mathbf{D}\boldsymbol{\sigma}\right). \tag{B.9}$$

Each one of this set of activation coefficients can be iteratively estimated by means of a multiplicative update (MU) rule

$$\hat{\sigma} \leftarrow \hat{\sigma} \odot \left\{ \mathbf{D}^T \left(\mathbf{D}\hat{\sigma} \right)^{[-2]} \odot \phi \right\} \oslash \left\{ \mathbf{D}^T \left(\mathbf{D}\hat{\sigma} \right)^{[-1]} \right\}, \tag{B.10}$$

where $\odot$ and $\oslash$ are element-wise product and division, respectively. The exponentiation is also an element-wise operation.

The observed periodogram matrix $\mathbf{\Phi} \in \mathbb{R}_{\geq 0}^{K \times R}$ can be expressed as $\mathbf{\Phi} \approx \mathbf{D}\mathbf{\Sigma}$, where R is the number of frames and $\mathbf{\Sigma} \in \mathbb{R}_{\geq 0}^{U \times R}$ is the activation matrix which contains in each one of its columns the activation coefficients for a single frame. Therefore, this corresponds to a supervised NMF where $\mathbf{D}$ contains the gain-normalized (i.e., unitary variance) parametrized AR spectral envelopes [13] in each one of its columns as $\widetilde{\mathbf{d}}_u = \left[\widetilde{d}_u(0) \ldots \widetilde{d}_u(k) \ldots \widetilde{d}_u(K-1) \right]^T$, where each frequency-bin element is given by

$$\widetilde{d}_u(k) = \frac{1}{\left| 1 + \sum_{i=1}^{P'} a_u(i) \exp\left(-\frac{2\pi j i k}{K} \right) \right|^2}, \tag{B.11}$$

where $\{a_u(i)\}_{i=1}^{P'}$ are the P' AR coefficients of the u^{th} spectral basis. The first U_s columns of $\mathbf{D}$ correspond to AR speech spectral envelopes and the last U_z ones to AR noise spectral envelopes, i.e., $\mathbf{D} = [\mathbf{D}_s \; \mathbf{D}_z]$.

Finally, after estimating $\mathbf{\Sigma}$, in order to estimate the noise PSD $\phi_z(k)$ as in (B.4), estimates $\hat{\lambda}_S$ and $\hat{\lambda}_Z$ can be obtained as $\hat{\lambda}_S^2(k) = [\mathbf{D}_s\mathbf{\Sigma}_s]_{(k+1),i}$ and $\hat{\lambda}_Z^2(k) = [\mathbf{D}_z\mathbf{\Sigma}_z]_{(k+1),i}$, where $\mathbf{\Sigma}_s$ corresponds to the first U_s rows of $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_z$ to the last U_z ones. Then, an estimate $\hat{\xi}(k) = \hat{\lambda}_S^2(k)/\hat{\lambda}_Z^2(k)$ is found.

For a more robust adaptive pre-whitener, which takes into account noise types or samples which may not be well represented in the pre-trained spectral basis, we also append as a last column in $\mathbf{D}$ a spectral envelope corresponding to the MMSE noise PSD based pre-whitener $\{a_{mmse}(i)\}_{i=1}^{P'}$ [10]

$$\widetilde{d}_{mmse}(k) = \frac{1}{\left| 1 + \sum_{i=1}^{P} a_{mmse}(i) \exp\left(-\frac{2\pi j i k}{K} \right) \right|^2}. \tag{B.12}$$

A summary of the pre-whitening process is given in Table I.

4 Experimental evaluation

In this section, we quantify how well the described pre-whitener works in terms of the spectral flatness measure (SFM), how it improves pitch estimation performance and how well it works for speech enhancement. For these purposes, a general speech codebook was trained from approximately 54 minutes of sentences from 4 different speakers of the CMU Arctic

Table B.1: Summary of the proposed pre-whitening scheme.

1. Train speech and noise codebooks on LSF coefficients, convert them to $\{a_u(i)\}_{i=1}^{P'}$ coefficients and build $\mathbf{D} = [\mathbf{D}_S \ \mathbf{D}_Z]$ whose columns are given by (D.13).
2. For every frame, estimate $\phi(k) = |X(k)|^2 / N, \ k = 1, ..., K$.
3. Add spectral envelope from MMSE PSD estimator to $\mathbf{D}$.
   (a) Estimate the MMSE noise PSD estimate from [10].
   (b) Estimate $r_{mmse}(n)$ from (B.3)
   (c) Estimate $\{a_{mmse}(i)\}_{i=1}^{P'}$ from Levinson-Durbin recursion and form spectral envelope as (B.12), for each frame.
4. Find $\hat{\sigma}_{est}$ per frame, and therefore $\mathbf{\Sigma}$.
   (a) Initialize $\hat{\sigma}_{est}$ with random positive numbers.
   (b) Compute $\hat{\sigma}_{est}$ with the MU rule in (D.22) for 40 iterations.
5. Compute $\hat{\lambda}_S^2(k) = [\mathbf{D}_s\mathbf{\Sigma}_s]_{(k+1),i}$, $\hat{\lambda}_Z^2(k) = [\mathbf{D}_z\mathbf{\Sigma}_z]_{(k+1),i}$.
6. Compute $\hat{\xi}(k) = \hat{\lambda}_S^2(k)/\hat{\lambda}_Z^2(k)$.
7. Compute pre-whitening filter based on estimated noise PSD.
   (a) Estimate noise PSD $\phi_z(k)$ per frame as in (B.4).
   (b) Estimate noise covariance from (B.3).
   (c) Estimate noise prediction coefficients from Levinson Durbin recursion which form filter in (E.1).

database [18], resampled from 16 to 8 kHz. The offline training of the codebooks was done using a standard vector quantization technique from speech coding [19] on the line spectral frequency (LSF) coefficients. The parameters for both the training and for the NMF based pre-whitening (LPC Par-NMF) are summarized in Table II. The noise codebook was trained on noise samples from the Aurora database [20] of restaurant, street, car and airport noise types. Excerpts from the Keele database [21], resampled to 8 kHz, with added babble or exhibition noise from the Aurora database, were used for the evaluation. It is important to note that these noise types were not included in the training stage, and also that the testing speech involves other speakers (i.e., of another database) different from those of the training stage. LPC pre-whiteners based on other noise PSD estimates (e.g., MS, MMSE, IMCRA), as well as the oracle (AR parameters directly computed

from the noise signal), with the same frame duration and overlap as in Table II, were also applied to compare their performance to our proposed pre-whitener.

Table B.2: NMF Pre-whitener parameters

Parameters	Value
sampling frequency(Hz)	8000
frame duration	32 ms
frame overlap	50%
speech order P'	14

Parameters	Value
noise order P'	14
U_s	32
U_z	14
MU iterations	40

4.1 Spectral flatness measure (SFM)

To demonstrate how well the described pre-whitener renders noise closer to white, the whiteness of the noise is quantified in terms of the SFM, defined as [13, 16] the ratio of the geometric mean to the arithmetic mean of the pre-whitened noise PSD ϕ_{zw}, i.e., SFM $= \frac{\left(\sqrt[K]{\prod_{k=0}^{K-1}\phi_{zw}(k)}\right)}{\left(\frac{1}{K}\sum_{k=0}^{K-1}\phi_{zw}(k)\right)}$. The SFM is bounded between 0 (more coloured noise) and 1 (perfect white noise). Babble and exhibition noise types were added at SNRs from -10 to 10 dB. Before pre-whitening, the mean SFM of babble noise was 0.07 and for exhibition noise it was 0.30 at all SNRs. Results of the pre-whitened noise SFM are shown in Fig. F.1 for two LPC pre-whitening orders ($P = 20$ and $P = 30$). It is observed that the highest SFM (closest to the oracle pre-whitener) can be achieved with the NMF based pre-whitening scheme for babble noise at all SNRs, while for exhibition this happens for SNRs below 5dB, since at greater SNRs a similar SFM to pre-whiteners based on MS and MMSE is observed. It is also noted that using a higher LPC pre-whitening order implies a higher SFM, i.e., the noise gets closer to white.

4.2 Pitch estimation

We now consider the task of estimating the pitch ω_0 of a periodic signal buried in additive coloured noise. Voiced speech segments can be modeled as a periodic signal $s(n)$ consisting of L harmonics whose frequencies are an integer multiple of ω_0, having a real amplitude C_l and phase $\psi_l \in [0, 2\pi)$. When such signal segments are contaminated by uncorrelated additive coloured gaussian noise $z(n)$, the signal model becomes $x(n) = \sum_{l=1}^{L} C_l \cos(n\omega_0 l + \psi_l) + z(n)$. In particular for speech, this model is valid for short time segments ($\sim$20-30 ms) where the speech is considered as stationary.

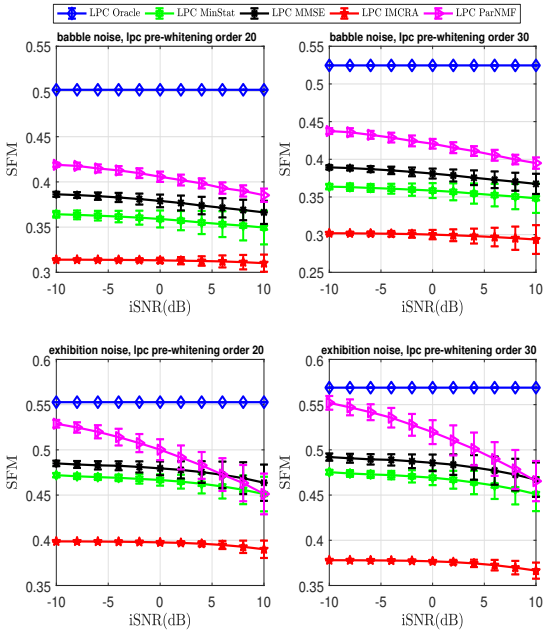

Fig. B.1: mean and 95% confidence interval of the SFM as a function of SNR.

When K noisy samples are stacked in a vector as $\mathbf{x} = [x(0) \ \ldots \ x(K-1)]^T$, the signal model becomes $\mathbf{x} = \mathbf{s} + \mathbf{z} = \mathbf{B}\mathbf{c} + \mathbf{z}$, where $\mathbf{c} = \frac{1}{2}[C_1 e^{j\psi_1} \ C_1 e^{-j\psi_1} \ \ldots \ C_L e^{j\psi_L} \ C_L e^{-j\psi_L}]$, $\mathbf{B} = [\mathbf{b}(\omega_0) \ \mathbf{b}^*(\omega_0) \ \ldots \ \mathbf{b}(\omega_0 L) \ \mathbf{b}^*(\omega_0 L)]$ and $\mathbf{b}(\omega_0 l) = [1 \ e^{-j\omega_0 l} \ \ldots \ e^{-j\omega_0 l(K-1)}]^T$. If $\mathbf{z} = [z(0) \ z(1) \ \ldots \ z(K-1)]^T$ is WGN, the ML pitch estimate $\hat{\omega}_0$ is [4, 13]

$$\hat{\omega}_0 = \arg\max_{\omega_0} \mathbf{x}^T \mathbf{\Pi}_B \mathbf{x}, \tag{B.13}$$

where $\mathbf{\Pi}_B = \mathbf{B}(\mathbf{B}^H\mathbf{B})^{-1}\mathbf{B}^H$ with $(\cdot)^H$ denoting the hermitian transpose. As we are here concerned with pitch estimation in coloured noise, an LPC pre-whitener can be applied to the noisy vector $\mathbf{x}$ since asymptotically this only modifies the complex amplitude vector $\mathbf{c}$ [6] and not ω_0. Solving (F.9) in a fast way is described in [12].

In the tested Keele database excerpts, the pitches which were manually annotated are considered here as the ground truth [21]. In order to match the available ground truth, segments of duration 30 ms and an overlap of 20 ms between them were used for the pitch estimation setup. Babble and exhibition noise were added to the testing sentences at SNRs from -4 to 10 dB. After pre-whitening the noisy signals, the pitch was estimated in an interval $[60, 380]$ Hz, with a maximum possible of 15 harmonics. The evaluation was done in terms of gross error rates (GER), which is the

proportion of frames where both the ground truth and the pitch estimator result in the presence of a pitch (i.e., $\hat{L} > 0$), where the relative error of the estimated pitch is larger than a certain percentage [22]. Here we use 10%. An LPC pre-whitening order $P = 30$ is used for both scenarios, since from the SFM experiment we saw that with a higher P the noise can get closer to white. As a reference, the pitch was also estimated without any pre-whitening (WGN assumption). The results are depicted in the first row of Fig. F.2.

We also conducted an experiment with a specific speaker of the CMU Arctic database, for which a codebook was trained on 24 minutes of speech material (with the same parameters as the general speech codebook), and then we evaluated the pitch estimates on 40 sentences from the same speaker, not included in the training. The evaluation was also done with 30 ms segments, with an overlap of 20 ms between them. For this case, the ground truth was obtained by estimating pitches from the clean speech segments using (F.9). We also evaluated the pitch estimation performance on 40 sentences from same speakers of the general speech codebook, which were not used for the training. Results for the specific speaker are depicted in the second row of Fig. F.2, and for general speakers in the last row of Fig. F.2.

It is seen that the suggested pre-whitener helps better in reducing the GER of the pitch estimates, in comparison to others based on well-known noise PSD estimates (MS and MMSE), since for both noise types, the parametric NMF pre-whitener performance is the closest to the oracle one. In fact, for exhibition noise type the performance gets very similar to the oracle pre-whitening, implying that a more accurate noise PSD could be captured. We speculate that this is due to that the exhibition noise is more stationary.

4.3 Speech enhancement

Table B.3: Results of segSNR improvement in babble noise

Enhanc. method w/ OM-LSA	**segSNR improvement (in dB)**			
	-2dB	***1dB***	***4dB***	***7dB***
MS pre-wh	2.74±0.20	2.61±0.20	2.43±0.23	2.17±0.28
MMSE pre-wh	3.15±0.22	2.94±0.24	2.63±0.30	2.30±0.40
ParNMF pre-wh	3.75±0.29	3.28±0.25	2.71±0.33	2.07±0.51
Conv.(no pre-wh)	3.41±0.25	3.06±0.26	2.63±0.36	2.19±0.52

Finally, we verify that using the proposed pre-whitener as a pre-processor will not ruin the signal. The approach is to do speech enhancement on the

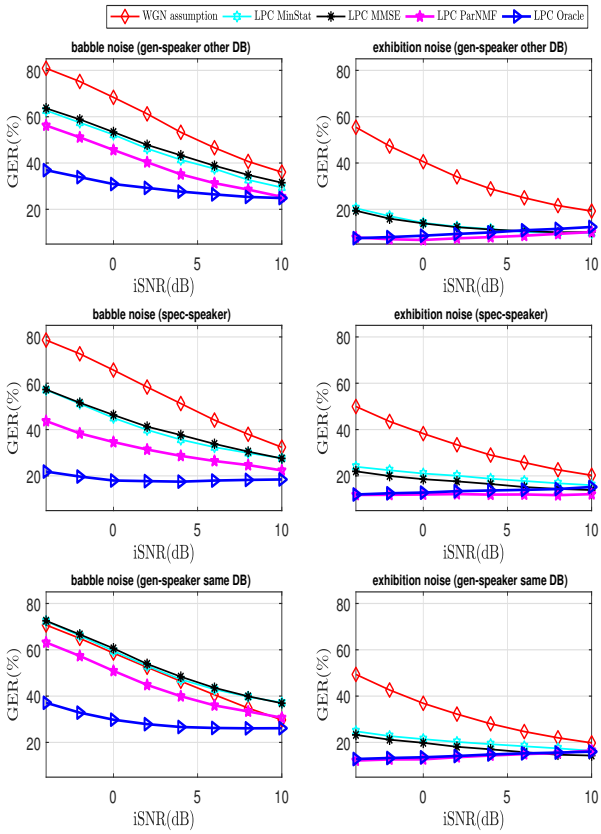

Fig. B.2: Gross error rate (GER) as a function of SNR.

pre-whitened noisy signal from a WGN assumption (i.e., with the WGN variance as the single noise parameter), and then undoing the pre-whitening by applying the inverse of the pre-whitening filter. It is important to note that we do not encourage to pre-whiten a noisy signal before enhancing it in a real setup, it only serves as a mean of verification of the presented pre-whitener. We use the optimally modified log-spectral amplitude estimator (OM-LSA) [23] algorithm for this enhancement task. The WGN variance is also calculated when one computes the noise prediction coefficients from the Levinson-Durbin recursion, as explained in Sec. II. In order that this WGN variance does not change abruptly, a recursive smoothing with a smoothing factor of 0.88 is used after computing the noise PSD from (B.4).

The evaluation is done under babble noise conditions, and again a pre-whitening order $P = 30$ is used, including also pre-whitening based on MS and MMSE. Noisy speech is also enhanced without applying a pre-whitener, i.e., by using a conventional noise PSD estimate [10] directly with OM-LSA.

Table B.4: Results of PESQ in babble noise

Enhanc. method w/ OM-LSA	PESQ			
	-2dB	***1dB***	***4dB***	***7dB***
Noisy Speech	1.63±0.15	1.77±0.12	1.96±0.11	2.16±0.10
MS pre-wh	1.71±0.08	1.93±0.08	2.16±0.07	2.39±0.07
MMSE pre-wh	1.72±0.08	1.94±0.07	2.17±0.06	2.40±0.05
ParNMF pre-wh	1.74±0.10	1.97±0.07	2.21±0.05	2.41±0.04
Conv.(no pre-wh)	1.73±0.09	1.95±0.08	2.18±0.06	2.41±0.05

Segmental SNR improvement and PESQ are reported in Tables III and IV, where 95% confidence intervals are seen for each value. In general, the performance from the proposed pre-whitener is better in comparison to the other pre-whiteners since it results in a higher average segSNR improvement (below 7dB) and higher average PESQ. Similar results to the conventional enhancement method are seen by using the presented pre-whitener, which indicates that a signal can be recovered even if it was pre-whitened for another purpose.

5 Conclusions

In this work, we proposed a new adaptive NMF based pre-whitener with pre-trained spectral envelopes parametrized with AR coefficients. The proposed pre-whitener achieves a higher spectral flatness in comparison to pre-whiteners based on classical noise PSD estimators, and therefore reduces considerably the pitch errors. Speech enhancement results based on the WGN assumption show that the pre-whitener can preserve the signal of interest. A fundamental question is why one would pre-whiten the signal instead of just enhancing it, so further research in answering this question should be conducted.

References

[1] B. G. Quinn, “Efficient estimation of the parameters in a sum of complex sinusoids in complex autoregressive noise,” in *2007 Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers*, Nov 2007, pp. 636–640.

[2] M. G. Christensen and S. H. Jensen, “Variable order harmonic sinusoidal parameter estimation for speech and audio signals,” in *2006*

Fortieth Asilomar Conference on Signals, Systems and Computers, Oct 2006, pp. 1126–1130.

[3] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 795–805, April 1991.

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

[5] M. G. Christensen, "Accurate estimation of low fundamental frequencies from real-valued measurements," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2042–2056, 2013.

[6] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "A study on how pre-whitening influences fundamental frequency estimation," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May, pp. 6495–6499.

[7] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Instantaneous fundamental frequency estimation with optimal segmentation for non-stationary voiced speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2354–2367, Dec 2016.

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

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

[10] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based Noise Power Estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[11] J. K. Nielsen, M. Kavalekalam, M. Christensen, and J. Boldt, "Model-based noise PSD estimation from speech in non-stationary noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

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

[13] P. Stoica and R. L. Moses, "Spectral analysis of signals," *Pearson*, 2005.

[14] M. S. Kavalekalam, J. K. Nielsen, L. Shi, M. G. Christensen, and J. Boldt, "Online parametric NMF for speech enhancement," in *2018 26th European Signal Processing Conference*, 2018, pp. 2320–2324.

[15] C. F. E. Vincent, A. Ozerov, "Single-channel audio source separation with nmf: Divergences, constraints and algorithms. audio source separation." *Springer*, pp. 1–24, 2018.

[16] S. M. Kay, *Modern spectral estimation*. Pearson Education, 1988.

[17] R. M. Gray, "Toeplitz and circulant matrices: A review," *Foundations and trends in communications and information theory, Vol. 2, Iss. 3, p.155-239*, vol. 2, pp. 155–239, 2005.

[18] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[19] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, January 1980.

[20] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[21] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *EUROSPEECH*, 1995.

[22] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *INTERSPEECH*, 2011.

[23] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113–116, 2002.

Paper C

Robust Fundamental Frequency Estimation in Coloured Noise

Alfredo Esquivel Jaramillo, Andreas Jakobsson, Jesper Kjær Nielsen, Mads Græsbøll Christensen

The paper has been published in the
Proceedings IEEE Int. Conf. Acoust., Speech, Signal Process. pp. 741–745, 2020.

Abstract

Most parametric fundamental frequency estimators make the implicit assumption that any corrupting noise is additive, white Gaussian. Under this assumption, the maximum likelihood (ML) and the least squares estimators are the same, and statistically efficient. However, in the coloured noise case, the estimators differ, and the spectral shape of the corrupting noise should be taken into account. To allow for this, we here propose two schemes that refine the noise statistics and parameter estimates in an iterative manner, one of them based on an approximate ML solution and the other one based on removing the periodic signal obtained from a linearly constrained minimum variance (LCMV) filter. Evaluations on real speech data indicate that the iteration steps improve the estimation accuracy, therefore offering improvement over traditional non-parametric fundamental frequency methods in most of the evaluated scenarios.

1 Introduction

The problem of estimating the fundamental frequency (a.k.a. pitch) of a periodic signal has received considerable attention during recent decades, and is of particular importance in many forms of audio and speech processing, such as speaker identification [1], audio coding [2], music transcription [3], and speech decomposition [4]. As opposed to correlation-based methods (e.g. YIN [5], RAPT [6]), parametric estimators [7] exploit a parametric model of the signal structure, which, if correct, allows for estimators that are more robust and that offer better resolution [8]. Many forms of parametric estimators, e.g., those based on subspace orthogonality [9], assume that the additive noise is white and Gaussian distributed (WGN), something that is rare in practice. A common consequence of this is that the found estimate is a rational number of the actual fundamental frequency when they are applied in practical noise scenarios, causing so-called octave errors. This effect may be alleviated by taking the spectral shape of the additive noise into account, which can, for example, be done by modelling the noise as an autoregressive (AR) process. Formulated mathematically, the problem may thus be expressed as follows: A set of harmonically related sinusoids, with frequencies $\{\omega_l\}$, are assumed to be observed corrupted by an additive AR noise, $e(n)$, for $n = 0, ..., N-1$, such that

$$x(n) = s(n) + e(n) = \sum_{l=-L, l\neq 0}^{L} \alpha_l e^{j\omega_l n} + e(n), \tag{C.1}$$

where L is the number of harmonics and $\alpha_l = \alpha_{-l}^*$ denotes the complex amplitude of the lth harmonic. For voiced speech segments, it is often

assumed that the harmonics are exact integer multiples of the fundamental ω_0, i.e., $\omega_l = \omega_0 l$, leading to the so-called harmonic model. Under the assumption that the additive noise may be well modelled as an AR process, it further holds that

$$e(n) = -\sum_{i=1}^{P} a_i e(n-i) + w(n), \tag{C.2}$$

where $\{a_i\}_{i=1}^{P}$ are the noise AR parameters and $w(n)$ is a driving zero-mean WGN process with variance σ_w^2.

Regrettably, jointly estimating the parameters detailing both the speech $(\{\omega_l\}, \{\alpha_l\}, L)$ and the noise $(\{a_i\}, \sigma_w^2)$ is computationally prohibitive, being a multimodal and multidimensional optimization problem [10], although, reminiscent of the mixed-spectrum estimation problem presented in [11], the problem described herein may be solved in a cascaded approach, where the sinusoidal parameters and the AR noise parameters are estimated separately. However, the problems differ in two significant ways. Firstly, in [11], the signal is assumed to consist of independent sinusoids (i.e., not harmonically related) in AR noise, whereas we strive to exploit the harmonic structure of the sinusoids to allow for improved estimates [12]. Secondly, in [11], a single iteration of the procedure was sufficient for convergence, since estimating independent sinusoids under the WGN assumption is asymptotically efficient, even for coloured noise [13] but not for fundamental frequency estimation.

In the problem considered herein, estimating ω_0 without taking the AR structure into account will increase the risk of selecting an erroneous peak as the estimate, causing the noted octave error [14], and from the above discussion it is suggested that the estimates of the noise and signal parameters should rather be done in an iterative manner. This may be done by first estimating the sinusoidal frequencies without exploiting the harmonic structure, which could then be incorporated using a weighting reminiscent of the extended invariance principle (EXIP) [15]. An alternative, which is examined here, is to first form an estimate of the noise shape, and then use this in a pre-whitening step prior to estimating ω_0 (such a filtering step will not change the frequency content of the signal, merely the corresponding amplitudes [14]). However, in order to allow for reliable estimates, accurate noise AR parameters are required. For this purpose, accurate noise statistics are needed and this topic has attracted significant interest, for instance in classical algorithms such as minimum statistics (MS) [16] and minimum MSE based on speech presence probabilities (MMSE-SPP) [17], both which perform well when the noise is fairly stationary. However, for non-stationary noise types, such as babble noise, the noise parameters accuracy and the pre-whitening performance can be improved by taking

into account prior spectral information on the AR-parameters of speech and noise sources [18, 19]. In this paper, extending upon the work in [11, 14, 19], we investigate two schemes for reducing the likelihood of octave errors using an iteratively refined pre-whitening filter. Both proposed methods are based on estimating the error sequence, from which a new pre-whitening filter may then be directly obtained.

2 Model, problem and proposed method

To introduce notation and properly formulate the problem, we proceed to introduce the fundamental frequency estimator along with useful matrix and vector definitions, and discuss how the noisy signal can be pre-whitened. Consider a signal segment of N samples,

$$\mathbf{x} = \begin{bmatrix} x(0) & x(1) & \cdots & x(N-1) \end{bmatrix}^T, \tag{C.3}$$

with $(\cdot)^T$ denoting the transpose. Then, (C.1) may be written as

$$\mathbf{x} = \mathbf{s} + \mathbf{e} = \mathbf{Z}_L(\omega_0)\boldsymbol{\alpha} + \mathbf{e}, \tag{C.4}$$

with $\mathbf{e}$ defined similar to $\mathbf{x}$, and

$$\mathbf{Z}_L(\omega_0) = \begin{bmatrix} \mathbf{z}(\omega_0) & \mathbf{z}^*(\omega_0) & \cdots & \mathbf{z}(\omega_0 L) & \mathbf{z}^*(\omega_0 L) \end{bmatrix}, \tag{C.5}$$

$$\mathbf{z}(\omega) = \begin{bmatrix} 1 & e^{-j\omega} & \cdots & e^{-j\omega(N-1)} \end{bmatrix}^T, \tag{C.6}$$

$$\boldsymbol{\alpha} = \frac{1}{2}\begin{bmatrix} A_1 e^{j\psi_1} & \cdots & A_L e^{j\psi_L} & A_L e^{-j\psi_L} \end{bmatrix}^T, \tag{C.7}$$

where $A_l > 0$ denotes the (real-valued) amplitude and $\psi_l \in [0, 2\pi)$ the initial phase. For a not-voiced speech segment (including unvoiced speech and pauses), the observed signal model thus reduces to $\mathbf{x} = \mathbf{e}$. Both models may be expressed jointly as $\mathbf{x} = u\mathbf{Z}_L(\omega_0)\boldsymbol{\alpha} + \mathbf{e}$, where $u = 1$ for a voiced segment, and 0 otherwise. For white Gaussian noise, the ML estimate of $\hat{\omega}_0$ is

$$\hat{\omega}_0 = \arg\max_{\omega_0} \mathbf{x}^T \boldsymbol{\Pi}_{Z_L}(\omega_0)\mathbf{x}, \tag{C.8}$$

where $\boldsymbol{\Pi}_{Z_L}(\omega_0) = \mathbf{Z}_L(\omega_0)\left[\mathbf{Z}_L^H(\omega_0)\mathbf{Z}_L(\omega_0)\right]^{-1}\mathbf{Z}_L^H(\omega_0)$, which depends on the (unknown) candidate model order, L. This is the estimator that we will here refer to as the NLS (nonlinear least-squares) estimator. Fortunately, (F.9) may be solved efficiently in an order-recursive manner [8], after which a suitable order may be selected using a model selection criteria such as the Bayesian Information Criteria (BIC) [20, 21]. The resulting estimate would only be statistically efficient if $\mathbf{e}$ was white. As we are here concerned with

coloured noise, the AR noise parameters need to be first estimated and used to pre-whiten the signal using the filter

$$A(\omega) = 1 + \sum_{i=1}^{P} a_i e^{-j\omega i}. \tag{C.9}$$

In order to estimate the noise parameters, the noise spectral density (PSD) $\phi_e(k),\ k = 0, 1, ...,\ N-1$, can be estimated using algorithms such as MS, MMSE-SPP, the parametric NMF (Par-NMF) [19], or the model-based method introduced in [18]. Using the estimated noise PSD, the noise autocovariance sequence is then estimated as $r_e(n) = \frac{1}{N}\sum_{k=0}^{N-1}\phi_e(k)\exp\left(j\frac{2\pi}{N}nk\right), 0 \leq n \leq P$. Finally, a Levinson-Durbin recursion of order P is applied on $r_e(n)$ to determine the $\{a_i\}_{i=1}^{P}$ filter coefficients. Then, this pre-whitening filter is applied to $\mathbf{x}$, and the initial $\hat{\omega}_0$ is obtained from (8).

What has been described up to now, does not involve a reestimation step, and we now proceed to detail on how the parameters are reestimated. In a first approach, using the harmonic structure, for a given $\hat{\omega}_0$, the least squares (LS) estimate of the amplitudes may be formed as [9]

$$\hat{\boldsymbol{\alpha}} = [\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{Z}_L(\hat{\omega}_0)]^{-1}\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{x}. \tag{C.10}$$

Using the resulting estimate, the additive noise may be estimated by removing the harmonic model contribution from the observed signal, such that $\hat{\mathbf{e}} = \mathbf{x} - \mathbf{Z}_L(\hat{\omega}_0)\hat{\boldsymbol{\alpha}}$. From this estimate, the AR noise parameters $\{\hat{a}_i\}_{i=1}^{P}$ may be reestimated using the *autocorrelation method* (see, e.g., [22]) of AR modeling, which may then be used to form a new pre-whitened signal vector, from which a new estimate $\hat{\hat{\omega}}_0$ can be obtained. This process can then be repeated until convergence, which is here defined as when the cost function (F.9) between two consecutive iterations is below a given threshold value. We refer to this method as the approximate ML approach.

The second possibility is to apply an optimal filter, capable of extracting a desired periodic signal, which satisfies the harmonic model. For this purpose, we make use of the noise covariance matrix, defined as $\mathbf{R_e} = E\left[\mathbf{e}\mathbf{e}^T\right]$, where $E[\cdot]$ is the mathematical expectation operator. The applied filter will be driven by the estimated fundamental frequency $\hat{\omega}_0$ and by the estimated model order $\hat{L}$. A linear filter is applied to $\mathbf{x}$ in order to extract an arbitrary signal sample $s(n-m)$, i.e.,

$$\hat{s}(n-m) = \mathbf{h}^T\mathbf{x} = \mathbf{h}^T\mathbf{Z}_L(\hat{\omega}_0)\boldsymbol{\alpha} + \mathbf{h}^T\mathbf{e}, \tag{C.11}$$

where $\mathbf{h} = [h_0\ h_1\ \ldots\ h_{N-1}]^T$. It is seen that the filter affects both the speech and noise components. In order to obtain a distortionless estimate of the voiced speech sample, the constraint $\mathbf{h}^T\mathbf{Z}_L(\hat{\omega}_0) = \mathbf{b}_m^T\mathbf{Z}_L(\hat{\omega}_0)$ is imposed, which implies that the harmonics of the desired signal will not be distorted.

Here, $\mathbf{b}_m^T$ corresponds to the m^{th} column of the $N \times N$ identity matrix. The problem for extracting a sample of the desired periodic signal is to minimize the residual noise variance (i.e., $E\left[(\mathbf{h}^T\mathbf{e})^2\right]$) with the above constraint, i.e.,

$$\min_{\mathbf{h}} \mathbf{h}^T\mathbf{R}_e\mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^T\mathbf{Z}_L(\hat{\omega}_0) = \mathbf{b}_m^T\mathbf{Z}_L(\hat{\omega}_0). \tag{C.12}$$

The filter resulting from this optimization problem is the linearly constrained minimum variance (LCMV) filter [23] and is given by

$$\mathbf{h}_{\text{LCMV}} = \mathbf{R}_e^{-1}\mathbf{Z}_L(\hat{\omega}_0)\left(\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{R}_e^{-1}\mathbf{Z}_L(\hat{\omega}_0)\right)^{-1}\mathbf{b}_m^T. \tag{C.13}$$

The constraints of the problem can also be modified to estimate the entire speech vector as

$$\hat{\mathbf{s}} = \mathbf{H}^T\mathbf{x} = \mathbf{H}^T\mathbf{Z}_L(\hat{\omega}_0)\boldsymbol{\alpha} + \mathbf{H}^T\mathbf{e}, \tag{C.14}$$

which for being distortionless must satisfy $\mathbf{H}^T\mathbf{Z}_L(\hat{\omega}_0) = \mathbf{Z}_L(\hat{\omega}_0)$. This leads to the optimization problem

$$\min_{\mathbf{H}} \text{Tr}\left\{\mathbf{H}^T\mathbf{R}_e\mathbf{H}\right\} \quad \text{s.t.} \quad \mathbf{H}^T\mathbf{Z}_L(\hat{\omega}_0) = \mathbf{Z}_L(\hat{\omega}_0). \tag{C.15}$$

The solution of this problem is given by

$$\mathbf{H}_{\text{LCMV}} = \mathbf{R}_e^{-1}\mathbf{Z}_L(\hat{\omega}_0)\left(\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{R}_e^{-1}\mathbf{Z}_L(\hat{\omega}_0)\right)^{-1}\mathbf{Z}_L^H(\hat{\omega}_0) \tag{C.16}$$

It is worth noting that one may here directly use the Gohberg-Semencul (GS) formula (see, e.g., [24]) to form the matrix inverses in closed form using the already estimated noise AR parameters:

$$\hat{\mathbf{R}}_e^{-1} = \frac{1}{\sigma_w^2}\left\{\begin{bmatrix} 1 & & & 0 \\ a_1 & \ddots & & \\ \vdots & \ddots & \ddots & \\ a_P & \dots & a_1 & 1 \end{bmatrix}\begin{bmatrix} 1 & a_1 & \dots & a_P \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_1 \\ 0 & & & 1 \end{bmatrix} - \begin{bmatrix} 0 & & & 0 \\ a_P & \ddots & & \\ \vdots & \ddots & \ddots & \\ a_1 & \dots & a_P & 0 \end{bmatrix}\begin{bmatrix} 0 & a_P & \dots & a_1 \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_0 \\ 0 & & & 0 \end{bmatrix}\right\} \tag{C.17}$$

The harmonic signal is then estimated as $\hat{\mathbf{s}} = \mathbf{H}_{\text{LCMV}}^T\mathbf{x}$, yielding the noise estimate $\hat{\mathbf{e}} = \mathbf{x} - \hat{\mathbf{s}}$, from which noise AR parameters can then be reestimated. A new pre-whitening filter is applied and a new estimate $\hat{\omega}_0$ is reestimated. As in the ML approach, a similar reiteration between estimating the noise AR parameters and estimating the fundamental frequency to drive the

LCMV filtering is possible, being repeated until convergence of the cost function (F.9). We refer to this as the LCMV filtering approach.

In both approaches, when the not-voiced model is favored (i.e., $\hat{L} = 0$), the estimated noise vector is $\hat{\mathbf{e}} = \mathbf{x}$, and if in the next iteration the segment is still detected as not-voiced, the process is stopped for that segment.

3 Experimental setup

We now proceed to experimentally evaluate the performance of the introduced method as compared to some well-known non-parametric methods, namely YIN, RAPT, and the Cepstrum-based method introduced in [25], here denoted Cepstrum. The speech material used for evaluation is the ten sentences in the Keele database [26], resampled to 8 kHz. This database has an annotated ground truth, which corresponds to an estimate obtained using RAPT. In the evaluation, we discard labeled segments with a negative value, i.e., we only considered voiced and not-voiced segments which have certainty of the annotated values (see [26] for further details). The ground truth values were obtained for segment lengths of 26.5 ms, with a shift of 10 ms between them. The same segment length and rate are used for all the methods. The signals are corrupted by additive babble, factory, and F-16 noise types from the NOISEX-92 database [27], at iSNRs of -5, 5, and 15 dB. The iSNR indicates the level of the clean speech signal relative to the noise component in the noisy signal, i.e. iSNR $= \frac{\sigma_s^2}{\sigma_e^2}$, where σ_s^2 is the variance of the speech signal, and σ_e^2 is the variance of the noise signal.

To assess the performance, both the fundamental frequency estimation accuracy and the voicing detection are of interest. Firstly, we use gross error rate (GER) to compute the proportion of segments where both the reference and the estimated values result in a voiced segment, and differ in more than 20%. The percentage of voiced/not-voiced detection errors is known as the voicing decision error (VDE). It is desirable to have low values of both the GER and the VDE, however some estimators may have a high VDE even if they presented a low GER as many not-voiced segments could be wrongly classified as voiced, and vice-versa. Therefore, in [28], a performance measure known as the full frame error (FFE) was proposed, which considers all kinds of possible errors: GERs, not-voiced segments wrongly classified as voiced, and voiced segments misclassified as not-voiced.

The ω_0 estimation is done on the interval [60,400] Hz for all methods. For the NLS estimator, a maximum model order of $L = 27$ is used, and the not-voiced case, i.e., $L = 0$ is considered as well. To allow that the fast NLS estimator yields accurate estimates, the signal is first pre-whitened. The AR pre-whitening order in (C.9) is set to $P = 25$. The applied pre-whitener is

the one based on the parametric-NMF noise PSD estimate described in [19], for which a dictionary that contains typical speech and noise spectral envelopes is required. To build the dictionary, speech and noise codebooks were trained offline using a standard vector quantization technique (i.e., the Lloyd algorithm) [29]. The training is done on LSF coefficients on segments of 26.5 ms duration, with a time shift between segments of 10 ms. The quantized LSF coefficients are converted back to linear prediction coefficients of order 12. Once the speech and noise codebooks are obtained, the spectral envelopes corresponding to each codebook entry can be arranged as columns of the dictionary matrix (as described in [19]). In our case, a speech codebook of 32 entries was trained on 54 minutes of several sentences from the CMU Arctic database [30], from 4 different speakers (2 female and 2 male), resampled from 16 to 8 kHz. A noise codebook of 16 entries was trained on samples from the NOISEX-92 database of babble, F-16, factory, and street noise, resampled at 8 kHz. It is important to note that the noise samples used for the training are not the same ones used for the evaluation, and also that the speech codebook involves different speakers from the evaluation.

4 Experimental Results

We first demonstrate that the proposed reiteration scheme is able to correct wrong initial estimates. Figure F.1 illustrates the ground truth and the estimated fundamental frequencies of 650 overlapping segments (approx 6.5 s) of a female speech signal excerpt of the Keele database. The clean signal is added factory noise at an iSNR of 15 dB. The ground truth is plotted in black circles, where a value of 0 corresponds to a not-voiced segment. In the top figure, the estimates which were obtained from the NLS estimator (after applying the pre-whitening), without the reiteration steps, are displayed. It may be seen that many segments which are not-voiced are wrongly estimated as voiced. Applying reestimation using the LCMV filtering technique (bottom figure), one may note that many of those segments are now correctly detected as not-voiced.

Next, the performance as a function of the iSNR is investigated, computed using 6 Monte-Carlo simulations, for each noise type, at each iSNR and for each one of the Keele files. The results for the three noise types in terms of GER, VDE, and FFE are shown in Figure F.2, including with the corresponding 95% confidence interval. As YIN does not perform voicing detection, it is here coupled with the voicing decisions of the summation of residuals harmonics (SRH), as was also done in [31]. The NLS-NMF notation implies no re-estimation, where ω_0 is estimated only one time from (8), after the pre-whitening filter from (9) is applied. The NLS-NMF

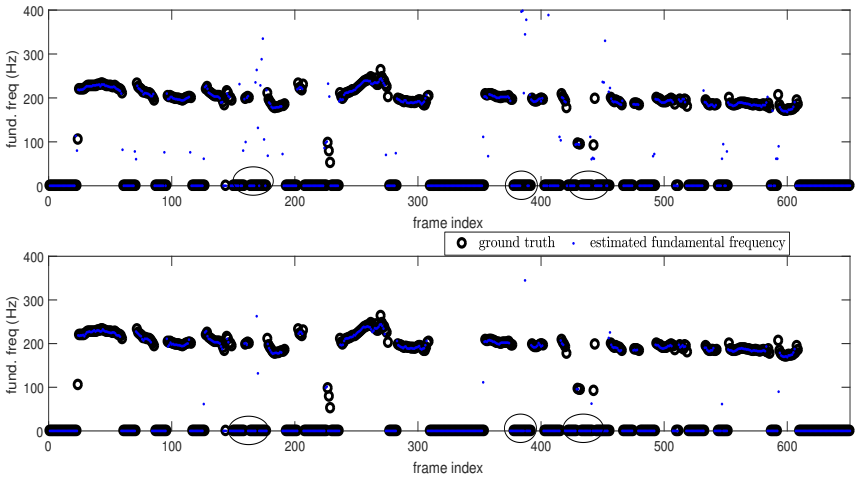

Fig. C.1: Fundamental frequency ground truth and estimates without (top) and with (bottom) the proposed LCMV filtering iteration scheme.

Iter1 and Iter2 notation correspond to the iterative scheme based on the approximate ML approach and the LCMV filtering approach, respectively. We have found that convergence in both approaches typically requires 4 to 5 iterations for a voiced segment and 2 to 3 for a not-voiced one. It is important to point out that these approaches result in independent ω_0 estimates between all segments, as opposed to the other methods which include a final step of refinement, using, for instance, dynamic programming or a best local estimate selection.

First, it is noted that both presented iterative schemes result in similar performance. Furthermore, the improvements from applying the reiteration step are more evident at higher SNRs (i.e., 5 and 15 dB), as the confidence intervals are not overlapping such as in the -5 dB case. Next, it is observed that the Cepstrum method presents the lowest GER, although it results in higher voicing detection errors than the NLS estimator, even if the reiteration is not applied. Both the YIN and RAPT results are worse in terms of GER than the NLS estimator, even without the reiteration, at -5 and 5 dB. RAPT seems to better in terms of GER at 15 dB compared to NLS if the reiteration is not applied, however the performance from the approximate ML and the LCMV filtering reestimation is improved. Lower voicing detection errors are seen from the proposed methods under babble noise conditions, even without the reestimation, as compared to the three non-parametric estimators. The NLS method, even without the reestimation, also has lower VDE as compared to YIN, in the F16 and factory noise scenarios. Comparing to RAPT, the proposed methods (with and

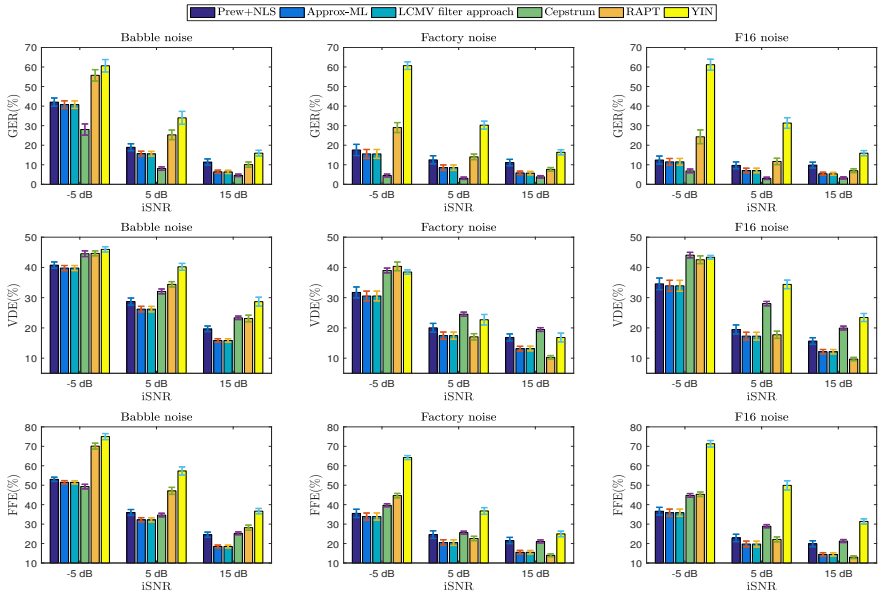

Fig. C.2: The GER, voicing detection errors and total frame error of the estimated fundamental frequency, for different SNRs, for the Keele database with different noise types.

without the re-estimation) have lower VDE at -5 dB, for both F16 and factory noise. However, at 5 and 15 dB, similar voicing detection errors are observed when using the re-estimation schemes. It is important to remember that RAPT makes use of a final dynamic programming stage, which also takes the neighbor values into account, which is not the case for the NLS estimator. In terms of full frame errors, in babble noise conditions, the proposed methods, even if there was no reiteration, have better performance than RAPT and YIN. Similar performance to the Cepstrum method is seen at -5 dB, while at 5 dB and 15 dB, the performance of NLS with reestimation is better. For factory and F16 noise scenarios, the proposed reiteration scheme yields lower FFE as compared to Cepstrum and YIN, at all SNRs, and also compared to RAPT at -5 and 5 dB. It may be noted that RAPT seems to be slightly better at 15 dB, although it should be recalled that the ground truth estimates were obtained with that method.

5 Discussion

This paper considered the topic of fundamental frequency estimation in coloured noise scenarios. Most estimators make an implicit assumption that the corrupting noise is an additive, white Gaussian process, for which case the least squares estimate is statistically efficient. In practice, the additive

noise shape should be taken into account in order to avoid octave errors, which may be done using a pre-whitening scheme using the estimated noise parameters. In this work, we do so by forming an AR model for the noise corrupting the speech segments, allowing us to form the required pre-whitening filter. By then estimating the harmonic components, the estimate of the additive noise may be improved, allowing for an improved pre-whitening filter, which in turn allows for an improved pitch estimate. By iteratively refining the estimates in this manner, one may reduce the risk of octave errors noticeably. Evaluated on measured speech data, we conclude that the NLS estimator reduces the number of full frame errors in most of the scenarios and therefore can offer better performance than the state-of-the-art non-parametric estimators, although only when the reiteration scheme is applied. Even without taking the correlation of consecutive estimates into account (i.e., tracking capabilities), the proposed method is more robust to the noise.

References

[1] A. O. T. Hogg, C. Evers, and P. A. Naylor, "Speaker change detection using fundamental frequency with application to multi-talker segmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5826–5830.

[2] E. Vincent and M. D. Plumbley, "Low bit-rate object coding of musical audio using Bayesian harmonic models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1273–1282, May 2007.

[3] N. Kroher and E. Gómez, "Automatic transcription of flamenco singing from polyphonic music recordings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 901–913, May 2016.

[4] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "On optimal filtering for speech decomposition," in *26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2325–2329.

[5] A. D. Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[6] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

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

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

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

[10] S. M. Kay and V. Nagesha, "Maximum likelihood estimation of signals in autoregressive noise," *IEEE Transactions on Signal Processing*, vol. 42, no. 1, pp. 88–101, Jan 1994.

[11] J. Li and P. Stoica, "Efficient mixed-spectrum estimation with applications to target feature extraction," *IEEE Transactions on Signal Processing*, vol. 44, no. 2, pp. 281–295, Feb 1996.

[12] F. Elvander, J. Ding, and A. Jakobsson, "On harmonic approximations of inharmonic signals," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 5360–5364.

[13] 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 Transactions on Signal Processing*, vol. 45, no. 8, pp. 2048–2059, 1997.

[14] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "A study on how pre-whitening influences fundamental frequency estimation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6495–6499.

[15] P. Stoica and T. Söderström, "On reparametrization of loss functions used in estimation and the invariance principle," *Signal processing*, vol. 17, no. 4, pp. 383–387, 1989.

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

[17] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based Noise Power Estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[18] J. K. Nielsen, M. S. Kavalekalam, M. G. Christensen, and J. Boldt, "Model-Based noise PSD estimation from speech in non-stationary noise," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5424–5428.

[19] A. E. Jaramillo, J. Nielsen, and M. Christensen, "Adaptive prewhitening based on parametric NMF," in *2019 27th European Signal Processing Conference*, September 2019.

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

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

[22] P. Stoica, *Introduction to spectral analysis*. Prentice Hall, 1997.

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

[24] P. Stoica and R. L. Moses, "Spectral analysis of signals," *Pearson*, 2005.

[25] A. M. Noll, "Cepstrum pitch determination," *The journal of the acoustical society of America*, vol. 41, no. 2, pp. 293–309, 1967.

[26] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *EUROSPEECH*, 1995.

[27] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[28] W. Chu and A. Alwan, "Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 3969–3972.

[29] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on communications*, vol. 28, no. 1, pp. 84–95, 1980.

[30] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[31] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

Paper D

An Adaptive Autoregressive Pre-whitener for Speech and Acoustic Signals Based on Parametric NMF

Alfredo Esquivel Jaramillo, Jesper Kjær Nielsen, Mads Græsbøll Christensen

The paper has been submitted to the
Applied Acoustics, 2021.

The layout has been revised.

Abstract

A common assumption in many speech and acoustic processing methods is that the noise is white and Gaussian (WGN). Although making this assumption results in simple and computationally attractive methods, the assumption is often too simple and crude in many applications. In this paper, we introduce a general purpose and online pre-whitener which can be used as a pre-processor with methods based on the WGN assumption, improving their reliability and performance in applications with colored noise. The pre-whitener is a time-varying FIR filter whose coefficients are found using a parametric non-negative matrix factorization (NMF), based on autoregressive (AR) mixture modeling of both the noise component and the signal component constituting the noisy signal. Compared to other types of pre-whiteners, we show that the proposed pre-whitener has the best performance, especially in applications with non-stationary noise. We also perform a large number of experiments to quantify the benefits of using a pre-whitener as a pre-processor for methods based on the WGN-assumption. The experiments focus on pitch estimation, where the WGN assumption is very popular, but examples with speech enhancement and time-of-arrival estimation are also included.

1 Introduction

In many speech and acoustic applications, the signal of interest is contaminated with noise. To cope with this, methods or estimators designed to extract the signal (or a quantity) of interest must be robust to the noise whose level and spectral shape are often unknown a priori. Like the signal of interest, a noise model can also be elicited from which a robust, joint estimator of the signal and noise model parameters can be derived (see examples in, e.g., [1–5]). A Gaussian noise model is popular, but estimating its covariance sequence or its parametrization jointly with the signal model parameters often leads to intractable estimators. Moreover, this approach is not very flexible since a new estimator has to be re-derived when the noise model changes. As an alternative to the joint approach, it is possible to keep using methods which were derived based on the much simpler WGN assumption, provided that a pre-whitener is used as a pre-processor so that the noise color of the pre-whitened signal is approximately white. Various acoustic and speech processing methods [6–13] have assumed that the noise is WGN to retain the mathematical simplicity of the problem and to achieve a fast implementation. However, if those WGN-based methods are applied without any form of pre-processing, certain problems may appear. An example of this can be found in pitch estimation where a pronounced

noise peak at low frequencies causes the pitch estimator to produce a pitch estimate which is an integer fraction of the true pitch [14]; an estimation error which is often referred to as the subharmonic error. To combat this, applying a pre-whitener as a pre-processor is desirable since the noise will be whitened, thereby better fulfilling the model assumptions made in the WGN-based method. As we show later, however, accurate information on the noise spectrum (or, equivalently, statistics) is needed to perform this pre-processing.

The task of applying a pre-whitening scheme has been important in several areas, such as wireless communications [15], remote sensing [16], sonar [17], biomedical engineering [18], speech [19–22] and acoustic array processing [23]. Pre-whitening of the noise component can be performed by, e.g., applying a general linear transformation. This can be a matrix such as the Cholesky factor of the inverse noise covariance matrix [14, 24] that will decorrelate the noise samples, but at the same time will modify the signal of interest, including the frequency content [20]. This was seen in [25], where the Cholesky factor was applied in the context of subspace-based speech enhancement. Also, in [14, 20] it was shown how applying the Cholesky factor as a whitening transformation to noisy signals will modify the harmonic model structure, often used in pitch estimation, of the desired voiced speech parts. An alternative way to pre-whiten the background noise is to apply a linear filter whose amplitude response is the inverse of the spectral shape of the noise. An example is the (autoregressive) AR pre-whitener [26] which is an FIR filter whose coefficients are the AR parameters [27] describing the noise spectral shape. This filter corresponds to the classical prediction error filter [28], and its application only modifies the sinusoidal amplitudes and phases of the desired signal and not its frequency content (asymptotically) [14]. Therefore, model-based estimators assuming WGN such as the (nonlinear least squares) NLS pitch estimator [6, 7, 29] can be reliably used after the signal has been pre-processed with that pre-whitening filter. It should be noted that fitting the noise (power spectral density) PSD with a smoothed spectrum, as the AR spectrum is, is preferable to directly using the estimated noise PSD coefficients to generate the FIR filter that counteracts the noise spectral shape, since this option could possibly lead to inaccurate estimates [14].

In this paper, we address the problem of how an AR pre-whitener should be computed so that statistical-based estimators assuming WGN [6, 7] achieve the best performance when cascaded with it. We do not only show that using a pre-whitener as a pre-processor improves the accuracy of such estimators, but also that using a pre-whitener, which relies on *a priori* spectral information, leads to a better performance than simply using a pre-whitener based on a noise PSD tracker, often used in speech enhancement applications [30, 31]. We also show that in some scenarios, non-parametric

pitch estimators [32, 33] can also benefit from a pre-whitener. To estimate the needed noise statistics, from which the pre-whitenining filter coefficients are derived, we use the recently introduced model in [34] where both the signal and noise are represented as a sum of time-varying AR processes. The estimation of the parameters of this model was performed using parametric NMF in [34] which is a generalisation of traditional NMF of superimposed Gaussian sources [35]. In the proposed pre-whitener, the noise statistics were obtained from the parametric NMF method and the AR dictionaries were pre-trained offline on typical envelopes of speech and noise sources represented by AR parameters. Given the pre-trained AR-dictionaries, the parametric NMF method continuously re-computes the activation coefficients which are the excitation noise variances of the pre-trained AR-spectra. The solution of the cascade of the AR pre-whitener with the NLS pitch estimator can be further improved by post-processing the initial estimates through iterative refinement, leading to an improved accuracy [36]. Some of the presented ideas have been outlined in [36, 37], on the basis of the parametric NMF formulation in [34], but we here investigated the conditions that lead for a higher noise whiteness and improved spectral estimation accuracy, which at the same time lead to an improved accuracy of pitch estimators as compared to the ones reported at [36, 37]. Moreover, we consider a source localization application, namely time of arrival (TOA) estimation [9].

The remainder of this paper is organized as follows. Section 2 introduces the related work. We describe how to obtain the AR pre-whitening filter on a segment-by-segment basis in Section 3. In Section 4, we describe how the pre-whitening filter coefficients can be obtained from a noise PSD estimate which relies on a parametric NMF that makes use of prior spectral information stored in AR dictionaries. The experimental setting, details for training the dictionaries, the performance measures and the discussion of the observed results are presented in Section 5. Finally, Section 6 concludes the presented work.

2 Related work

To cope with the non-stationary nature of both the signal of interest and the noise, a pre-whitener should update its parameters, e.g., on a segment-by-segment basis [24]. Unfortunately, the noise statistics needed in the pre-whitener are not known and estimating them from the noisy mixture is difficult. In the literature, the parameters of the pre-whitener are usually determined only from segments in which the desired signal is absent, i.e., where only the noise is present. For example, in a sonar application to detect a low-Doppler target [26], an AR pre-whitener obtained its parameters only

when the reverberation was assumed to be present, thus ignoring a more realistic scenario in which both the reverberation and the signal of interest coexist. Similarly, in [25], the noise statistics were only computed during speech silent periods obtained from a voice activity detector (VAD) [38]. Other works, such as [39], have assumed that the noise AR parameters describing the noise spectral shape are known beforehand, but this is unrealistic specially in non-stationary noise situations where the noise spectral shape changes quickly between segments.

For non-stationary noise, the noise statistics may change significantly during speech presence, and this will not be tracked when a VAD is used, potentially leading to a poor performance of the pre-whitener as well as the estimator assuming WGN. During speech presence, information about the noise spectrum can be tracked across time using various well-known state-of-the art methods (e.g., minimum statistics (MS) [30] and MMSE based on speech presence probabilities (SPP) [31]). Pre-whitening reliant on these approaches results in a good performance when the noise is stationary or slowly time-varying, but not when the noise is highly non-stationary. For pitch estimation, this was demonstrated in [14].

To cope with noise statistics estimation in nonstationary noise, a model-based estimator [40] using *a priori* spectral information about typical speech and noise AR parameters stored in pre-trained codebooks has been found to improve the noise PSD estimation accuracy compared to traditional noise tracking methods. As opposed to traditional codebook-based approaches which use a log-spectral distortion approximation and make use of noise classification [41, 42], multiplicative-update (MU) based approaches [43] have been shown to result in more accurate excitation variance estimates, i.e., they capture better the noise spectral envelope. For the introduced pre-whitener based on parametric NMF, the minimization of the spectral distance between the periodogram and the modelled PSD leads to an MU rule of the activation coefficients. It should be noted that the parametric NMF method differs from unsupervised approaches such as [35, 44] by parametrising the dictionary with normalised AR-envelopes.

3 AR pre-whitener

This section describes the basic principle of how an AR pre-whitening filter can be applied when the noise statistics are available. The details on how such noise statistics can be estimated are discussed in the next section. An observed signal $x(n)$ is assumed to be formed by the mixture of a clean signal of interest (e.g., speech) $s(n)$ and a colored noise signal $c(n)$, i.e.,

$$x(n) = s(n) + c(n) \; . \tag{D.1}$$

Furthermore, we assume that $c(n)$ is well modeled as an AR process, i.e.,

$$c(n) = -\sum_{i=1}^{P} w_c(i)c(n-i) + e(n)\ . \tag{D.2}$$

This means that $c(n)$ is modelled to be generated by passing white Gaussian excitation noise $e(n)$ with variance σ_e^2 through an all-pole filter with response

$$H(\omega) = \frac{1}{W(\omega)} = \frac{1}{1+\sum_{i=1}^{P} w_c(i)e^{-j\omega i}}, \tag{D.3}$$

where $\{w_c(i)\}_{i=1}^{P}$ denote the AR parameters which describe the spectral shape of the colored noise, and P is the AR order. This generative noise model is illustrated in the left part of Fig.(F.1)). For a stable AR process, the original white Gaussian excitation noise can be retrieved using the FIR filter $W(\omega)$ in the right part of Fig.(F.1)). Thus, the filter $W(\omega)$ is a whitening filter, and the prefix "pre" denotes that it is applied before some other method. The noise AR parameters $\{w_c(i)\}_{i=1}^{P}$ and the excitation

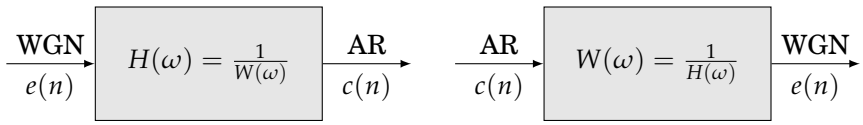

Fig. D.1: Generative noise model (left) and whitening FIR filter (right)

variance σ_e^2 are seldom known and must, therefore, be computed from the noise statistics. If the noise covariance sequence $\{r_c(i)\}_{i=0}^{P}$ is available, the AR parameters can be computed by solving the Yule-Walker equations [27]

$$\underbrace{\begin{bmatrix} r_c(0) & r_c(1) & \dots & r_c(P) \\ r_c(-1) & r_c(0) & \dots & r_c(P-1) \\ \vdots & \vdots & \vdots & \vdots \\ r_c(-P) & r_c(-P+1) & \dots & r_c(0) \end{bmatrix}}_{\mathbf{R}_{P+1}} \underbrace{\begin{bmatrix} 1 \\ w_c(1) \\ \vdots \\ w_c(P) \end{bmatrix}}_{\mathbf{w}_P} = \begin{bmatrix} \sigma_e^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{D.4}$$

which can be implemented efficiently using the Levinson-Durbin algorithm [27, 28]. If instead N uniform samples from the noise PSD $\{\Phi_c(k)\}_{k=0}^{N-1}$ is available, the noise covariance sequence can be computed as [27]

$$r_c(n) = \frac{1}{N}\sum_{k=0}^{N-1} \Phi_c(k)\exp\left(j\frac{2\pi}{N}nk\right), \quad 0 \le n \le P\ . \tag{D.5}$$

Due to the time-varying noise statistics, the AR parameters will be time-varying. In practice, we implement this by dividing the data into

overlapping segments, each of length N. Given such N data samples

$$\mathbf{x}(l) = \begin{bmatrix} x(0,l) & x(1,l) & \cdots & x(N-1,l) \end{bmatrix}^T \tag{D.6}$$

and time-varying AR parameters $\mathbf{w}_P(l)$ in segment l, the pre-whitener is implemented in the frequency domain. That is, the discrete Fourier transform (DFT) of the pre-whitened signal is computed as

$$\hat{X}_W(k,l) = W(k,l)X(k,l) \tag{D.7}$$

where $W(k,l)$ and $X(k,l)$ are the k^{th} bin of the N-length DFT of time-varying AR parameters $\mathbf{w}_P(l)$ and the data segment $\mathbf{x}(n+lM)v(n)$, respectively, where M denotes the hop size in samples between segments and $v(n)$ is the analysis window. The whitened signal in the time domain $x_W(n,l)$ is then obtained by computing an inverse DFT of $\{\hat{X}_W(k,l)\}_{k=0}^{N-1}$. As the processing is done on overlapping segments, a synthesis window $v(n)$ is applied to update the full pre-whitened signal as $x_W(n+lM) = x_W(n+lM) + v(n)x_W(n,l)$.

4 Noise PSD estimation based on parametric NMF

As mentioned above, segment-wise estimates of the noise PSD $\Phi_c(k)$ are required to compute the AR-coefficients used in the pre-whitening filter. In this section, we describe how the noise PSD is estimated in the proposed pre-whitener from a segment of data. Note that we omit the segment index l in this section to simplify the notation. To get good performance in even non-stationary noise conditions, we here propose that the noise PSD estimate is obtained by taking typical spectral shapes of speech and noise into account. For this purpose, we model the data vector in (D.6) as a summation of U AR processes $\{\mathbf{t}_u\}_{u=1}^{U}$ where each AR-process describe a typical spectral shape. Specifically, the data vector is modelled as

$$\mathbf{x} = \sum_{u=1}^{U} \mathbf{t}_u = \sum_{u=1}^{U_s} \mathbf{t}_u + \sum_{u=U_s+1}^{U} \mathbf{t}_u, \tag{D.8}$$

where the first U_s AR processes model clean signals (e.g., speech), and the last U_c AR processes model noise signals. A stationary and stable AR process can be described as a realisation from a multivariate Gaussian probability density function (pdf) [42], i.e., $\mathbf{t}_u \sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \mathbf{R}_u(\mathbf{a}_u))$, where σ_u^2 is the excitation variance, $\mathbf{R}_u(\mathbf{a}_u)$ is its gain normalized covariance matrix, and

$$\mathbf{a}_u = \begin{bmatrix} 1 & a_u(1) & \cdots & a_u(P') \end{bmatrix}^T \tag{D.9}$$

is the vector containing the AR parameters of the u^{th} spectral basis. Here P' is the AR order.

The likelihood of the observation $\mathbf{x}$ as a function of U excitation variances and U spectral shapes is given by

$$p(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{D}) \sim \mathcal{N}\left(\mathbf{0}, \sum_{u=1}^{U} \sigma_u^2 \mathbf{R}_u(\mathbf{a}_u)\right) \tag{D.10}$$

where

$$\boldsymbol{\sigma} = \begin{bmatrix}\sigma_1^2 & \cdots & \sigma_U^2\end{bmatrix}^T \tag{D.11}$$

is a $U \times 1$ vector containing the U excitation variances and is referred to as the vector of activation coefficients. The matrix $\mathbf{D}$ of dimension $N \times U$ is referred as either the spectral basis matrix or the AR dictionary, and its column vectors are the U gain normalized PSDs parametrised by the AR parameters, i.e.,

$$\mathbf{D} = \begin{bmatrix} d_1(0) & \cdots & d_u(0) & \cdots & d_U(0) \\ \vdots & & \vdots & & \vdots \\ d_1(k) & \cdots & d_u(k) & \cdots & d_U(k) \\ \vdots & & \vdots & & \vdots \\ d_1(N-1) & \cdots & d_u(N-1) & \cdots & d_U(N-1) \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{d}}_0 \\ \vdots \\ \widetilde{\mathbf{d}}_k \\ \vdots \\ \widetilde{\mathbf{d}}_{N-1} \end{bmatrix} \tag{D.12}$$

$$= \begin{bmatrix}\mathbf{d}_1 & \cdots & \mathbf{d}_u & \cdots & \mathbf{d}_U\end{bmatrix} . \tag{D.13}$$

where $\widetilde{\mathbf{d}}_k$ and $\mathbf{d}_u$ the k^{th} row vector and u^{th} column vector of $\mathbf{D}$, respectively. As shown in the Appendix, the $(k,u)^{\text{th}}$ element of $\mathbf{D}$ is given by

$$d_u(k) = \frac{1}{\left|1 + \sum_{i=1}^{P'} a_u(i) \exp(-\frac{2\pi j i k}{N})\right|^2} \tag{D.14}$$

which is the k^{th} bin of the u^{th} gain normalized PSD. The U different sets of AR parameters $\{a_u(i)\}_{i=1}^{P'}$ are obtained from a training stage which is detailed in the next section. The matrix $\mathbf{D}$ can be partitioned as $\mathbf{D} = \begin{bmatrix}\mathbf{D}_s & \mathbf{D}_c\end{bmatrix}$, where $\mathbf{D}_s$ of size $N \times U_s$ contains only the U_s signal spectral envelopes, and $\mathbf{D}_c$ of size $N \times U_c$ contains only the U_c noise spectral envelopes. The k^{th} row of $\mathbf{D}$ can be partitioned similarly, and we write this as $\widetilde{\mathbf{d}}_k = \begin{bmatrix}\widetilde{\mathbf{d}}_{s,k} & \widetilde{\mathbf{d}}_{c,k}\end{bmatrix}$.

The AR parameters describing the spectral shapes contained in $\mathbf{D}$ are obtained offline. Thus, only the activation coefficients in σ have to be estimated online which we do by maximizing the likelihood in (D.10) w.r.t. σ, i.e.,

$$\hat{\boldsymbol{\sigma}} = \arg\max_{\boldsymbol{\sigma} \geq 0} p(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{D}) = \arg\max_{\boldsymbol{\sigma} \geq 0} \mathcal{N}\left(0, \sum_{u=1}^{U} \sigma_u^2 \mathbf{R}_u(\mathbf{a}_u)\right) . \tag{D.15}$$

As shown in the Appendix, the log-likelihood function can be expanded as

$$\ln p(\mathbf{x}|\boldsymbol{\sigma},\mathbf{D}) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}D_{\text{IS}}\left(\boldsymbol{\Phi}|\mathbf{D}\boldsymbol{\sigma}\right) - \frac{1}{2}\sum_{k=0}^{N-1}\ln\Phi(k) + \frac{N}{2} \tag{D.16}$$

where we have defined

$$\Phi(k) = \frac{1}{N}\left|X(k)\right|^2 = \frac{1}{N}\left|\sum_{n=0}^{N-1}x(n)\exp\left(-j2\pi\frac{nk}{N}\right)\right|^2 \tag{D.17}$$

$$\boldsymbol{\Phi} = \begin{bmatrix}\Phi(0) & \cdots & \Phi(N-1)\end{bmatrix}^T \tag{D.18}$$

$$D_{\text{IS}}(\boldsymbol{\psi}_1|\boldsymbol{\psi}_2) = \frac{1}{N}\sum_{k=0}^{N-1}\left(\frac{\psi_1(k)}{\psi_2(k)} - \ln\frac{\psi_1(k)}{\psi_2(k)} - 1\right)\,. \tag{D.19}$$

The function $D_{\text{IS}}(\boldsymbol{\psi}_1|\boldsymbol{\psi}_2)$ is the Itakura-Saito (IS) distortion measure between two discrete spectra $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ [45]. As $\sum_{k=0}^{N-1}\ln\Phi(k)$ does not depend on $\boldsymbol{\sigma}$, maximising the likelihood with respect to $\boldsymbol{\sigma}$ simply corresponds to minimizing the IS distance between the periodogram $\boldsymbol{\Phi}$ and the modelled PSD $\mathbf{D}\boldsymbol{\sigma}$, under the constraint that $\Phi(k) \geq 0\ \forall k$. That is, the ML estimate of $\boldsymbol{\sigma}$ is obtained by solving the supervised non-negative matrix factorisation (NMF) problem

$$\hat{\boldsymbol{\sigma}} = \begin{bmatrix}\hat{\boldsymbol{\sigma}_s}^T & \hat{\boldsymbol{\sigma}_c}^T\end{bmatrix}^T = \underset{\boldsymbol{\sigma}\geq 0}{\arg\min}\, D_{\text{IS}}\left(\boldsymbol{\Phi}|\mathbf{D}\boldsymbol{\sigma}\right). \tag{D.20}$$

We remark here that unlike in [46], the matrix $\mathbf{D}$ is parametrised by pre-trained AR-envelopes for which reason the problem here is referred to as parametric NMF [34]. We also remark that the optimization problem can easily be extended to the case where $V > 1$ segments are available which is the case in, e.g., offline processing. The modelled PSD can be written as the matrix product $\mathbf{D}\boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma}$ of dimension $U \times V$ is the activation matrix containing the activation coefficients of a segment as a column vector, i.e., $\boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\sigma}(1) & \cdots & \boldsymbol{\sigma}(V)\end{bmatrix}$.

Focusing on estimating $\boldsymbol{\sigma}$ for a single segment from (D.20), it is well-known in the NMF-literature that (D.20) cannot be solved analytically. Instead, the multiplicative gradient descent (MU) [47, 48] is typically used to iteratively approach the solution. Specifically, the value of the variable of interest at the $(i+1)^{\text{th}}$ iteration is updated by multiplying its value at the previous i^{th} iteration by the ratio of the negative part to the positive part of the gradient of the criterion with respect to this variable, namely $\theta^{(i+1)} \leftarrow \theta^{(i)}\frac{[\nabla f(\theta)]_-}{[\nabla f(\theta)]_+}$, where θ is the variable of interest and the gradient is decomposed as $\nabla f(\theta) = [\nabla f(\theta)]_+ - [\nabla f(\theta)]_-$. Taking the derivative of $D_{\text{IS}}\left(\boldsymbol{\Phi}|\mathbf{D}\boldsymbol{\sigma}\right)$ with respect to $\boldsymbol{\sigma}$ leads to

$$\frac{\partial D_{\text{IS}}\left(\boldsymbol{\Phi}|\mathbf{D}\boldsymbol{\sigma}\right)}{\partial\boldsymbol{\sigma}} = \mathbf{D}^T\left[(\mathbf{D}\boldsymbol{\sigma})^{[-2]}\odot(\mathbf{D}\boldsymbol{\sigma}-\boldsymbol{\Phi})\right] = \mathbf{D}^T(\mathbf{D}\boldsymbol{\sigma})^{[-1]} - \mathbf{D}^T(\mathbf{D}\boldsymbol{\sigma})^{[-2]}\odot\boldsymbol{\Phi} \tag{D.21}$$

which leads to that $\boldsymbol{\alpha}$ is computed iteratively from

$$\hat{\boldsymbol{\sigma}}^{(i+1)} \leftarrow \hat{\boldsymbol{\sigma}}^{(i)} \odot \frac{\mathbf{D}^T \left(\mathbf{D}\hat{\boldsymbol{\sigma}}^{(i)}\right)^{[-2]} \odot \boldsymbol{\Phi}}{\mathbf{D}^T \left(\mathbf{D}\hat{\boldsymbol{\sigma}}^{(i)}\right)^{[-1]}}, \tag{D.22}$$

where $\odot$ denotes element-wise multiplication. The division and exponentiation are also element-wise, and the number of iterations is denoted as I.

Having computed $\hat{\boldsymbol{\sigma}} = \begin{bmatrix} \hat{\boldsymbol{\sigma}_s}^T & \hat{\boldsymbol{\sigma}_c}^T \end{bmatrix}^T$, we can compute an SNR estimate as

$$\xi(k) = \frac{\lambda_s^2(k)}{\lambda_c^2(k)} \tag{D.23}$$

where

$$\lambda_s^2(k) = \widetilde{\mathbf{d}}_{s,k}\hat{\boldsymbol{\sigma}}_s \tag{D.24}$$

$$\lambda_c^2(k) = \widetilde{\mathbf{d}}_{c,k}\hat{\boldsymbol{\sigma}}_c\ . \tag{D.25}$$

In the noise PSD estimation literature, these quantities are often referred to as the *a priori* SNR, the prior speech PSD, and the prior noise PSD, respectively. Given values for these quantities, it can be shown that the MMSE estimator of the noise PSD is [31]

$$\Phi_c(k) = \left(\frac{1}{1+\xi(k)}\right)\Phi(k) + \left(\frac{\xi(k)}{1+\xi(k)}\right)\lambda_c^2(k), \tag{D.26}$$

for $k = 0, ..., N-1$. The differences between the estimate in (D.26) and the MMSE-based estimate in [31] are how the a priori SNR and the prior noise PSD are computed. While we here obtain values for these via the parametric NMF method, the approach in [31] relies on speech presence probabilities (SPPs).

To add robustness to cases where the observed noise samples are not well-represented by the pre-trained spectral envelopes in $\mathbf{D}$, it can be augmented with a single time-varying entry corresponding to the normalized AR spectral envelope that is fitted to the MMSE-SPP [31] noise PSD based pre-whitener $\{w_{\text{MMSE-SPP}}(i)\}_{i=1}^{P'}$, in which each frequency-bin entry is given by

$$d_{\text{MMSE-SPP}}(k) = \frac{1}{\left|1 + \sum_{i=1}^{P'} w_{\text{MMSE-SPP}}(i) \exp\left(-\frac{2\pi j i k}{N}\right)\right|^2}. \tag{D.27}$$

A summary on how the pre-whitening filter is updated for a single segment is outlined in Alg. 1. Note that the computational complexity of each step is given using big $\mathcal{O}$ notation. A block diagram of the pre-whitening

Algorithm 1 Proposed Pre-whitening for a single segment, based on parametric NMF noise PSD estimate, assuming U_s signal and U_c noise spectral envelopes whose columns given by (D.14) are contained on **D**.

1: Obtain $\Phi(k),\ k = 0, ..., N-1$ from (D.17) $\triangleright \mathcal{O}(N \log N)$
2: Estimate MMSE-SPP noise PSD [31] and fit it to an AR spectrum using (D.5) and (D.4). Augment **D** with envelope whose elements are given by (D.27) $\triangleright \mathcal{O}[(N+P)\log N] + \mathcal{O}(P^2)$
3: Initialize $\hat{\sigma}^{(i)}$ with random positive numbers $\triangleright \mathcal{O}(1)$
4: **for** `i=1:I` **do** $\triangleright \mathcal{O}(UNI)$
5: Compute $\hat{\sigma}^{(i)}$ using (D.22)
6: **end for**
7: Compute $\lambda_s^2(k)$ and $\lambda_c^2(k)$, for $k = 0, ..., N-1$, from (D.24) and (D.25) $\triangleright \mathcal{O}(UN)$
8: Obtain $\zeta(k)$ from (D.23), for $k = 0, ..., N-1$ $\triangleright \mathcal{O}(N)$
9: Estimate $\Phi_c(k), k = 0, ..., N-1$ from (D.26) $\triangleright \mathcal{O}(N)$
10: Fit noise PSD to AR spectrum of order P via (D.5) and (D.4). The pre-whitening filter is $W(\omega) = 1 + \sum_{i=1}^{P} w_c(i) e^{-j\omega i}$ $\triangleright\ \mathcal{O}(P \log N) + \mathcal{O}(P^2)$

method based on parametric NMF is shown in Fig. (D.2). The proposed pre-whitening method has a time complexity of $\mathcal{O}[(N+P)\log N] + O(P^2) + \mathcal{O}(NUI)$, while pre-whitening based on MMSE-SPP and MS has simply an order of $\mathcal{O}[(N+P)\log N] + \mathcal{O}(P^2)$.

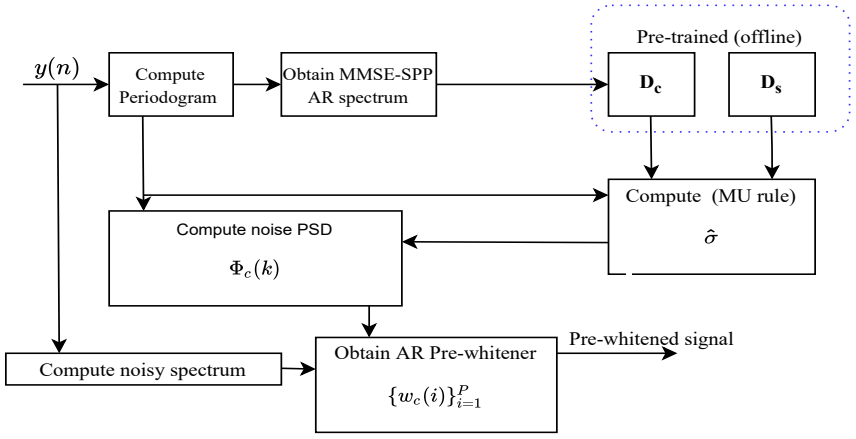

Fig. D.2: The block diagram of a parametric NMF-based AR pre-whitener.

5 Experimental setup and results

In this section, we present an extensive performance evaluation of the proposed pre-whitener on real signals under different colored noise scenarios. Except for the last experiment, which is concerned with time-of-arrival (TOA) estimation, we focus on speech processing problems. Specifically, the results of the following experiments are presented.

1. We seek to answer if whitening the noise using a pre-whitener is preferable to removing the noise using a speech enhancement (or noise reduction) algorithm [49]. Specifically, we evaluated the accuracy of the nonlinear least squares (NLS) pitch estimator [6, 7], which is optimal under a WGN assumption, when its input speech signal has either been pre-whitened or enhanced. The comparison also included the baseline approach where no pre-processing is performed.

2. We demonstrate that the proposed pre-whitener outperforms other pre-whiteners in terms of whiteness and spectral distance to an oracle pre-whitener. The oracle pre-whitener is the pre-whitener obtained from the AR parameters computed directly from the noise signal.

3. We investigate how the pre-whitening performance depends on the AR-order and the number of spectral shapes of the pre-trained dictionaries.

4. We aimed to verify that a better estimation accuracy of the NLS pitch estimator could be obtained when the signal was pre-processed with the proposed pre-whitener, especially in non-stationary noise conditions. The comparison also included the case in which a fixed (i.e., non-adaptive) pre-whitener is applied. Moreover, for a fairer comparison to typical non-parametric pitch estimators (e.g., RAPT), we then conducted an experiment in which we applied either speech enhancement or pre-whitening before the pitch was estimated with those classical approaches, thus allowing us to determine whether there is a greater benefit with certain types of pre-processing. The computational complexity of the different pre-processing approaches was also evaluated.

5. We applied a last stage of post-processing in order to contrast the performance to individual pitch estimators.

6. Finally, the last experiment dealt with TOA estimation, and it was assessed how much the proposed pre-whitener improved the estimation accuracy.

5.1 Codebook training

As we have alluded to earlier, the matrix $\mathbf{D}$ in (D.12), containing spectral envelopes of typical speech and noise segments, must first be obtained via a training step. The spectral envelopes are determined from autoregressive parameters which were obtained by using a standard vector quantization technique of speech coding. Specifically, the generalized Lloyd algorithm [50] was used to obtain cluster centers of line spectral frequency (LSF) coefficients of order $P' = 12$ computed from a large number of windowed data segments. The LSF parametrization was used in the clustering to ensure that the cluster centers corresponded to stable AR-processes. The obtained cluster centers were converted into AR parameters of order $P' = 12$. The collection of cluster centers converted into AR-parameters are often referred to as a codebook [41, 42]. A speech codebook of U_s entries was obtained from training on 54 minutes of sentences uttered by four speakers (two male and two female) from the CMU Arctic database [51], which were re-sampled from 16 to 8 kHz. We note that another database was used in the evaluation. Similarly, a noise codebook of U_c entries was obtained from training samples from the NOISEX-92 database [52], and we used the noise types babble, factory, F-16 and street, all resampled to 8 kHz. The duration of segments for the training was 32 ms, with an overlap of 50 %. The codebook sizes U_s and U_c are intentionally kept as variables as we evaluate the pre-whitening performance for different codebook sizes.

5.2 Performance measures

To compare the different pre-whiteners, both the spectral flatness measure (SFM) and the IS distortion are used. The SFM is defined as the ratio between the geometric mean and the arithmetic mean of a PSD [53], i.e., as

$$\text{SFM} = \frac{\sqrt[N]{\prod_{k=0}^{N-1} \Phi_{c,w}(k)}}{\frac{1}{N}\sum_{k=0}^{N-1} \Phi_{c,w}(k)}. \tag{D.28}$$

where $\mathbf{\Phi}_{c,w}$ is the noise PSD of the pre-whitened noise. The SFM indicates how correlated the noise samples are and, therefore, the degree of coloring. A value of 0 means that the noise is very correlated (colored), whereas an SFM of 1 means that the noise is perfectly white (the samples are perfectly uncorrelated). Therefore, an important goal of a pre-whitener is to increase the SFM, and we can also use the SFM to quantify the performance of a pre-whitener. Another approach to quantifying pre-whitening performance is to measure a spectral distance between a pre-whitener and the oracle pre-whitener. We here measure this spectral distance using the IS distortion defined in (D.19). Note that both the SFM and IS distortion are computed on a segment-by-segment basis and averaged over the test set.

While the SFM and IS distortion can be used to evaluate a pre-whitener directly, we can also evaluate it indirectly by measuring the performance improvement of the estimator cascaded with a pre-whitener. For pitch estimation, typical performance measures are [54]:

- Gross Error Rate (*GER*): GER is defined as

$$GER = \frac{N_g}{N_{VV}} \times 100\ \% \tag{D.29}$$

  where N_{VV} is the number of voiced segments and N_g is the number of voiced segments in which the magnitude of the relative difference between the estimate and the ground truth is greater than a threshold. Here, we used a relative threshold of 20 %. Note that only segments which a correctly classified as being voiced are included in the GER.

- Voicing Detection Error (*VDE*): VDE is defined as

$$VDE = \frac{N_{VU} + N_{UV}}{N} \times 100\ \% \tag{D.30}$$

  where N_{VU}, N_{UV}, and N are the number of segments misclassified as voiced, the number of segments misclassified as not-voiced (i.e., as unvoiced or pauses), and the total number of segments, respectively.

- Full Frame Error (*FFE*) [54]: FFE is defined as

$$FFE = \frac{N_{VU} + N_{UV} + N_g}{N} \times 100\ \% \tag{D.31}$$

  where the different quantities are the same as in the GER and VDE. Note that FFE is a composite metric which is simply the sum of the VDE and the FFE when all segments are voiced and correctly classfied as being voiced (i.e., when $N = N_{VV}$).

5.3 Experimental results with the Keele speech database

The set of experiments in this subsection was conducted on the Keele database [55], which consists of speech recordings of around 40 seconds from five male and five female speakers. The signals were resampled from 20 kHz to 8 kHz. Pitch estimates[1] extracted from laryngograph measurements segmented into 26.5 ms frames with 16.5 ms overlap are available in the database, and we treat these as being the ground truth estimates. We used the same segment length and overlap for the pitch estimators. Note that we in the evaluation have ignored segments for which

[1]The pitch estimates were computed using the RAPT [32] pitch estimation method and manually checked afterwards.

the ground truth estimate has been labelled unreliable. These segments represents only approximately 3 % of the total number of segments.

Different noise types such as babble, factory, F-16 and street noise from the NOISEX-92 database [52] were added at different values of iSNR. The iSNR indicates the power level of the clean speech signal relative to the noise power component, i.e.,

$$\text{iSNR} = \frac{\sigma_s^2}{\sigma_c^2}, \tag{D.32}$$

where σ_s^2 is the variance of the speech signal, and σ_c^2 is the noise signal variance. The samples used for the testing were different from those used in the training of the noise codebooks. In addition, samples of restaurant noise from the Aurora database [56], already sampled at 8 kHz, were used in the evaluation to assess the robustness against new encountered noise types.

Comparison to no pre-processing and to speech enhancement

First, the accuracy of a WGN-based method, namely the nonlinear least squares (NLS) pitch ω_0 estimator [6, 7], was assessed for the cases where the input signal to the estimator is either pre-whitened, enhanced using a speech enhancement method (an approach suggested in [49]), or unprocessed. Fig. (D.3) illustrates the case where a pre-whitener is used as a pre-processor. Note that a final post-processing step can be used to refine the initially obtained parameter estimates. Such post-processing step is ignored in all the subsections, except the last one, as we want first to verify that a pre-whitener applied as a pre-processor will result in a better accuracy of the pitch estimator. The NLS estimator of ω_0 [6, 29] corresponds to

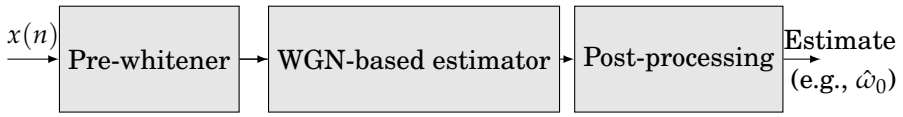

Fig. D.3: Structure diagram for obtaining estimates on colored noise scenarios based on a WGN method.

the ML estimator under the WGN assumption and is given by

$$\hat{\omega}_0 = \arg\max_{\omega_0} \ \underline{\mathbf{x}}^T \mathbf{Z}_L(\omega_0) \left[\mathbf{Z}_L^H(\omega_0)\mathbf{Z}_L(\omega_0)\right]^{-1} \mathbf{Z}_L^H(\omega_0)\underline{\mathbf{x}}, \tag{D.33}$$

where $\mathbf{Z}_L(\omega_0) = [\mathbf{z}(\omega_0)\ \mathbf{z}^*(\omega_0)\ ...\ \mathbf{z}^*(\omega_0 L)]$ is a Fourier matrix constructed from $2L$ complex exponential vectors $\mathbf{z}(\omega_0 l) = [1\ e^{j\omega_0 l}\ ...\ e^{j\omega_0 l(N-1)}]^T$. Here, $\underline{\mathbf{x}}$ is the vector used in the estimation, either the vector of noisy speech (i.e.,

discarding the pre-processing block), or enhanced speech or pre-whitened speech. An important feature of this problem is that it jointly estimates ω_0 and the number of real sinusoids L. The frequency of each sinusoid is an integer multiple of the fundamental ω_0, in contrast to the case of independent sinusoids, which are not harmonically related [1]. To obtain L, some Bayesian model comparison methods (e.g., based on maximum a posteriori) [57] can be used to find the most likely model order, after estimates of ω_0 have been obtained for all candidate model orders. The model comparison is the key in reducing the sub-harmonic error problems, such as doublings or halvings. Such model comparison also includes the case $L = 0$, i.e., it is possible to do voicing detection. In all the experiments related to pitch estimation, the pitch range was [60,400] Hz, and a maximum model order of $L = 30$ harmonics is set for the model comparison.

To ensure that the differences in noise PSD estimates do not influence the result, both the applied AR pre-whitener and the speech enhancement method were based on the introduced parametric NMF (Par-NMF) noise PSD estimate. In the first experiment, $U_s = 32$ and $U_c = 16$ pre-trained spectral shapes were used, whereas we assessed the performance as a function of the number of entries in the second experiment. The pre-whitening order was set to $P = 30$, and the pre-whitening filter coefficients were updated on segments of length 32 ms, with a time shift of 16 ms between them. In the MU rule, $I = 40$ iterations were used. The pre-processing based on speech enhancement was performed with the optimally modified LSA (OM-LSA) speech estimator [58]. The iSNR was varied from -5 to 10 dB, and three Monte-Carlo simulations (MCS) were run for each noise type at each iSNR for each file from the Keele database. The performance measures in (D.29)-(D.31) were computed and are depicted in Fig. F.2 with 95 % confidence intervals. The average number of pitch errors was very high when $\underline{x}$ is the unprocessed input signal, even in high SNR conditions. When $\underline{x}$ is produced by the speech enhancement method, the pitch estimation accuracy improved considerably in most cases compared to when the input signal was unprocessed. When $\underline{x}$ is the pre-whitened input signal, this gives the overall best performance, as noted from the non-overlapping confidence intervals between speech enhancement and pre-whitening. This is more evident at lower iSNRs, but still a considerable gap is seen at high SNRs, specially for non-stationary noise types, such as babble and restaurant noise.

Comparison of AR pre-whiteners

We investigated the pre-whitening performance of AR pre-whiteners based on three noise PSD estimates: MS [30], MMSE-SPP [31], and the proposed Par-NMF based approach. Since the codebooks were trained on segments

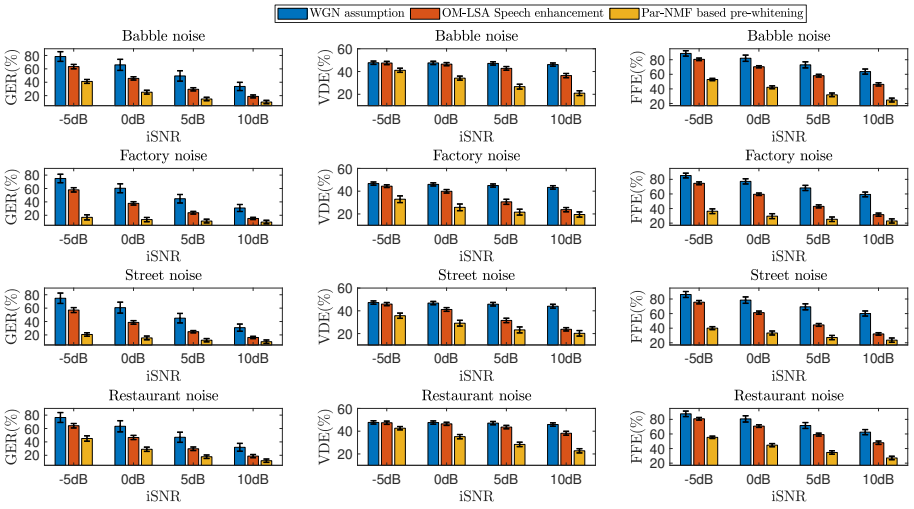

Fig. D.4: The gross pitch errors, voicing detection errors and full frame errors of the estimated pitch computed by the NLS pitch estimator, for different iSNRs, with different colored noise types, after assuming that the noise is WGN, applying speech enhancement, or applying pre-whitening.

of 32 ms, overlapped by 50 %, the same segment length and overlapping percentage were used in the different pre-whiteners. The SFM of the pre-whitened noise in (D.28) and the IS distortion between the frequency responses of the oracle and the estimated pre-whiteners were evaluated.

First, we studied the performance as a function of the AR-order P. The iSNR used in this setup was 0 dB. The SFM results include four curves, as the performance from applying the oracle pre-whitener is also included while the IS distortion plots only involve three curves, as comparing the response of the oracle pre-whitener to itself leads to an ISD of 0. $U_s = 32$ and $U_c = 16$ pre-trained spectral shapes were used. Before pre-whitening, the average SFM values at all the iSNRs were: babble noise (0.065), factory noise (0.045), restaurant noise (0.129), and F-16 noise (0.115). The results are depicted in Fig. F.3. For the proposed Par-NMF pased pre-whitener, the SFM increased as a function of P, but at the same time, the ISD between the oracle and the estimated pre-whitener increased. Thus, even if the noise gets closer to being white by increasing P, the spectral response of the pre-whitener becomes more different from the oracle pre-whitener response, given that fitting the estimated noise PSD to an AR spectrum of a higher order P is more prone to overfitting. With a lower P, the noise PSD can be more easily fitted to a much smoother spectral shape, which may lack some detail. The performance of pre-whitening based on MMSE [31] is better than the one based on MS [30]. The SFM from Par-NMF based pre-

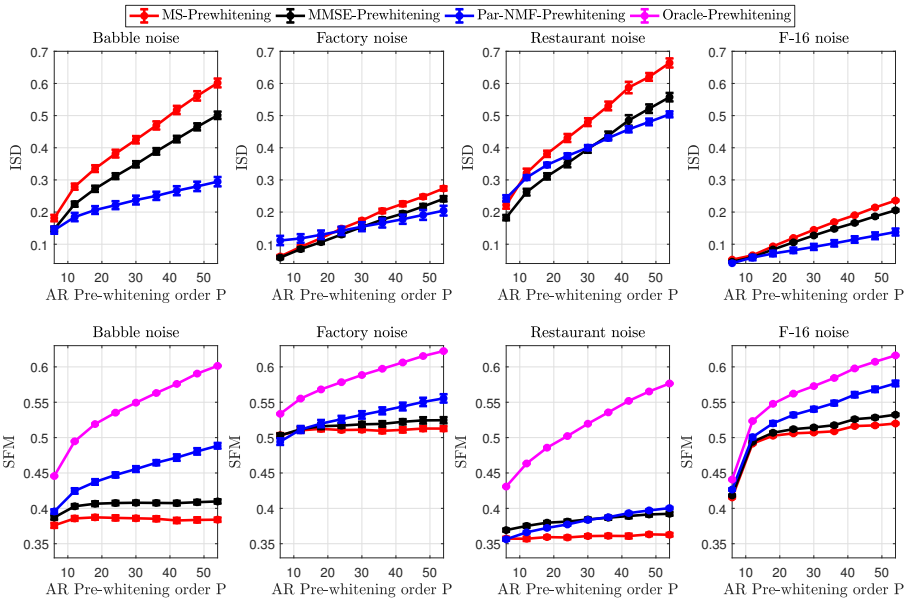

Fig. D.5: The IS distortion between the oracle and estimated pre-whiteners and the spectral flatness measure of the pre-whitened noise as a function of the AR pre-whitening order, at iSNR = 0 dB, under different colored noise scenarios. A lower IS distortion is preferred, and a higher SFM is desirable. Results are reported in 95% confidence intervals.

whitening was always higher than that based on MS and MMSE for both babble and F-16 noise. An important observation is that increasing P did not appear to significantly improve the noise whiteness by pre-whitening based on MS or MMSE. For factory noise, the Par-NMF pre-whitener can achieve better performance when P is not too low. For the restaurant noise scenario, with a $P > 30$, the Par-NMF based pre-whitener had lower ISD than the MMSE based pre-whitener. To balance the noise whiteness and the accuracy of the pre-whitener response, we found using a value of P in the interval [30,40] to be convenient. We used $P = 36$ in the next experiments.

Next, we evaluated how the performance of the Par-NMF pre-whitener depends on how many speech and noise entries are used in $\mathbf{D}$. The performance was evaluated for different combinations of noise and speech AR dictionaries sizes where the speech AR dictionary $\mathbf{D}_s$ could have 2^{b_s} spectral shapes for $b_s \in \{5,6,7,8\}$ and the noise AR dictionary $\mathbf{D}_c$ could have 2^{b_c} spectral shapes for$b_c \in \{3,4,5,6,7,8,10\}$. Again, the iSNR was fixed at 0 dB. The results are displayed in Fig. D.6. Using 32 speech spectral envelopes gives the best results, and increasing U_s degrades the performance due to overfitting, as it was also seen in [59] in a speech enhancement framework. Using $U_s = 32$ with $U_c \geq 128$ spectral shapes lead to the best performance

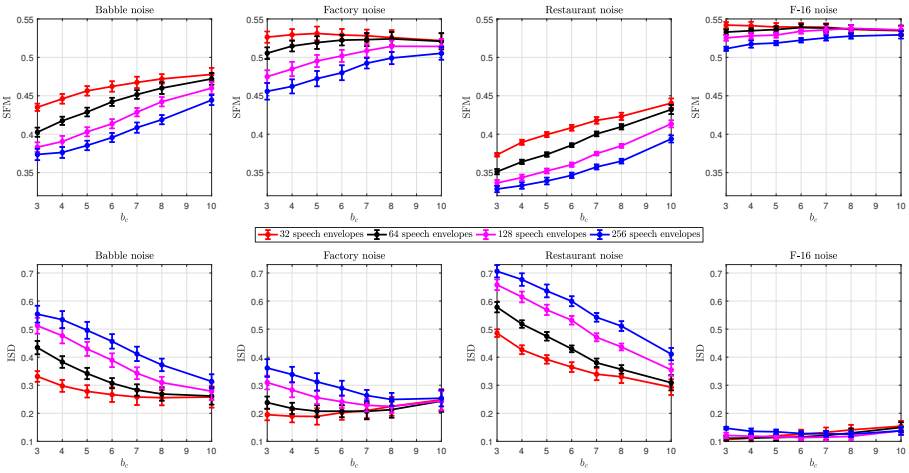

Fig. D.6: The spectral flatness measure of the pre-whitened noise and the IS distortion between the oracle and estimated Par-NMF based pre-whitener as a function of b_c, for different number 2^{b_s} speech spectral envelopes, at iSNR = 0 dB, under different colored noise scenarios. A lower IS distortion is preferred, and a higher SFM is desirable.

for both babble and restaurant noise, although increasing from 256 to 1024 noise spectral envelopes, did not decrease the ISD significantly, as the confidence intervals overlapped. Using $U_c = 16$ entries seemed to be enough for factory and F-16 noise types, although using a higher number of entries did not degrade the performance too much. We therefore used both combinations $U_s = 32, U_c = 16$ and $U_s = 32, U_c = 256$ in the next experiments.

We then conducted an evaluation of the pre-whitening performance as a function of the iSNR, and the results are depicted in Fig. D.7. For babble noise, Par-NMF based pre-whitening had the best performance, regardless how many U_c entries were used, although with $U_c = 16$, a considerable lower ISD was observed at higher iSNRs. However, $U_c = 256$ allowed for a slightly better SFM. For restaurant noise, a similar SFM was achieved for Par-NMF based pre-whitening using $U_c = 16$ entries and MMSE pre-whitening, with a slightly lower ISD for the Par-NMF. For this noise type, not included in the training step, $U_c = 256$ entries lead to a much better performance. For factory and F-16 noise, using a lower number of U_c entries is more convenient. In these cases, the benefit of Par-NMF pre-whitening (with $U_c = 16$) was seen at lower iSNRs, because at higher ones, the ISD from MMSE or MS based pre-whitening became lower, although the noise whiteness from the three approaches were similar. Again, in most cases, MS based pre-whitening was outperformed by either MMSE-SPP or Par-NMF based pre-whitening.

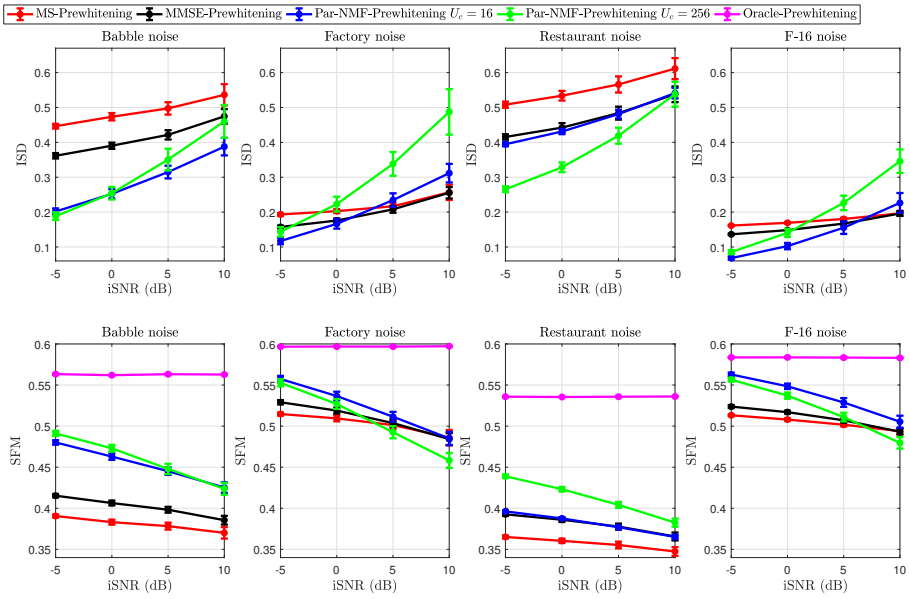

Fig. D.7: The IS distortion between the oracle and estimated pre-whiteners, and the spectral flatness measure of the pre-whitened noise as a function of the iSNR, for an AR pre-whitening order P=36. A lower IS distortion is preferred, and a higher SFM is desirable.

Evaluation of pitch estimation accuracy with pre-whitening

In the next experiment, we compared four different pre-whiteners by evaluating the performance improvement of the NLS pitch estimator when its input signal is the pre-whitened signal. The four pre-whiteners are based on noise PSD estimates obtained by MS, MMSE, the proposed Par-NMF (with both $U_c = 16$ and $U_c = 256$ entries), and a fixed noise PSD computed from the long-term averaged spectrum of the samples of the noise of interest used in the codebooks training. That is, a fixed pre-whitening filter is applied to verify that an adaptive pre-whitener based on the local characteristics of speech and noise signals should be preferred as a pre-processor. Only in the restaurant noise case, which was not included in the training, the samples used for the testing were used to determine the long-term average spectrum. The post-processing block in Fig. D.3 is still not used. Babble, factory, street, and restaurant noise were added at different iSNRs from -5 to 10 dB, and three MCS were run for each file at each iSNR. The performance measures in (D.29)-(D.31) were computed after estimating the pitch. The results are depicted in Fig. D.8 with 95 % confidence intervals. Clearly, using a fixed pre-whitener resulted in poorer pitch estimates and voicing detections than using the time-varying pre-whiteners. For babble and restaurant noise, the

best accuracy of the NLS pitch estimator was achieved when the cascading was done with the Par-NMF based pre-whitener, because the confidence intervals of the FFE were clearly separated from those of MMSE-SPP or MS based pre-whitening. When using $U_c = 256$ entries, a slightly better performance is seen than when using $U_c = 16$ entries. For street noise, using $U_c = 256$ entries has a positive effect in reducing the GER, although it will not benefit the VDE. In terms of FFE, for factory and street noise, which are more stationary noise types, the accuracy after pre-whitening based on the three approaches is very similar, thus indicating that the proposed Par-NMF based pre-whitener is of greater benefit in non-stationary noise scenarios.

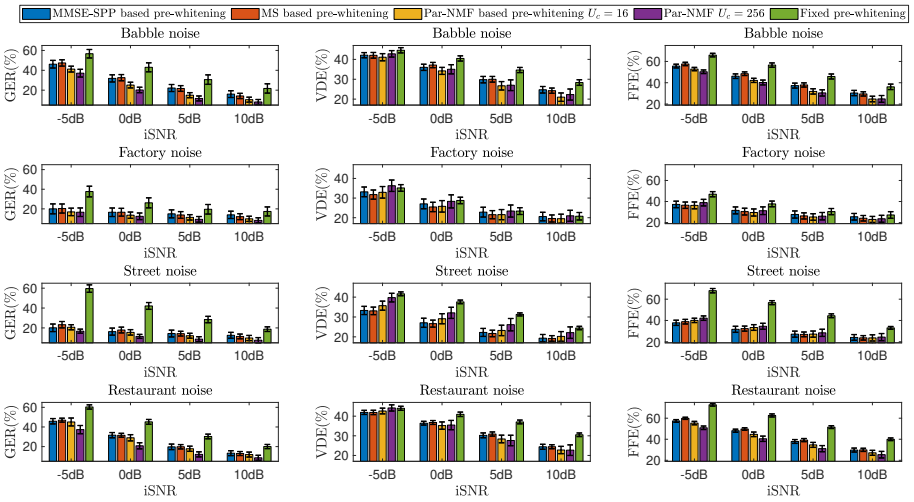

Fig. D.8: Estimation accuracy of the NLS pitch estimator under different noise conditions, after application of AR pre-whitening based on different noise PSD estimates, and also after applying a fixed pre-whitening filter.

We then investigated if various non-parametric pitch estimators, which are not derived under a WGN assumption, improved their accuracy by using either the proposed pre-whitener or an enhancement system as a pre-processor. Particularly, the Cepstrum-based method [60], RAPT [32], SHRP [61], and SWIPE' [33], all of them with a final smoothing step, were used in the evaluation. To determine which is the best pre-processing method on average, we present the averaged performance from those four estimators in three different ways: in its naïve (out-of-the-box) form (i.e., without a pre-processor), and when either an OMLSA based enhancement system or the proposed Par-NMF pre-whitener were used as a pre-processor. The performance of the recently introduced robust Bayesian pitch tracker [62], also derived under a WGN assumption, was also evalu-

ated. This method models the dynamic evolution of the pitch, the number of harmonics, and the voicing state by using first-order Markov processes. The already introduced NLS pitch estimates were again included, but a Ney smoothing step between consecutive independent-segment values [63] was applied, to be fairer in the comparison, as all the other methods had tracking or refinement capabilities. The NLS and the Bayesian pitch tracker estimates were evaluated only when the Par-NMF pre-whitener was used as a pre-processor, as we verified in subsection 5.3 that the performance of the NLS estimator, has a better improvement from pre-whitening. An example of the estimates produced by the Bayesian pitch tracker for a female speaker in babble noise at an iSNR of 3 dB is shown in Fig. D.9. The pitch was estimated after either OMLSA-based enhancement or the Par-NMF based pre-whitening were used as pre-processors. Clearly, the resulting estimates after enhancement showed a large number of not-voiced (e.g., silent) segments wrongly detected as voiced, and also a high number of octave errors. When pre-whitening is instead applied, the pitch contour is better captured as less octave errors and less voicing detection errors are obtained.

The overall performance is assessed by adding either babble or factory noise at iSNRs from -5 to 11 dB. Three MCS were run for each file from the Keele database at each iSNR. The results are depicted in Fig. D.10. The best performance was achieved by cascading the pre-whitener with the Bayesian pitch tracker, although in some cases under factory noise conditions, the confidence intervals overlapped with the NLS pitch estimates. The performance in the babble noise case was worse than the factory noise case, which is expected due to the fact that babble noise is a random mixture of human speech signals, making more challenging the pitch estimation task. On average, under babble noise conditions, the GER from non-parametric estimators was improved by pre-processing via pre-whitening, and in the factory noise, also from pre-whitening based on $N_c = 16$ spectral shapes, for iSNRs below 7 dB. In contrast, the VDE was improved by applying enhancement below 7 dB for babble noise, and in the factory noise case, at iSNRs below 11 dB. The FFE slightly decreased when pre-whitening based on $U_c = 16$ entries was used, but only at iSNRs lower than 3 dB in babble noise. In the factory noise case, the full frame errors were reduced by applying enhancement at iSNRs below 7 dB. However, although the performance of non-parametric pitch methods was improved by either enhancing or pre-whitening, the best performance was achieved from pre-whitening followed by the Bayesian pitch tracking. Pre-whitening combined with NLS pitch estimation followed by nonlinear smoothing also resulted in less full frame errors than non-parametric pitch estimators (even if they obtained a benefit from a pre-processing step) for babble noise at iSNRs below 7 dB.

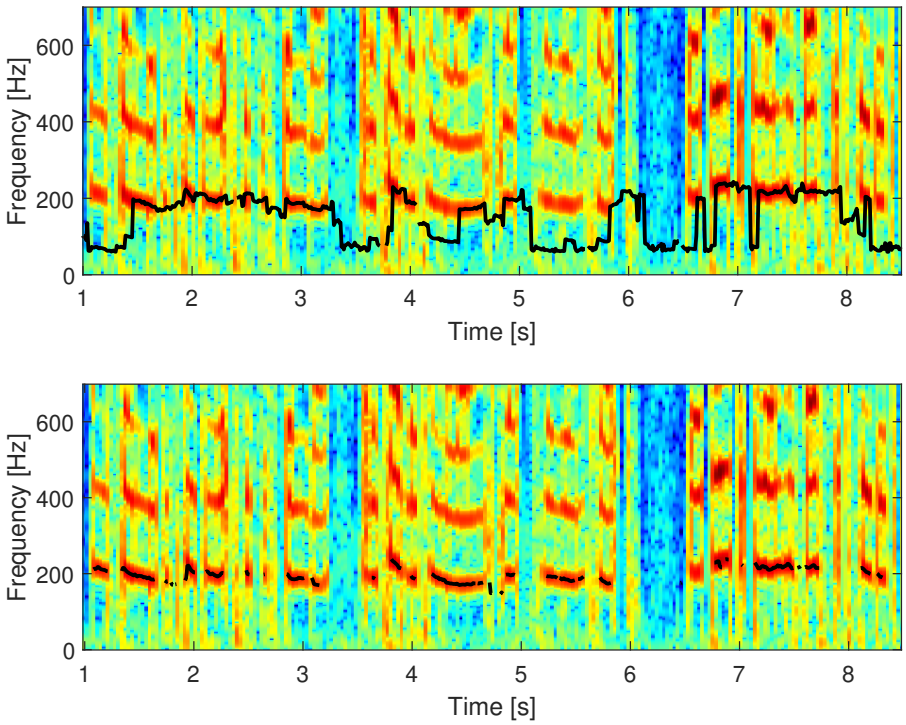

Fig. D.9: Pitch estimates from the Bayesian tracker for a female excerpt in 3 dB babble noise after either enhancement (top) or pre-whitening (bottom) is applied as pre-processor. Note that only the spectrograms of the clean signal are shown to facilitate an easier visual evaluation of the produced pitch estimates.

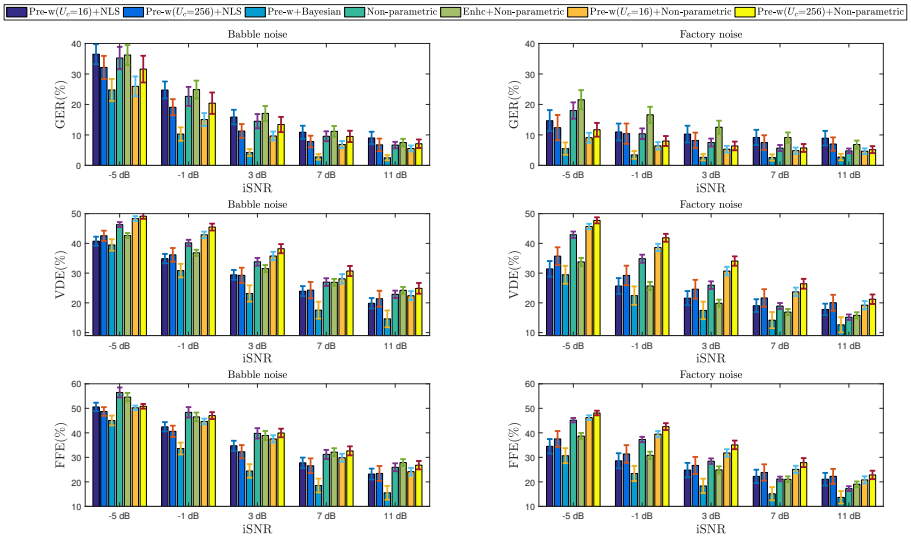

Fig. D.10: Estimation accuracy of parametric pitch estimators (NLS and Bayesian) after pre-whitening, and averaged performance of four non-parametric pitch estimators (Cepstrum, SWIPE, SHRP and RAPT) in their naïve states, and when either speech enhancement or pre-whitening was previously applied, under different noise conditions.

OMLSA	MS	MMSE	Par-NMF ($U_c = 16$)	Par-NMF ($U_c = 256$)
5.274	1.057	0.981	6.820	10.272

Table D.1: Computation time in [s] for different pre-processing schemes

Evaluation of computational complexity of pre-processors

We also evaluated the computation time of the various pre-processors, including OMLSA-based enhancement based on the parametric NMF noise PSD estimate. The testing was done with one excerpt of the Keele database with a duration of 40.3 seconds. The total time for each type of pre-processing (enhancement and pre-whitening based on different noise PSD estimates) is reported in Table D.1. The other approaches, specially the pre-whiteners based on MS or MMSE, are computationally faster than the pre-whitener based on parametric NMF. However, as seen from previous experiments, such an increase in computation time for the proposed pre-processing scheme resulted in an improvement in the accuracy of the WGN-based estimators, specially under non-stationary noise.

Evaluation of pitch estimation accuracy including post-processing

In the last evaluation of pitch estimation, we included the last post-processing block of Fig. D.3. As previously shown experimentally, using the Par-NMF pre-whitener as a pre-processor leads to the largest improvement of the accuracy of the NLS pitch estimator. However, there might still be some segments in which the solution produces estimates resulting in either a gross error or a voicing detection error. To further reduce these errors, post-processing of the initial pitch estimates is performed by iterating the following two steps [36]:

1. The harmonic amplitudes for a given estimated $\hat{\omega}_0$ and order $\hat{L}$ are $\hat{\mathbf{a}} = \left[\mathbf{Z}_{\hat{L}}^H(\hat{\omega}_0)\mathbf{Z}_{\hat{L}}(\hat{\omega}_0)\right]^{-1}\mathbf{Z}_{\hat{L}}^H(\hat{\omega}_0)\mathbf{x}$ [6], so the residual representing what is not captured by the harmonic model, including the stochastic parts of the speech, is $\hat{\mathbf{c}} = \mathbf{x} - \mathbf{Z}_{\hat{L}}^H(\hat{\omega}_0)\hat{\mathbf{a}}$. Thus, the AR parameters for an updated pre-whitener can be directly re-estimated from this residual using the autocorrelation method [27].

2. The re-estimated AR parameters of the residual are directly used as the coefficients of a new pre-whitening filter, which is applied to the noisy signal. From the new pre-whitened signal, the pitch ω_0 and L are again estimated with the NLS estimator (D.33).

In this iterative process, the new pre-whitener is no longer computed using the parametric NMF based noise PSD estimator, as it is now instead computed from the residual. As seen below, however, the key to achieve the final best pitch estimation accuracy is having applied a better pre-whitener as a pre-processor.

The full setup in Fig. D.3 was evaluated, and although this involves an even higher computational complexity than the one reported in Sec. 5.3, it also leads to an improved pitch estimation accuracy. Both pre-whiteners based on MMSE-SPP and on the proposed Par-NMF were applied as a pre-processor. For the Par-NMF one, $U_s = 32$ and either $U_c = 16$ or $U_c = 256$ spectral shapes were considered, for factory and babble, respectively. For the iterative estimation, the iteration was performed a maximum of 10 times. Moreover, if a frame was detected as being not-voiced (i.e., $\hat{L} = 0$), the estimation was stopped for that segment. We compared to the performance of individual non-parametric estimators SWIPE', PEFAC [64], SHRP, and RAPT which all include a final smoothing step between consecutive estimates. Their individual performance was also assessed after pre-processing the noisy signal, being pre-whitened in the babble noise case, and enhanced using OMLSA for factory noise, according to the averaged preferred pre-preprocessing that was noted previously. The FFE for babble and factory noise are depicted in Fig. D.11. By including the post-processing

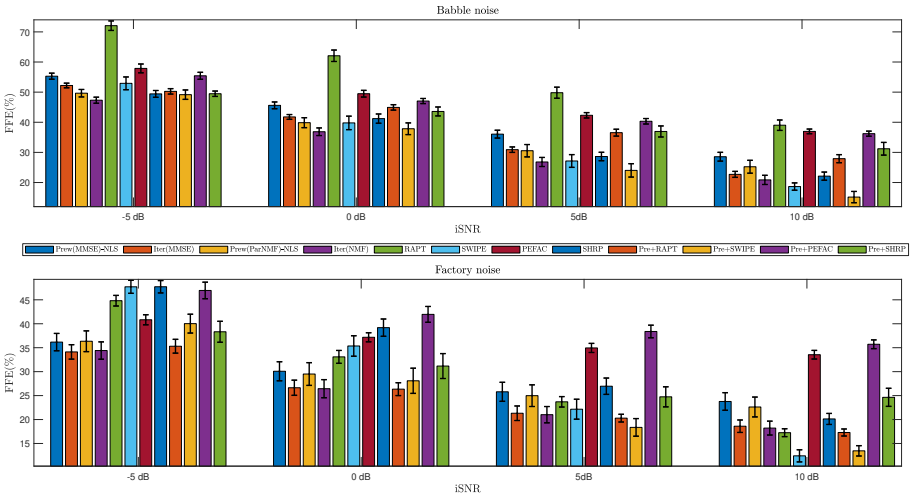

Fig. D.11: Estimation accuracy of NLS pitch estimation by considering both pre-processing and post-processing, and of non-parametric pitch estimators (PEFAC, SWIPE', SHRP, and RAPT) in their naïve states and with pre-processing, under different noise conditions.

step, there was a reduction between 2.2 and 4.5 % compared to the single cascade of pre-whitening and NLS pitch estimation. The independent consecutive pitch estimates were not smoothed in this case, but by doing so we expect that the performance could be improved.

It is seen that only pre-processing using parametric NMF combined with the NLS pitch estimates was not enough to have better accuracy than SWIPE' in babble noise. However, by post-processing the initially obtained estimates, a better performance was achieved below 5 dB, and similar one at 5 and 10 dB. That did not occur if the initial pre-whitener based on MMSE-SPP was applied, even if the post-processing based on iterative refinement was later applied. Also, SWIPE' improved its accuracy when its input signal was the pre-whitened signal, achieving similar or better accuracy than the proposed method at SNRs above 0 dB. Similarly, for factory noise, by applying the post-processing in the proposed method, RAPT was outperformed for SNRs below 10 dB and SWIPE' was outperformed at -5 and 0 dB. However, when the estimates are obtained from the pre-processed (enhanced) signal, RAPT was able to achieve a similar performance to the proposed method for al the SNRs. Also, when SWIPE' was applied to the enhanced signal, it achieved a similar performance at 0 and 5 dB, and a better one at 10 dB.

5.4 Experimental results regarding TOA estimation

Finally, we evaluated the accuracy of an estimator of the time-of-arrival (TOA) of a signal emitted by a source such as a loudspeaker and received by a receiver such as a microphone. For this application, we use a model in which the received signal is modeled as

$$x(n) = gs(n-\tau) + c(n) \tag{D.34}$$

where $s(n)$ is a known signal emitted by the source, g is the attenuation of the signal, and τ is the time it takes for the signal to propagate from the source to the receiver (i.e., the TOA). If the noise term $c(n)$ is assumed to be WGN, the ML estimator of τ and g are the solutions to

$$\{\hat{\tau}, \hat{g}\} = \min_{\tau \in T, g>0} \|x(n) - gs(n-\tau)\|_2^2, \tag{D.35}$$

over the possible set T of TOAs. If the analysis window is long relative to the size of $s(n)$, the TOA estimator can be accurately approximated as

$$\hat{\tau} = \arg\max_{\tau \in T} x(n)s(n-\tau) \tag{D.36}$$

which is often referred to as the matched filter [9]. In practical setups, the noise is likely to be non-white. In such cases, pre-whitening should be applied as a pre-processor, but we remark that it has to be applied to both $x(n)$ and $s(n)$ since applying the pre-whitener to only $x(n)$ would introduce an additional delay, resulting in a biased estimator.

We used the recorded signals from the SMARD database [65] at both the loudspeaker and the single microphone, both of them separated 3.13 m, with configuration number 0001. The known source signal was an artificial white noise synthetic signal, and the size of the burst was 3500 samples at a sampling frequency of 48 kHz. The rooms where the signals were recorded had a reverberation time of approximately 0.15 s. The colored noise was taken from the DREGON database [66]. Specifically, rotor noise from a drone running at 70 rounds per second was added to the signal picked up by the single microphone at different signal-to-noise ratios (SNR) before the TOA was estimated. 200 MCS were run at each SDNR. The rotor noise was resampled from 44.1 to 48 kHz to match the rate of the source signal. In the evaluation, we compared the performance of the matched filter with and without a pre-whitener. To pre-whiten the observation, a spectral basis matrix of four AR spectra shapes of the rotor noise was built by training a noise codebook on samples of the rotor noise. The training samples were different from those for testing. The testing samples were randomized at each MCS. An additional entry, corresponding to the known source signal, was also included as the clean signal spectral shape which

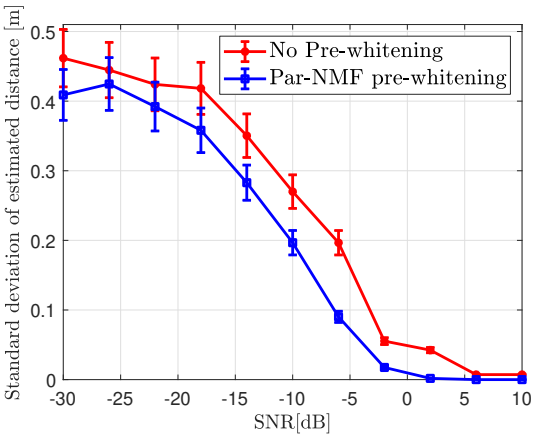

Fig. D.12: Estimation accuracy of single-channel TOA estimation with pre-whitening applied or not applied, versus the SDNR. Results are reported as 95% confidence intervals.

was simply a flat PSD. The training was performed with an order $P' = 35$ on segments with a duration of 20 ms with an overlap of 50 % between them. This order was chosen according to our observation that the best oracle performance was obtained with a higher order, as important envelope components that might be present at medium and high frequencies were not smoothed out. In the MU rule (D.22), the number of iterations was set to $I = 30$. From the estimated TOA, the distance between the loudspeaker and the microphone was obtained, and we computed the standard deviation of the measured distance at each SDNR. The results are shown in Fig. D.12. For SDNRs above -22 dB, pre-whitening resulted in a lower variance of the estimated distances than when pre-whitening was omitted. Below -22 dB, the confidence intervals from ignoring and performing pre-whitening overlapped.

6 Conclusion

The accuracy of statistical-based estimators based on the WGN assumption in real acoustic scenarios can be considerably improved when an AR pre-whitener is applied as a pre-processor of the noisy observation. In this paper, we introduced a time-varying pre-whitener which requires the activation coefficients of pre-trained spectral shapes in the parametric NMF method. Through numerous simulations, we have shown that using an AR pre-whitener based on the parametric NMF method results in a higher noise whiteness and a more similar spectral response to that of the oracle pre-whitener compared to conventional noise PSD estimators, specially in non-

stationary noise situations. Although the training stage of the spectral shapes may initially require additional effort compared to using traditional noise trackers, it offers a consistent way of including prior information about speech and noise types, resulting in a better performance of WGN-based estimators such as the NLS pitch estimator. Although well-known non-parametric pitch estimators can improve their accuracy from some pre-processing, the combination of pre-whitening with fast and efficient statistical-based WGN methods gives the best performance in terms of pitch errors and voicing detection errors, specially in scenarios of high noise levels (i.e., low SNRs). An additional improvement can be obtained by post-processing the resulting NLS pitch estimates, and this will result in a better overall accuracy than individual non-parametric pitch estimators, even if they are using a pre-processor, specially under low SNR conditions. This may require high computation time, but it allows to extract both the harmonic and autoregressive components of the speech signal, which is useful, e.g., in the speech decomposition problem [21]. The pre-whitener was also applied before a time-of-arrival estimation method formulated under the WGN assumption. In that case, the TOA estimation accuracy is improved by a pre-whitening step which relies on pre-trained shapes of the involved source signal and of the real noise in the recording environment, such as wind noise or drone ego-noise.

7 Appendix

To maximize the data likelihood $p(\mathbf{x}|\sigma,\mathbf{D}) \sim \mathcal{N}\left(\mathbf{0}, \sum_{u=1}^{U} \sigma_u^2 \mathbf{R}_u(\mathbf{a}_u)\right)$, we use the well-known fact [67] that $\mathbf{R}_u(\mathbf{a}_u)$ can be approximated as circulant and therefore diagonalized by the Fourier transform if N is much larger than the AR-order P'. Thus, the approximate diagonalisation of the covariance is

$$\mathbf{R}_u(\mathbf{a}_u) \approx \frac{1}{N}\mathbf{F}\mathbf{D}_u(\mathbf{a}_u)\mathbf{F}^H \tag{D.37}$$

where $\mathbf{F}$ is the DFT matrix whose entries are given by $[\mathbf{F}]_{n,l} = \exp\left(j2\pi nl/N\right),\ n,l = 0,1,...,N-1$, and

$$\mathbf{D}_u(\mathbf{a}_u) = \left(\mathbf{\Lambda}_u^H(\mathbf{a}_u)\mathbf{\Lambda}_u(\mathbf{a}_u)\right)^{-1},\ \ \mathbf{\Lambda}_u(\mathbf{a}_u) = \mathrm{diag}\left(\mathbf{F}^H\left[\mathbf{a}_u^T\ \mathbf{0}^T\right]^T\right). \tag{D.38}$$

The diagonal entries of $\mathbf{D}_u(\mathbf{a}_u)$ represent the eigenvalues of $\mathbf{R}_u(\mathbf{a}_u)$, which correspond to the normalized PSD of the u^{th} AR process.

Using the above definitions, the log-likelihood can be written as [34]

$$\ln p(\mathbf{x}|\boldsymbol{\sigma},\mathbf{D}) = -\frac{N}{2}\ln 2\pi - \frac{1}{2}\ln\left|\sum_{u=1}^{U}\frac{\sigma_u^2\mathbf{F}\mathbf{D}_u(\mathbf{a}_u)\mathbf{F}^H}{N}\right| - \frac{1}{2}\mathbf{x}^T\left[\sum_{u=1}^{U}\frac{\sigma_u^2\mathbf{F}\mathbf{D}_u(\mathbf{a}_u)\mathbf{F}^H}{N}\right]^{-1}\mathbf{x}, \quad \text{(D.39)}$$

which can be simplified as

$$\begin{aligned}\ln p(\mathbf{x}|\boldsymbol{\sigma},\mathbf{D}) = & -\frac{N}{2}\ln 2\pi - \frac{1}{2}\ln\prod_{k=0}^{N-1}\sum_{u=1}^{U}\sigma_u^2 d_u(k) \\ & -\frac{1}{2N}\mathbf{x}^T\mathbf{F}\left[\sum_{u=1}^{U}\sigma_u^2\mathbf{D}_u(\mathbf{a}_u)\right]^{-1}\mathbf{F}^H\mathbf{x} \quad \text{(D.40)} \\ = & -\frac{N}{2}\ln 2\pi - \frac{1}{2}\sum_{k=0}^{N-1}\ln\sum_{u=1}^{U}\hat{\Phi}_u(k) - \frac{1}{2}\sum_{k=0}^{N-1}\frac{\Phi(k)}{\sum_{u=1}^{U}\hat{\Phi}_u(k)} \quad \text{(D.41)}\end{aligned}$$

where $\Phi(k)$ is the k^{th} element of the periodogram of $\mathbf{x}$ and $\hat{\Phi}_u(k) = \sigma_u^2 d_u(k)$. Each $d_u(k)$ is the k^{th} diagonal element of $\mathbf{D}_u(\mathbf{a}_u)$. The summation over U spectral basis, i.e., $\sum_{u=1}^{U}\hat{\Phi}_u(k) = \widetilde{\mathbf{d}}_k^T\boldsymbol{\sigma}$ is the modeled PSD at frequency bin k. The expression in (D.41) can now be re-written into (D.16).

References

[1] B. G. Quinn, "Efficient estimation of the parameters in a sum of complex sinusoids in complex autoregressive noise," in *2007 Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers*, Nov 2007, pp. 636–640.

[2] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of quasi-harmonic sounds in colored noise," in *10th Int. Conf. on Digital Audio Effects (DAFx-07)*, 2007, pp. 1,5.

[3] K. Yoshii and M. Goto, "Infinite composite autoregressive models for music signal analysis." in *ISMIR*. Citeseer, 2012, pp. 79–84.

[4] Z. Dou, C. Shi, Y. Lin, and W. Li, "Modeling of non-gaussian colored noise and application in CR multi-sensor networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2017, no. 1, pp. 1–11, 2017.

[5] B. G. Quinn, J. K. Nielsen, and M. G. Christensen, "Fast algorithms for fundamental frequency estimation in autoregressive noise," *Signal Processing*, vol. 180, p. 107860, 2021.

[6] M. G. Christensen, "Accurate estimation of low fundamental frequencies from real-valued measurements," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2042–2056, 2013.

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

[8] J. Swärd, H. Li, and A. Jakobsson, "Off-grid fundamental frequency estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 296–303, 2017.

[9] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 4, pp. 477–489, 1988.

[10] Y. Zou and H. Liu, "A simple and efficient iterative method for Toa localization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 4881–4884.

[11] L. Blanco and M. Nájar, "Sparse covariance fitting for direction of arrival estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, p. 111, 2012.

[12] Y. Y. Al-Aboosi and A. Z. Sha'ameri, "Improved underwater signal detection using efficient time–frequency de-noising technique and prewhitening filter," *Applied Acoustics*, vol. 123, pp. 93–106, 2017.

[13] J. R. Jensen, U. Saqib, and S. Gannot, "An EM method for multichannel Toa and Doa estimation of acoustic echoes," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2019, pp. 120–124.

[14] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "A study on how pre-whitening influences fundamental frequency estimation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6495–6499.

[15] H. Yazdani, A. M. Rabiei, and N. C. Beaulieu, "On the benefits of MAI-plus-noise whitening in TH-BPSK IR-UWB systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 7, pp. 3690–3700, July 2014.

[16] A. Jakobsson, M. Mossberg, M. D. Rowe, and J. A. S. Smith, "Frequency-selective detection of nuclear quadrupole resonance signals," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 11, pp. 2659–2665, Nov 2005.

[17] A. Trucco, "Experimental results on the detection of embedded objects by a prewhitening filter," *IEEE Journal of Oceanic Engineering*, vol. 26, no. 4, pp. 783–794, Oct 2001.

[18] G. E. Birch, P. D. Lawrence, J. C. Lind, and R. D. Hare, "Application of prewhitening to AR spectral estimation of EEG," *IEEE Transactions on Biomedical Engineering*, vol. 35, no. 8, pp. 640–645, Aug 1988.

[19] Y. Zhao, R. Hu, and S. Nakamura, "Whitening processing for blind separation of speech signals," in *Proc. ICABSS*, 2003, pp. 331–336.

[20] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Instantaneous fundamental frequency estimation with optimal segmentation for non-stationary voiced speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2354–2367, Dec 2016.

[21] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "Speech decomposition based on a hybrid speech model and optimal segmentation," in *Interspeech*, 2021.

[22] ——, "On optimal filtering for speech decomposition," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2325–2329.

[23] T. Okamoto, Y. Iwaya, and Y. Suzuki, "Wide-band dereverberation method based on multichannel linear prediction using prewhitening filter," *Applied acoustics*, vol. 73, no. 1, pp. 50–55, 2012.

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

[25] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP Journal on Advances in Signal Processing*, pp. 092 953–, 2007.

[26] S. Kay and J. Salisbury, "Improved active sonar detection using autoregressive prewhiteners," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1603–1611, 1990.

[27] P. Stoica and R. L. Moses, "Spectral analysis of signals," *Pearson*, 2005.

[28] C. W. Therrien, *Discrete random signals and statistical signal processing*. Prentice Hall PTR, 1992.

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

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

[31] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based Noise Power Estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[32] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[33] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[34] M. S. Kavalekalam, J. K. Nielsen, L. Shi, M. G. Christensen, and J. Boldt, "Online parametric NMF for speech enhancement," in *26th European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 2320–2324.

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

[36] A. E. Jaramillo, A. Jakobsson, J. K. Nielsen, and M. G. Christensen, "Robust fundamental frequency estimation in coloured noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 741–745.

[37] A. E. Jaramillo, J. Nielsen, and M. Christensen, "Adaptive prewhitening based on parametric NMF," in *2019 27th European Signal Processing Conference*, September 2019.

[38] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan 1999.

[39] J. Huang and Y. Zhao, "An energy-constrained signal subspace method for speech enhancement and recognition in white and colored noises," *Speech Communication*, vol. 26, no. 3, pp. 165–181, 1998.

[40] J. K. Nielsen, M. Kavalekalam, M. Christensen, and J. Boldt, "Model-based noise PSD estimation from speech in non-stationary noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[41] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 163–176, Jan 2006.

[42] ——, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 441–452, Feb 2007.

[43] Q. He, F. Bao, and C. Bao, "Multiplicative update of auto-regressive gains for codebook-based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 457–468, March 2017.

[44] R. Hennequin, R. Badeau, and B. David, "Nmf with time–frequency activations to model nonstationary audio events," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 744–753, 2011.

[45] F. Itakura, "Analysis synthesis telephony based on the maximum likelihood method," in *The 6th international congress on acoustics, 1968*, 1968, pp. 280–292.

[46] C. Févotte, E. Vincent, and A. Ozerov, "Single-channel audio source separation with NMF: divergences, constraints and algorithms," in *Audio Source Separation*. Springer, 2018, pp. 1–24.

[47] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[48] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[49] Q. Huang, C. Bao, X. Wang, and Y. Xiang, "Speech enhancement method based on multi-band excitation model," *Applied Acoustics*, vol. 163, p. 107236, 2020.

[50] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, January 1980.

[51] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[52] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[53] N. Madhu, "Note on measures for spectral flatness," *Electronics letters*, vol. 45, no. 23, pp. 1195–1196, 2009.

[54] W. Chu and A. Alwan, "Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 3969–3972.

[55] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *EUROSPEECH*, 1995.

[56] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

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

[58] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113–116, 2002.

[59] F. Bao, H.-j. Dou, M.-s. Jia, and C.-c. Bao, "Speech enhancement based on a few shapes of speech spectrum," in *2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP)*. IEEE, 2014, pp. 90–94.

[60] A. M. Noll, "Cepstrum pitch determination," *The journal of the acoustical society of America*, vol. 41, no. 2, pp. 293–309, 1967.

[61] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–333–I–336.

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

[63] H. Ney, "A dynamic programming algorithm for nonlinear smoothing," *Signal Processing*, vol. 5, no. 2, pp. 163–173, 1983.

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

[65] J. K. Nielsen, J. R. Jensen, S. H. Jensen, and M. G. Christensen, "The single- and multichannel audio recordings database (SMARD)," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2014, pp. 40–44.

[66] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "Dregon: Dataset and methods for uav-embedded sound source localization," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–8.

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

Paper E

On Optimal Filtering for Speech Decomposition

Alfredo Esquivel Jaramillo, Jesper Kjær Nielsen, Mads Græsbøll Christensen

The paper has been published in the
Proceedings European Signal Processing Conf.. pp. 2325–2329, 2018.

The layout has been revised.

Abstract

Optimal linear filtering has been used extensively for speech enhancement. In this paper, we take a first step in trying to apply linear filtering to the decomposition of a noisy speech signal into its components. The problem of decomposing speech into its voiced and unvoiced components is considered as an estimation problem. Assuming a harmonic model for the voiced speech, we propose a Wiener filtering scheme which estimates both components separately in the presence of noise. It is shown under which conditions this optimal filtering formulation outperforms two state-of-the-art speech decomposition methods, which is also revealed by objective measures, spectrograms and informal listening tests.

1 Introduction

The decomposition of speech into its major components, i.e., voiced and unvoiced, is a challenge in many speech processing applications. An accurate recovery of these components is important in speech coding [1], analysis [2], synthesis [3], enhancement [4], as well as for diagnosing illnesses [5]. The presence of noise is inevitable in most acoustic scenarios, so a major challenging problem is the robust estimation of both components in the presence of additive noise. This is useful, for example, in remote voice assessment applications [6], and therefore, for a proper diagnosis of voice pathologies. Many clinical assesment systems have used sustained vowel phonations to detect voice pathologies, and recently [7] the need of assessing natural speech has been considered.

With the classical speech production model, speech is classified as voiced or unvoiced depending on whether the source is a periodic impulse or a white noise sequence [8]. However, specially for a good quality of synthetic speech, it has been shown [9] that a mixed excitation can produce a more natural sounding speech. This is the case for voiced fricatives (e.g. /z/). Additionally, for clinical assessment of voice impairment, it is necessary to take into account the presence of the white noise source (i.e. unvoiced component) in the vocal apparatus which results in breathy vowels and other forms of vocal dysphonia [10].

Some efforts to separate the voiced and unvoiced components from a speech signal have been developed. There are methods which make a binary voiced/unvoiced decision per frequency bin such as the one based on the multiband excitation vocoder [1] and the harmonic plus noise (HNS) model [11], and methods [12, 13] which consider both components can coexist in the speech frequency bands, which is more accurate from a speech production perspective [12]. The well-known methods at this respect are an iterative

method for a periodic and aperiodic excitation decomposition [12] and a pitch scaled harmonic filtering (PSHF) based method [13]. The iterative method operates on an assumed mixed excitation to the vocal tract by reconstructing the unvoiced part excitation in the harmonic regions which are obtained from the cepstrum. The PSHF method is based on a pitch-scaled least-squares separation of the speech signal in the frequency domain. These speech decomposition methods, which decompose speech signals into stochastic and deterministic components, do not take the presence of background noise into account in the decomposition, and thus do not distinguish between and deal with unvoiced speech and noise, which may be present at the same time.

In the speech enhancement literature, a common approach to estimate a clean signal corrupted by noise is optimal filtering, such as the classical Wiener filter [14]. Traditionally, the filter design requires estimates of the second-order statistics of the noisy signal and the noise. In this paper, we investigate if the speech decomposition problem can also be tackled via an optimal filtering way. To use optimal filtering for decomposing speech into its components, we need estimates of their second-order statistics. To obtain these, we assume a periodic signal model, namely the harmonic model, for the voiced component [15, 16]. By assuming stationarity in a short time segment, the statistics of the voiced component will depend on the fundamental frequency, the number of harmonics and the power of the harmonics. If the noise is stationary, its statistics can be estimated during periods where no voice activity is detected. Otherwise, they can be obtained through the principle of minimum tracking [17], for example. Knowing the statistics of the voiced part, of the noise and those of the observed signal, the statistics of the unvoiced part can be estimated, and, therefore, a Wiener filter can be employed to extract separately the voiced and unvoiced component.

The remainder of the paper is structured as follows. Section II introduces the signal model, the assumptions and the optimal filtering formulation. Section III establishes the proposed filtering approach and details the main parts of the statistics estimation for each of the components. Section IV gives the performance measures and the experimental results. Finally, the paper is concluded in section V.

2 Model, problem and proposed method

The speech decomposition problem considered in this paper is to extract both the zero-mean voiced $v(n)$ and unvoiced $u(n)$ components, from the noisy observation $y(n)$, i.e,

$$y(n) = s(n) + z(n) = v(n) + u(n) + z(n), \tag{E.1}$$

where $n = 0, 1, \ldots, N-1$ is the discrete-time index, $s(n) = v(n) + u(n)$ is the clean speech signal which is buried in a zero-mean additive white or colored noise $z(n)$. We assume that the voiced and unvoiced parts as well as the noise are uncorrelated.

When we adopt a linear filtering approach to recovering the desired speech components, we consider the M recent successive samples. Therefore, the signal model in (E.1) can be expressed in a vector form as

$$\mathbf{y}(n) = \mathbf{v}(n) + \mathbf{u}(n) + \mathbf{z}(n), \tag{E.2}$$

where $\mathbf{y}(n) = [y(n)\ y(n-1)\ \ldots\ y(n-M+1)]^T$ is a vector of length M, $[\cdot]^T$ denotes the transpose of a vector or matrix, and $\mathbf{v}(n)$, $\mathbf{u}(n)$ and $\mathbf{z}(n)$ are defined in a similar way to $\mathbf{y}(n)$. The objective of speech decomposition is to estimate one or more samples of $v(n)$ and $u(n)$ from the noisy vector $\mathbf{y}(n)$ by the application of two different optimal filters to the observed signal vector, i.e.

$$\begin{aligned} \hat{v}(n) &= \sum_{k=0}^{M-1} h_{v,k} y(n-k) \\ &= \mathbf{h}_v^T \mathbf{y}(n) = \mathbf{h}_v^T \mathbf{v}(n) + \mathbf{h}_v^T \mathbf{u}(n) + \mathbf{h}_v^T \mathbf{z}(n), \end{aligned} \tag{E.3}$$

$$\hat{u}(n) = \mathbf{h}_u^T \mathbf{y}(n) = \mathbf{h}_u^T \mathbf{u}(n) + \mathbf{h}_u^T \mathbf{v}(n) + \mathbf{h}_u^T \mathbf{z}(n) \tag{E.4}$$

where $\mathbf{h}_v = [h_{v,0}\ \ldots\ h_{v,M-1}]^T$, $\mathbf{h}_u = [h_{u,0}\ \ldots\ h_{u,M-1}]^T$, and $\hat{v}(n)$, $\hat{u}(n)$ are estimates of $v(n)$ and $u(n)$ respectively.

For speech decomposition, the problem is to find the optimal filters $\mathbf{h}_v$ and $\mathbf{h}_u$ which make the level of the undesired components as small as possible while passing the desired component with as little distortion as possible. The undesired components are the sum of the two last right-hand terms of (E.3) and (E.4). With the assumption that $v(n)$, $u(n)$ and $z(n)$ are uncorrelated, the $M \times M$ covariance matrix of the observed signal can be expressed as

$$\mathbf{R}_\mathbf{y} = E\left[\mathbf{y}(n)\mathbf{y}^T(n)\right] = \mathbf{R}_\mathbf{v} + \mathbf{R}_\mathbf{u} + \mathbf{R}_\mathbf{z}, \tag{E.5}$$

where $E[\cdot]$ denotes expectation, $\mathbf{R}_\mathbf{v} = E\left[\mathbf{v}(n)\mathbf{v}^T(n)\right]$, $\mathbf{R}_\mathbf{u} = E\left[\mathbf{u}(n)\mathbf{u}^T(n)\right]$, $\mathbf{R}_\mathbf{z} = E\left[\mathbf{z}(n)\mathbf{z}^T(n)\right]$ are the covariance matrices of $\mathbf{v}(n)$, $\mathbf{u}(n)$, and $\mathbf{z}(n)$, respectively.

3 Optimal Filtering and Statistics Estimation

By considering the error between the true voiced and the estimated voiced component, i.e., $e_v(n) = \mathbf{h}_v^T \mathbf{y}(n) - v(n)$, and the error between the true

unvoiced and the estimated unvoiced component, i.e. $e_u(n) = \mathbf{h}_u^T\mathbf{y}(n) - u(n)$, the mean-squared-error (MSE) criteria can be defined as

$$J_v(\mathbf{h}_v) = E\left[e_v^2(n)\right] = \sigma_v^2 - 2\mathbf{h}_v^T\mathbf{R}_v\mathbf{i}_1 + \mathbf{h}_v^T\mathbf{R}_y\mathbf{h}_v, \tag{E.6}$$

$$J_u(\mathbf{h}_u) = E\left[e_u^2(n)\right] = \sigma_u^2 - 2\mathbf{h}_u^T\mathbf{R}_u\mathbf{i}_1 + \mathbf{h}_u^T\mathbf{R}_y\mathbf{h}_u \tag{E.7}$$

where $\mathbf{i}_1$ is the first column of the $M \times M$ identity matrix $\mathbf{I}_M$, and σ_v^2 and σ_u^2 are the variances of $v(n)$ and $u(n)$, respectively. If we take the gradient of each MSE with respect to $\mathbf{h}_v$ and $\mathbf{h}_u$, and equate the results to $\mathbf{0}$, we find the Wiener filters for estimating the voiced and unvoiced speech components to

$$\mathbf{h}_v = \mathbf{R}_y^{-1}\mathbf{R}_v\mathbf{i}_1, \tag{E.8}$$

$$\mathbf{h}_u = \mathbf{R}_y^{-1}\mathbf{R}_u\mathbf{i}_1. \tag{E.9}$$

To compute these filters, the different statistics in (E.5) are required. In order to avoid problems over frame transitions of the noisy signal, we adopt a recursive approach [18], in which a short-term sample estimate and a moving average is used for computing an estimate at the time frame n as

$$\hat{\mathbf{R}}_y(n) = \alpha_y\hat{\mathbf{R}}_y(n-1) + (1-\alpha_y)\bar{\mathbf{R}}_y(n), \tag{E.10}$$

where $0 < \alpha_y < 1$ is a forgetting factor and

$$\bar{\mathbf{R}}_y(n) = \frac{1}{N-M+1}\sum_{n=0}^{N-M}\mathbf{y}(n)\mathbf{y}^H(n). \tag{E.11}$$

For the voiced part $v(n)$, we use the harmonic model, i.e.

$$v(n) = \sum_{l=1}^{L} A_l\cos(l\omega_0 n + \phi_l), \tag{E.12}$$

where L is the number of harmonics, ω_0 is the fundamental frequency, A_l denotes the real amplitude of the lth harmonic with its corresponding phase $\phi_l \in [0, 2\pi)$. As an extension to the vector model in (2), the voiced signal vector is expressed as $\mathbf{v}(n) = \mathbf{Z}\mathbf{a}$, with the definitions

$$\mathbf{a} = \frac{1}{2}[A_1e^{j\phi_1}\ A_1e^{-j\phi_1}\ \ldots\ A_Le^{j\phi_L}\ A_Le^{-j\phi_L}]^T, \tag{E.13}$$

$$\mathbf{Z} = [\mathbf{z}(\omega_0)\,\mathbf{z}^*(\omega_0)\ \ldots\ \mathbf{z}(\omega_0 L)\,\mathbf{z}^*(\omega_0 L)], \tag{E.14}$$

$$\mathbf{z}(\omega_0 l) = [1\ e^{jl\omega_0}\ldots e^{jl(M-1)\omega_0}]^T. \tag{E.15}$$

The voiced part covariance matrix $\mathbf{R_v} = E\{\mathbf{v}(n)\mathbf{v}^H(n)\} = E\{(\mathbf{Za})(\mathbf{Za})^H\}$ can be expressed as $\mathbf{R_v} \approx \mathbf{ZPZ}^H$ [19], where $[\cdot]^H$ denotes complex conjugate transpose and the amplitude covariance matrix $\mathbf{P}$ has the form [19]

$$\mathbf{P} = E\{\mathbf{aa}^H\} = \frac{1}{4}\,\mathrm{diag}([A_1^2\, A_1^2 \ldots\, A_L^2\, A_L^2]). \tag{E.16}$$

Clearly, $\mathbf{R_v}$ depends on ω_0, the model order L and the amplitude vector $\mathbf{a}$, which need to be estimated. The amplitude vector can be estimated using the principle of least-squares [20] as $\hat{\mathbf{a}} = (\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{y}$, and the fundamental frequency and model order L are estimated by using a fast nonlinear least squares (NLS) algorithm [21]. However, the NLS method assumes that the signal is observed in white gaussian noise, which is not always true in many real acoustic cases. Therefore, after estimating the noise power spectral density, a linear prediction scheme suggested in [22] is used to prewhiten the noisy signal. Then, the fundamental frequency is estimated from the prewhitened signal resulting in better frequency estimates than without prewhitening, when dealing with speech corrupted in colored noise.

As voiced speech is non-stationary across a segment of length N, a similar recursive approach to (E.10) can be used to smooth the voiced frame covariance matrix

$$\hat{\mathbf{R}}_v(n) = \alpha_v\hat{\mathbf{R}}_v(n-1) + (1-\alpha_v)\bar{\mathbf{R}}_v(n), \tag{E.17}$$

where $\bar{\mathbf{R}}_v(n) = \mathbf{ZPZ}^H$ and $0 < \alpha_v < 1$ is another forgetting factor. A noise estimator based on optimal smoothing and minimum statistics [17], for example, can be used to estimate $\mathbf{R}_z$. From (E.5), after the voiced part and noise covariance matrices are estimated, an estimate of the unvoiced component covariance matrix at the time frame n can be computed as $\hat{\mathbf{R}}_u(n) = \hat{\mathbf{R}}_y(n) - \hat{\mathbf{R}}_v(n) - \hat{\mathbf{R}}_z(n)$. To ensure that this matrix is positive definite, an eigenvalue decomposition is applied and its negative eigenvalues are replaced with a very small positive number [23].

4 Experimental results

In this section, the performance of the proposed filtering approach (optimal) is compared to the iterative periodic-aperiodic decomposition (ITER) [12] and the pitch scaled harmonic filter (PSHF) [13] based method for noisy speech signal decomposition. The state-of-the-art methods evaluated their performance in a quantitative way only for synthetic speech signals, since the individual speech components are not available separately for real speech [24]. As it is difficult to evaluate the quality of a given decomposition in an objective way, we consider an intermediate approach, where we mix

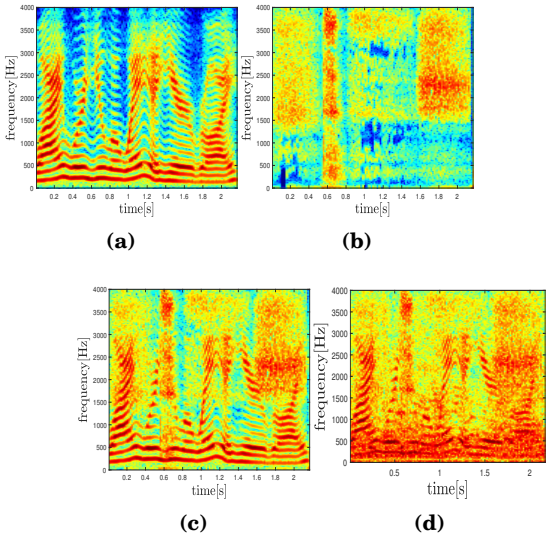

Fig. E.1: Spectrograms of (a) the true voiced component, (b) the true unvoiced component (concatenation of sounds in the order /f/, /t/, /s/, /sh/, (c) the clean speech (true voiced plus true unvoiced), and (d) the noisy speech with iSNR = 4 dB.

fully voiced and fully unvoiced utterances, so that we know the ground truth components of speech. The mixing of the two signals may not sound as natural as one would expect for common speech, but it will allow us to compare the decomposition performance with objective measures.

In the experiments, we consider five fully voiced utterances [25] (4 male and 1 female) as a ground truth for the voiced component, resampled to a sampling frequency of 8 kHz. In Fig. E.1(a), the spectrogram of the fully voiced female utterance "Why were you away a year, Roy?" is shown. For the unvoiced speech component, we consider the concatenation of five sounds /sh/, /f/, /s/, /t/, /p/ from the audio recordings of a free ebook about the full range of sounds used in general British English pronunciation [26], also resampled to 8 kHz. These sounds are either unvoiced fricatives or unvoiced stops [8]. As can be seen from Fig. E.1(b), this recording does lack a harmonic structure and has the appearance of rectangular red patterns instead of horizontal striations [8], which is representative of unvoiced speech. The clean speech for the experiment is the sum of the voiced speech and unvoiced speech, where different combinations of the five voiced sentences and ten orderings of the unvoiced sounds are considered, and the results will be averaged across the different realizations. An example of a clean signal, which contains both voiced and unvoiced parts, is shown in

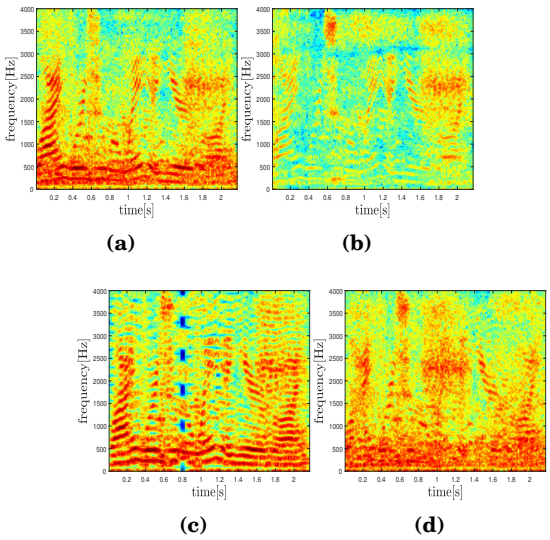

Fig. E.2: Spectrograms of (a) the estimated voiced component using the proposed filtering approach, (b) the estimated unvoiced component using the proposed filtering approach, (c) the voiced component obtained by ITER algorithm, and (d) the unvoiced component by PSHF method.

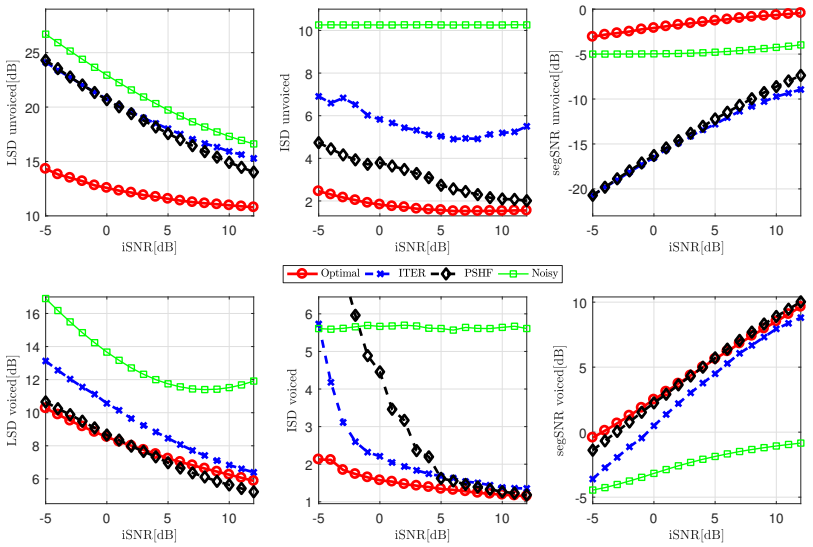

Fig. E.3: Average measured LogSpectrum distortion (LSD), Itakura-Saito distance (ISD) and segmental SNR(segSNR) for the proposed approach and state-of-the-art methods in different iSNRs and noise types. Comparison also includes the case of noisy speech as an estimate.

Fig. E.1(c). Three types of noise are considered: white, street and babble. The recordings of the street and babble noises are taken from the AURORA database [27]. In Fig. E.1(d), is shown the noisy speech spectrogram, which is formed by adding babble noise to the clean speech, the input SNR is 4 dB.

For the comparison, we add different types of noise to the speech signal at certain iSNR, ranging from -5 dB to 12 dB, two different noisy realizations at each iSNR are considered for each possible combination of voiced and unvoiced speech. For the proposed approach, the segment length is set to $N = 200$ and the filter length to $M = 25$, the forgetting factors to $\alpha_y = \alpha_v = 0.75$ in the white noise scenario and $\alpha_y = \alpha_v = 0.96$ in the street and babble noise scenario. The noise statistics are estimated using the minimum statistics (MS) [17] principle.

The spectrograms of the voiced and unvoiced component obtained by optimal filtering principle application, for the case of speech in babble noise (Fig. E.1(d)), are shown in Fig. E.2(a) and (b), the voiced component obtained by the ITER algorithm in Fig. E.2(c) and the unvoiced component which results from the PSHF method in Fig. E.2(d). The spectrogram in Fig. E.2(b) shows that the herein developed approach generates an unvoiced estimated component which looks more similar to the original unvoiced speech signal (opposed to that of Fig. E.2(d) in the sense that its spectrogram has similar red patterns as opposed to the PSHF method, which looks distributed in other frequency bins. Similar observations can be made with the ITER method. Even if some frequency bins do not appear in the spectrogram,

informal listening tests reveal that the unvoiced stops or sounds can be perceived, so the main features of the unvoiced component are preserved.

The decomposition performance was evaluated quantitatively in terms of segmental SNR (segSNR), Itakura-Saito distance (ISD), [28] and LogSpectrum Distortion (LSD) [29]. Due to space constraint, we here show the result at each iSNR averaged across the different realizations and across all the three noise types. The results are plotted in Fig. E.3. The comparison also includes the case of the noisy speech as an estimate, in order to see if the methods perform better or worse than the case of no processing of the noisy speech at all. Next, we describe what can be observed from the different plots of Fig. E.3.

The presented approach not only outperforms the other two in terms of segSNR for unvoiced speech, but it also results in a better measure against the case in which the noisy speech is considered as an estimate. This does not happen for the other methods, whose performance is below the curves of noisy speech as an estimate. In fact, as can be seen from Fig. E.2(d), the other methods show low-frequency content which is not present in the true unvoiced speech component, and that results in more signal content than this ground truth. The informal listening of their outputs does not allow to perceive all the unvoiced sounds, and some remaining of the female sentence with a high level of distortion can be listened in these unvoiced estimates. This does not occur by decomposing speech with the optimal filtering approach, in which the different unvoiced fricative and stop sounds can be perceived. In the white noise case for the ITER method, all the phonemes are lost, and for the PSHF method, only one of the phonemes is preserved, but in a very distorted manner. Much lower values of LogSpectrum distortion (LSD) and lower Itakura-Saito (ISD) distance values are also obtained with the optimal filtering formulation.

In the voiced speech case, the optimal filtering approach results in higher segSNR than the ITER method, and similar values with respect to the PSHF method at all iSNRs. It is important to mention that babble noise is one of the most difficult noise types to remove, since it is highly nonstationary and contains similar spectral content to speech. In this paper, we considered the noise statistics estimated with the default settings of the minimum statistics approach [17], but in a future improvement, the developed principle herein can be combined with a codebook-based approach [30], in order to get better estimates of the noise statistics. With respect to the Itakura-Saito distance (ISD), the ISD of the voiced component obtained by the optimal filtering formulation is lower than the other methods. This measure is more perceptually relevant than the segSNR [28]. The spectrogram of the voiced component processed by the ITER algorithm reveals some higher frequency components ($>$3000Hz), which were not present in the true voiced speech, and also some harmonics below this frequency

range, which were not present in the original speech. Informal listening test reveals that the voiced output of the ITER algorithm sounds more artificially distorted than the one obtained from the optimal filtering principle. For the developed approach, although the voiced estimate (Fig. E.2(a)) has still some noise present, it preserves the original features of the ground truth and sounds less distorted than the other methods. Finally, with the optimal filtering decomposition approach, we observe similar LogSpectrum distortion (LSD) values to the PSHF method for all iSNRs, and the proposed approach also has lower LSD values than the ITER method. Even if LSD and segSNR are similar for both approaches (optimal and PSHF), the ISD of the voiced PSHF estimate is higher for low SNRs.

5 Conclusions

In this paper, we have considered the speech decomposition problem employing the principle of optimal filtering with the corresponding statistics estimation for each one of the components of the noisy observation. We investigated if the presented approach is more robust and convenient for speech decomposition in noisy conditions. Based on the informal listening tests, spectrogram analysis and the objective measures, we found that the optimal filtering approach seems to work well.

References

[1] D. W. Griffin and J. Lim, "Multiband excitation vocoder," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 8, pp. 1223–1235, aug 1988.

[2] P. Cook, "Noise and aperiodicity in the glottal source: A study of singer voices," in *Twelfth International Congress of Phoenetic Sciences*, no. STAN-M-75, 08/1991 1991.

[3] N. B. Pinto, D. G. Childers, and A. L. Lalwani, "Formant speech synthesis: improving production quality," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 1870–1887, Dec 1989.

[4] J. Hardwick, C. D. Yoo, and J. S. Lim, "Speech enhancement using the dual excitation speech model," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, April 1993, pp. 367–370 vol.2.

[5] Y. Qi, R. E. Hillman, and C. Milstein, "The estimation of signal-to-noise ratio in continuous speech for disordered voices," *The Journal of the Acoustical Society of America 105*, vol. 105, pp. 2532–2535, 1999.

[6] R. J. Moran, R. B. Reilly, P. de Chazal, and P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 468–477, March 2006.

[7] Z. Ali, I. Elamvazuthi, M. Alsulaiman, and G. Muhammad, "Automatic voice pathology detection with running speech by using estimation of auditory spectrum and cepstral coefficients based on the all-pole model," *Journal of voice*, vol. 30, pp. 757.e7–757.e19, 2016.

[8] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. Wiley-IEEE Press, 2000.

[9] B. Atal and J. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *ICASSP. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7, 1982, pp. 614–617.

[10] D. Mehta and T. F. Quatieri, "Synthesis, analysis, and pitch modification of the breathy vowel," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.*, 2005, pp. 199–202.

[11] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, Jan 2001.

[12] B. Yegnanarayana, C. D'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 1–11, 1998.

[13] P. J. B. Jackson and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 713–726, Oct 2001.

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

[15] M. G. Christensen and A. Jakobsson, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Transactions on Signal Processing*, vol. 58, no. 12, pp. 5969–5983, Dec 2010.

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

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

[18] J. Chen, J. Benesty, and Y. A. Huang, "Study of the noise-reduction problem in the Karhunen-Loeve expansion domain," *IEEE Transactions on Speech and Audio Processing*, vol. 17, pp. 787–802, 2009.

[19] M. G. Christensen, "Accurate estimation of low fundamental frequencies from real-valued measurements," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2042–2056, 2013.

[20] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: survey, new results, and an application," *IEEE Transactions on Signal Processing*, vol. 48, pp. 338–352, 2000.

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

[22] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Instantaneous fundamental frequency estimation with optimal segmentation for non-stationary voiced speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2354–2367, Dec 2016.

[23] J. Benesty and J. Chen, *A Conceptual Framework for Noise Reduction*. Germany: Springer, 2015.

[24] C. d'Alessandro, V. Darsinos, and B. Yegnanarayana, "Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 12–23, Jan 1998.

[25] M. Cooke, *Modelling Auditory Processing and Organisation*. New York, NY, USA: Cambridge University Press, 1993.

[26] P.S., *45 Sounds of GB English*. Pronunciation Studio, 2017.

[27] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[28] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, Inc., 2013.

[29] I. McLoughlin, *Applied Speech and Audio Processing: With Matlab Examples*, 1st ed. New York, NY, USA: Cambridge University Press, 2009.

[30] J. K. Nielsen, M. Kavalekalam, M. Christensen, and J. Boldt, "Model-based noise PSD estimation from speech in non-stationary noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

Paper F

Speech Decomposition Based on a Hybrid Speech Model and Optimal Segmentation

Alfredo Esquivel Jaramillo, Jesper Kjær Nielsen, Mads Græsbøll Christensen

The paper was accepted in the
Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2021.

The layout has been revised.

Abstract

In a hybrid speech model, both voiced and unvoiced components can coexist in a segment. Often, the voiced speech is regarded as the deterministic component, and the unvoiced speech and additive noise are the stochastic components. Typically, the speech signal is considered stationary within fixed segments of 20-40 ms, but the degree of stationarity varies over time. For decomposing noisy speech into its voiced and unvoiced components, a fixed segmentation may be too crude, and we here propose to adapt the segment length according to the signal local characteristics. The segmentation relies on parameter estimates of a hybrid speech model and the maximum a posteriori (MAP) and log-likelihood criteria as rules for model selection among the possible segment lengths, for voiced and unvoiced speech, respectively. Given the optimal segmentation markers and the estimated statistics, both components are estimated using linear filtering. A codebook-based approach differentiates between unvoiced speech and noise. A better extraction of the components is possible by taking into account the adaptive segmentation, compared to a fixed one. Also, a lower distortion for voiced speech and higher segSNR for both components is possible, as compared to other decomposition methods.

1 Introduction

The problem of decomposing speech into its voiced and unvoiced components is useful in applications such as speech coding, analysis, synthesis, modification and diagnosing of illnesses [1–7]. As implied by hybrid speech models (e.g., harmonic plus noise model) [6, 8], deterministic and stochastic components may coexist in a speech segment. The deterministic part corresponds to voiced speech, which is well represented by a sum of harmonically related sinusoids [9–12], whose frequencies are an integer multiple of the pitch. The stochastic parts cover what is not described by the harmonic model, including for example, glottal turbulences and friction. The traditional speech decomposition methods [1, 13], however, do not distinguish between the unvoiced speech and the additive noise, and this distinction may be relevant, e.g., in remote voice assessment applications [14]. The method in [1] considers the colored nature of the stochastic parts of speech in order to estimate the pitch using fixed window lengths. However, the authors hypothesize that the decomposition performance can be improved by using adaptive windows instead and by estimating the number of harmonics. The iterative decomposition method in [13] is based on the cepstrum to obtain pitch information, which is not very robust under high noise conditions [15]. Moreover, in [5] it was found to converge to the wrong solution. Often, a

speech signal is assumed to be stationary within segments of a fixed length which last between 20 and 40 ms [16]. However, due to its non-stationary nature, the speech signal characteristics might change quickly during short periods of time [17]. Therefore, the optimal choice should be a time-varying segment length which better accommodates the local characteristics and has a better fit to a specified model. For example, if the pitch remains nearly constant, the segment length should be longer than when it exhibits fast variations [18].

An effort based on linear filtering to estimate separately the voiced and unvoiced parts from noisy speech was presented in [3]. In this paper, instead of relying on conventional noise tracking methods (e.g., [19]), the noise statistics were estimated using approaches which rely on prior spectral information contained in codebooks [20, 21]. In order to have a better recovery of the components, it is here proposed to do the extraction based on optimal segmentation [18, 22], instead of using a fixed one. To find the best possible segmentation, the parameters of the deterministic and stochastic parts are iteratively estimated as described in [15] for the different candidate segments and candidate models. Such approach is more robust to high noise levels than classical pitch estimators [15]. Estimates of the unvoiced and noise AR parameters are obtained from a codebook-based procedure [20] which is able to assign a zero excitation variance in silent segments for the speech part if necessary. The parameter estimates on the optimal segments are used to apply linear filtering to yield estimates of the individual components.

2 Signal model and filtering for speech decomposition

In this section, we describe how noisy speech can be decomposed into its components using linear filtering which require the knowledge of the different statistics. A speech segment of length N is described with a hybrid speech model. The model assumes that for a clean speech signal

$$s(n) = v(n) + u(n), \tag{F.1}$$

where the unvoiced part $u(n)$ is represented as an AR process and the voiced part $v(n)$ is described by the harmonic model. In speech decomposition, the goal is to extract both $v(n)$ and $u(n)$ when $s(n)$ is degraded by additive colored noise $c(n)$, i.e.,

$$y(n) = v(n) + u(n) + c(n). \tag{F.2}$$

The additive noise is also modelled as an AR process. The observation $y(n)$ can also be expressed as $y(n) = v(n) + x(n)$, where $x(n) = u(n) + c(n)$ is the

residual containing the stochastic parts of noisy speech. By considering a vector of M ($< N$) samples $\mathbf{y} = \mathbf{v} + \mathbf{u} + \mathbf{c}$, where $\mathbf{v}$, $\mathbf{u}$ and $\mathbf{c}$ are assumed to be uncorrelated, the $M \times M$ covariance matrix of the observation is expressed as the sum of covariance matrices of each component, i.e. $\mathbf{R_y} = E\left[\mathbf{y}\mathbf{y}^T\right] = \mathbf{R_v} + \mathbf{R_u} + \mathbf{R_c} = \mathbf{R_v} + \mathbf{R_x}$. Here, $E[\cdot]$ denotes expectation and $(\cdot)^T$ denotes the transpose.

Initially, we want to extract an estimate of the desired voiced speech vector $\mathbf{v}$, by applying a linear filtering $M \times M$ matrix to $\mathbf{y}$, i.e., $\hat{\mathbf{v}} = \mathbf{H}_v\mathbf{y} = \mathbf{H}_v\mathbf{v} + \mathbf{H}_v\mathbf{x}$, where

$$\mathbf{H}_v = [\mathbf{h}_{v,1}^H \; \mathbf{h}_{v,2}^H \; \ldots \; \mathbf{h}_{v,M}^H]^T, \tag{F.3}$$

$\mathbf{h}_{v,m}, m = 1, 2, ..., M$ are complex valued filters of length M and $(\cdot)^H$ is the conjugate transpose. The filtering applied in the time domain is commonly used when voiced speech parts described by the harmonic model are considered [23]. Several filter designs are possible from a recently introduced variable span linear filtering framework (VSLF) [24], by choosing a number of eigenvectors and eigenvalues of the joint diagonalization of $\mathbf{R_v}$ and $\mathbf{R_x}$. We here use the M eigenvectors $\mathbf{b}_q$ and eigenvalues λ_q to form an $M \times M$ Wiener filtering matrix [24]

$$\mathbf{H}_v = \mathbf{R}_v \sum_{q=1}^{M} \frac{\mathbf{b}_q \mathbf{b}_q^H}{1 + \lambda_q}. \tag{F.4}$$

In order to consider prior spectral information stored in codebooks for estimating $u(n)$, we make use of the corresponding representation of $\mathbf{R}_x$ in the frequency domain, i.e., the power spectral density (PSD) $\Phi_x(\omega) = \Phi_u(\omega) + \Phi_c(\omega)$, where $\Phi_u(\omega)$ is the unvoiced component PSD and $\Phi_c(\omega)$ is the noise PSD. Estimates of these PSDs can then be used to apply a frequency domain Wiener filter $H_u(\omega) = \frac{\hat{\Phi}_u(\omega)}{\hat{\Phi}_u(\omega)+\hat{\Phi}_c(\omega)}$ to yield an estimate of the unvoiced component $U(\omega) = H_u(\omega)\hat{X}(\omega)$, where $\hat{X}(\omega)$ is the spectrum of the residual which is obtained as described in the next section.

3 Statistics and parameters estimation

We now describe how to estimate the required statistics and parameters in order to apply previously described linear filtering. The harmonic model of voiced speech assumes that this component is represented as a set of sinusoids having frequencies which are an integer multiple of the pitch f_0 [9, 10], i.e.,

$$v(n) = \sum_{l=1}^{L} \left[\alpha_l e^{j2\pi f_0 l n} + \alpha_l^* e^{-j2\pi f_0 l n}\right], \tag{F.5}$$

for a segment of length N. This model, however, will have a more accurate fit for a particular segment length N_{opt}, which will be known after an

optimal segmentation of the signal has been obtained. Here, L is the unknown number of harmonics, $\alpha_l = \frac{A_l}{2} e^{j\psi l}$ is the complex amplitude of the l'th harmonic with $A_l > 0$ the real amplitude, ψ_l the initial phase and * the complex conjugate. A voiced vector of M succesive samples can be written as $\mathbf{v} = \mathbf{Z}(f_0)\boldsymbol{\alpha}$, where $\mathbf{Z}$ is a matrix of Fourier vectors, i.e.,

$$\mathbf{Z}(f_0) = [\mathbf{z}(f_0)\ \mathbf{z}^*(f_0)\ \ldots\ \mathbf{z}(Lf_0)\ \mathbf{z}^*(Lf_0)], \tag{F.6}$$

$$\mathbf{z}(lf_0) = [1\ e^{jl2\pi f_0}\ \ldots\ e^{jl2\pi f_0(M-1)}]^T, \tag{F.7}$$

and $\boldsymbol{\alpha} = \frac{1}{2}[A_1 e^{j\psi_1}\ A_1 e^{-j\psi_1}\ \ldots\ A_L e^{j\psi_L}\ A_L e^{-j\psi_L}]^T$ is a vector containing the amplitudes of the harmonics. The unvoiced parts of speech are modelled as an AR process of order P (often, set to a fixed value [6, 8]), i.e.,

$$u(n) = -\sum_{i=1}^{P} \beta_{u_i} u(n-i) + e(n), \tag{F.8}$$

where $\{\beta_{u_i}\}_{i=1}^P$ are the P AR coefficients of the unvoiced speech and $e(n)$ is the excitation WGN process with variance σ_e^2. Similarly, the colored noise $c(n)$ is modelled as an AR process with the P AR coefficients $\{\gamma_{c_i}\}_{i=1}^P$.

The estimated voiced part covariance matrix $\mathbf{R}_\mathbf{v} = E\left[\mathbf{v}\mathbf{v}^T\right]$ can be expressed as $\hat{\mathbf{R}}_\mathbf{v} = \mathbf{Z}(\hat{f}_0)\hat{\mathbf{P}}\mathbf{Z}(\hat{f}_0)^H$ [25], where the estimated amplitude covariance matrix has the form $\hat{\mathbf{P}} = E\{\hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}}^H\} = \frac{1}{4}\,\mathrm{diag}([\hat{A}_1^2\ \hat{A}_1^2 \ldots\ \hat{A}_L^2\ \hat{A}_L^2])$. It is therefore required to have estimates of the pitch f_0 and of the linear parameters. At a first instance, we would need to have estimates of these parameters from the optimal segment length N_{opt} when we do the processing based on the optimal segmentation. However, to estimate N_{opt}, we first need to estimate the parameters for all the possible segment lengths. The optimal segment length maximises the a posteriori probability of the observed data [18], as described in the next section. First, based on the estimated noise statistics (e.g., [19, 21]), a pre-processor is applied in order to pre-whiten the noise component [21], yielding the pre-whitened signal $y_{\text{W}}(n)$. This will allow to have better pitch estimates (i.e., reduce the sub-harmonic errors) from the nonlinear least-squares (NLS) estimator based on WGN assumption [11]. The parameters inside each possible candidate segment length are estimated by an approximate joint estimator of the voiced speech and the stochastic parts parameters, by iterating between these two steps: [15]

1. The f_0 is obtained from the NLS estimator [11], i.e.,

$$\hat{f}_0 = \arg\max_{f_0} \underline{\mathbf{y}}_{\text{W}}^T \mathbf{Z}(f_0) \left[\mathbf{Z}^H(f_0)\mathbf{Z}(f_0)\right]^{-1} \mathbf{Z}^H(f_0)\underline{\mathbf{y}}_{\text{W}} \tag{F.9}$$

for all candidate model orders, including $L = 0$ as a candidate to do voicing detection. The final model order L is selected using model selection criteria

such as Bayesian Information Criteria (BIC) [26]. Here $\underline{\mathbf{y}}_{\mathrm{W}}$ denotes the pre-whitened signal vector, where an underlined vector has $\overline{N}$ (or even N_{opt}) samples.

2. The amplitude vector is estimated via least-squares as $\hat{\boldsymbol{\alpha}} = [\mathbf{Z}^H(\hat{f}_0)\mathbf{Z}(\hat{f}_0)]^{-1}\mathbf{Z}(\hat{f}_0)^H\mathbf{y}$ [10], after which the AR parameters of the residual $\underline{\mathbf{x}} = \underline{\mathbf{y}} - \mathbf{Z}(\hat{f}_0)\hat{\boldsymbol{\alpha}}$ (and also $\hat{\mathbf{R}}_x$) are directly obtained [25]. These are directly used as the coefficients of a new AR pre-whitening filter, which is applied to yield $\underline{\mathbf{y}}_{\mathrm{W}}$.

The iterations are stopped when the difference of the cost function in (F.9) between two consecutive iterations is below a threshold value, or a maximum number of iterations is reached [15]. The estimation of the parameters for the different segment lengths allows us to obtain the segmentation markers for voiced speech extraction as described in the next section. Once these markers have been obtained, the noisy speech is processed to estimate the parameters inside the segments of length N_{opt}, from which $\mathbf{v}$ can be extracted using the matrix (F.4).

To obtain an estimate of the unvoiced part, we consider the modelled stochastic sequence $\underline{\mathbf{x}} = \underline{\mathbf{y}} - \mathbf{Z}(\hat{f}_0)\hat{\boldsymbol{\alpha}}$. The processing to estimate $u(n)$ is also obtained from an adaptive segmentation, but which is different from the one employed to extract the voiced part, i.e., the model in (F.8) will have a more accurate fit for an optimal segment length $N'_{\mathrm{opt}} (\neq N_{\mathrm{opt}})$. From pre-trained spectral shapes with the corresponding excitation variances, the modelled spectrum of the stochastic part is written as $\hat{\Phi}_x(\omega) = \frac{\sigma_u^2}{|B_u(\omega)|^2} + \frac{\sigma_c^2}{|\Gamma_c(\omega)|^2}$, where σ_u^2 and σ_c^2 are the excitation variances of unvoiced speech and noise, and $B_u(\omega) = 1 + \sum_{i=1}^{P} \beta_{u_i} e^{-j\omega i}$, $\Gamma_c(\omega) = 1 + \sum_{i=1}^{P} \gamma_{c_i} e^{-j\omega i}$. The parameters to be estimated are $\{\sigma_u^2, \sigma_c^2, \{\beta_{u_i}\}_{i=1}^{P}, \{\gamma_{c_i}\}_{i=1}^{P}\}$. Denoting $\beta_u^i(\omega)$ and $\gamma_c^j(\omega)$ the spectra of the i^{th} and j^{th} unvoiced speech and noise codebook entries, the single indices corresponding to the approximate ML estimate of the AR spectral shapes are obtained as

$$\{i^*, j^*\} = \arg\min_{i,j} \min_{\sigma_u^2, \sigma_c^2} d_{\mathrm{IS}}(\hat{\Phi}_x, \frac{\sigma_u^2}{\left|B_u^i(\omega)\right|^2} + \frac{\sigma_c^2}{\left|\Gamma_c^j(\omega)\right|^2}), \tag{F.10}$$

where d_{IS} is the Itakura-Saito distance. For all combinations, the excitation variances are needed and are obtained as described in [20]. Similarly to the voiced case, in order to estimate N'_{opt}, we first need to estimate the parameters for all the possible segment lengths. The optimal length will maximise the log-likelihood function, as described later. Having obtained the optimal codebook entries and excitation variances on the segment of

length N'_{opt}, they are used to form a Wiener filter

$$H_u(\omega) = \frac{\frac{\sigma_u^{2*}}{\left|B_u^{i*}(\omega)\right|^2}}{\frac{\sigma_u^{2*}}{\left|B_u^{i*}(\omega)\right|^2} + \frac{\sigma_c^{2*}}{\left|\Gamma_c^{j*}(\omega)\right|^2}}, \tag{F.11}$$

which is applied to $\underline{\underline{\mathbf{x}}}$ (of length N'_{opt}) in order to extract $u(n)$.

4 Criteria for optimal segmentation

Based on the principle of [22], in [18] it was proposed to segment the signal based on the MAP criterion which assumes a WGN condition. To deal with colored noise, it is therefore required to pre-whiten $y(n)$ [21]. The segmentation markers are required before applying the linear filtering to extract the voiced part. Each way in which the signal can be segmented (i.e., a segment composed of a number of minimum-length segments) is considered as a model, among a set of candidate models $\mathcal{M}$. Under the MAP criterion, the model which maximizes the model a posteriori probability given the observation, will be selected. The criterion [10, 18, 27] consists of a data log-likelihood term, and a term which penalizes model complexity. The estimated model order $\hat{L}$ is a function of N, in which $\hat{L}(N)$ and $\hat{f}_0(N)$ are estimated for each candidate segment with the iterative procedure described in [15]. For a candidate segment detected as voiced, i.e., $\hat{L}(N) \neq 0$, and considering the real signal harmonic model, the MAP cost function is

$$J_1(N) = \frac{N}{2} \ln \frac{1}{N} ||\underline{\mathbf{y}}_{\mathrm{W}} - \mathbf{Z}\boldsymbol{\alpha}_{\mathrm{W}}||_2^2 + \frac{3}{2} \ln N + \hat{L}(N) \ln N, \tag{F.12}$$

in which the amplitude vector $\boldsymbol{\alpha}_{\mathrm{W}}$ is obtained in this case from the pre-whitened signal. If a candidate segment is detected as not-voiced, i.e., $\hat{L}(N) = 0$, the MAP cost function involved in the comparison is instead $J(N) = \frac{N}{2} \ln ||\underline{\mathbf{y}}_{\mathrm{W}}||_2^2$. After the extraction of voiced speech, the modelled residual $x(n)$ is segmented based on the log-likelihood

$$J_2(N) = \frac{N}{2} d_{\mathrm{IS}}(\hat{\Phi}_x, \frac{\sigma_u^2}{\left|B_u^i(\omega)\right|^2} + \frac{\sigma_c^2}{\left|\Gamma_c^j(\omega)\right|^2}) + \frac{1}{2} \sum_{k=1}^{N} \ln \hat{\Phi}_x. \tag{F.13}$$

The model which maximises the log-likelihood given the observed residual will be selected. The markers are required before applying the filter in (F.11).

The segmentation requires that the cost is additive and independent over the segments, which is satisfied for both previous criteria. The optimal

lengths N_{opt} and N'_{opt} are found by comparing the cost of all the possibilities from the set of segment lengths and choosing the one minimizing the cost over all candidates, i.e., $\widehat{\mathcal{M}} = \arg\min_{\mathcal{M}} J_i$, $i \in \{1,2\}$. A minimal segment length, N_{min}, is defined, generating a subsegment of N_{min} samples and dividing the signal into S subsegments. This gives 2^{S-1} ways of segmenting the signal into S subsegments, and a maximum number of subsegments B_{max} is set. A dynamic programming algorithm is then used to find the optimal number of subsegments in a segment, b_{opt}, for all subsegments, $s = 1,...,S$, starting at $s = 1$ moving continuously to $s = S$ [22]. For every subsegment, the cost of all new subsegment combinations are reused from earlier subsegments. When the end of the signal is reached, the optimal segmentation of the signal is found, starting at the last subsegment and jumping backwards through the signal until reaching the beginning. This is done by starting at $s = S$ and setting the number of subsegments in the last segment to $b_{\text{opt}}(S)$. Thereby, the next segment ends at subsegment $s = S - b_{\text{opt}}(M)$ and includes $b_{\text{opt}}(S - b_{\text{opt}}(S))$ subsegments. This is continued until $s = 0$. The segmentation algorithm is described in [18].

To summarize, the steps to decompose (offline) noisy speech into its voiced and unvoiced components are:

1. The noisy signal is pre-processed with an adaptive autoregressive pre-whitener [21], yielding $y_{\text{W}}(n)$.

2. Parameter estimates of $v(n)$ and $x(n)$ are jointly obtained [15] for all candidate segment lengths. Followingly, based on (F.12), the markers of the optimal segmentation for voiced speech and N_{opt} are obtained.

3. Parameter estimates of $v(n)$ and $x(n)$ and statistics $\mathbf{R}_{\mathbf{v}}$, $\mathbf{R}_{\mathbf{x}}$ are obtained from the segments of length N_{opt}. If $\hat{L}(N_{\text{opt}}) \neq 0$, estimate $\mathbf{v}$ using (F.4) after joint diagonalization of $\mathbf{R}_{\mathbf{v}}$ and $\mathbf{R}_{\mathbf{x}}$.

4. Obtain the modelled residual $\underline{\mathbf{x}} = \underline{\mathbf{y}} - \mathbf{Z}(\hat{f}_0)\hat{\boldsymbol{\alpha}}$ in all the different obtained optimal lengths $\{N_{\text{opt}}\}$. Once the whole modelled $x(n)$ is obtained, estimate unvoiced speech parameters $\{\sigma_u^2, \{\beta_{u_i}\}_{i=1}^{P}\}$ for all candidate segment lengths.

5. Based on (F.13), obtain the markers of the optimal segmentation for unvoiced speech and N'_{opt}.

6. The unvoiced speech parameters $\{\sigma_u^2, \{\beta_{u_i}\}_{i=1}^{P}\}$ are obtained from the segments of length N'_{opt}. Extract $\underline{\mathbf{u}}$ using (F.11).

5 Experimental evaluation

We first illustrate the extracted speech components of one of the clean female excerpts from the Keele database [28], after the voiced speech

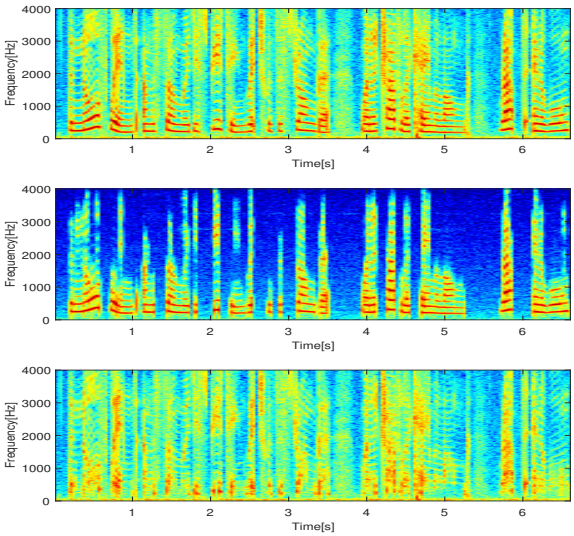

Fig. F.1: From top to bottom: Spectrograms of the observed signal, extracted voiced part, and extracted unvoiced part.

segmentation markers were obtained. Although subsegments increasing in steps of 5 ms (i.e., $N = 40$ at 8 kHz) are considered for $v(n)$, only segments from 20 to 50 ms (i.e. $N = 160$ to $N = 400$ in steps of 40) are possible in the segmentation. That is, the maximum number of possible subsegments is $B_{\max} = 10$, and the cost for $b = 1$, $b = 2$ or $b = 3$ is set to infinity, as the pitch estimator does not work well for very short segment lengths and low pitch f_0. For the optimal segments for which $\hat{L}(N) \neq 0$, the filtering matrix (F.4) with $M = 40$ is applied, and the filtering is updated every 20 samples, i.e., there is a 50 % of overlap. $M = 40$ was chosen as it is an integer divisor of all the candidate segment lengths, which facilitates the processing. The difference from the clean signal and $v(n)$ corresponds to $u(n)$, and this corresponds to a ground truth for the unvoiced speech component. The spectrograms of $s(n)$ and its corresponding $v(n)$ and $u(n)$ are displayed on Fig.1. It is seen that $v(n)$ has an appearance with horizontal striations and that $u(n)$ is displayed by rectangular patterns over a wide range of frequencies. Around 2.1 s, the harmonics up to around 3 kHz are obtained in $v(n)$. An example of how the time series of $v(n)$ and $u(n)$ look like if either a segment of fixed length is used (here 20 ms), or if the extraction of $v(n)$ is done using the optimal segmentation, is displayed in Fig. 2. The unvoiced part obtained from the optimal segmentation used for $v(n)$ exhibits a more stochastic nature compared to the one obtained from using segments of a fixed size to extract $v(n)$. The marked region exhibits a periodic nature, which corresponds to $v(n)$. The optimal segmentation results in a better

modelling of the periodic parts in the extracted voiced component.

We now proceed to evaluate the decomposition performance in noisy conditions. Four excerpts of 4 s of the Keele database files were added babble, factory, street and restaurant noise, at iSNRs of 0, 5 and 10 dB. The performance per iSNR is presented averaged across all the noise types, and two runs are done per excerpt and noise type. That is, a total of 32 runs are considered per iSNR. Before applying the segmentation, the signal was pre-whitened from the setup described in [21], which relies on a parametric NMF noise PSD estimate. The codebook of AR entries of unvoiced speech (including also from silent segments) was obtained from the training on samples which correspond to the difference of clean signals and the voiced speech extracted from the Wiener filter. And as stated before, this corresponds to a ground truth of unvoiced speech. The samples used for training were different than those at evaluation. The training was done on segments of length $N = 160$ (i.e., 20 ms) with an overlap of 50 % between them, with an AR order $P = 14$. Similarly, the codebook of AR entries of noise (including babble, F-16, restaurant and factory [29]) was trained. A total of 64 unvoiced speech and 16 noise entries were obtained from a standard vector quantization technique [30]. When evaluating (F.10), the modelled $\underline{x}$ was fitted to an AR spectrum Φ_x of order 28 [20]. To extract $u(n)$, segments from 15 to 40 ms were made possible for the segmentation. The results are shown in Figure 3, comparing the performance of applying the optimal segmentation to the extraction based on a traditional fixed one (20 ms). The decomposition performance is evaluated in terms of segSNR and Log Spectral distance (LSD) [20]. It is also compared to the decomposition methods [1, 13] after OMLSA speech enhancement [31] was applied as a pre-processor. This is done to attenuate the noise which is not taken into account in them. The comparison also has the case where noisy speech is obtained as an estimate, in order to see if the methods perform better than the case of not processing the signal. At an iSNR of 10 dB, the extraction based on adaptive segments leads to a higher segSNR for the case of $v(n)$. Also, with respect to the LSD, lower values are obtained for both the extracted $v(n)$ and $u(n)$ based on adaptive segments. At an iSNR of 5 dB, although the confidence intervals of segSNR overlap, as seen from the extreme intervals, there is higher probability that the adaptive segmentation leads to a better recovery of the components, and also the LSD values are clearly separated. At 0 dB, both ways of segmenting lead to similar performance. From using optimal segmentation, it is possible to get lower LSD for $v(n)$ compared to [1], which based its processing on segments of fixed size. Although it is possible to achieve higher segSNR with the proposed, it is seen that the other methods combined with enhancement achieve lower LSD for $u(n)$, at lower SNRs. However, there is a potential to trade off distortion and noise reduction by considering other filters in the

VSLF framework [24].

6 Discussion

The use of an optimal segmentation combined with parameter estimates of an hybrid speech model allow to have a more accurate recovery of the voiced and unvoiced speech parts, compared to the use of fixed segments. Specifically, an adaptive segmentation results in a better modelling of the periodic parts in the voiced component with a higher probability of improved segSNR and also of a lower LSD of both extracted voiced and unvoiced parts. We considered prior spectral information stored in codebooks in order to differentiate between unvoiced speech and noise. A higher segSNR and lower LSD for the voiced part is possible when compared to reference methods, with a potential to reduce the LSD for the extracted unvoiced part. As future work, we will consider deriving the segmentation based on the recently introduced joint pitch-AR estimator [12].

References

[1] B. Elie and G. Chardon, "Robust tonal and noise separation in presence of colored noise, and application to voiced fricatives," in *22nd International Congress on Acoustics (ICA)*, 2016.

[2] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 8, pp. 1223–1235, 1988.

[3] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "On optimal filtering for speech decomposition," in *26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2325–2329.

[4] D. Mehta and T. F. Quatieri, "Synthesis, analysis, and pitch modification of the breathy vowel," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.* IEEE, 2005, pp. 199–202.

[5] P. J. Jackson, "Characterisation of plosive, fricative and aspiration components in speech production," Ph.D. dissertation, University of Southampton, 2000.

[6] Y. Stylianou, J. Laroche, and E. Moulines, "High-quality speech modification based on a harmonic+ noise model," in *Fourth European Conference on Speech Communication and Technology*, 1995.

[7] E. Belalcazar-Bolanos, J. Orozco-Arroyave, J. Arias-Londono, J. Vargas-Bonilla, and E. Nöth, "Automatic detection of Parkinson's disease

using noise measures of speech," in *Symposium of Signals, Images and Artificial Vision-2013: STSIVA-2013*. IEEE, 2013, pp. 1–5.

[8] Y. Stylianou, "Modeling speech based on harmonic plus noise models," in *International School on Neural Networks, Initiated by IIASS and EMFCSC*. Springer, 2004, pp. 244–260.

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

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

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

[12] B. G. Quinn, J. K. Nielsen, and M. G. Christensen, "Fast algorithms for fundamental frequency estimation in autoregressive noise," *Signal Processing*, vol. 180, p. 107860, 2021.

[13] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 1–11, 1998.

[14] R. J. Moran, R. B. Reilly, P. de Chazal, and P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 468–477, 2006.

[15] A. E. Jaramillo, A. Jakobsson, J. K. Nielsen, and M. G. Christensen, "Robust fundamental frequency estimation in coloured noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 741–745.

[16] K. K. Paliwal, J. G. Lyons, and K. K. Wójcicki, "Preference for 20-40 ms window duration in speech analysis," in *2010 4th International Conference on Signal Processing and Communication Systems*, 2010, pp. 1–4.

[17] F. R. Drepper, "A two-level drive–response model of non-stationary speech signals," in *International Conference on Nonlinear Analyses and Algorithms for Speech Processing*. Springer, 2005, pp. 125–138.

[18] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Instantaneous fundamental frequency estimation with optimal segmentation for non-stationary voiced speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2354–2367, Dec 2016.

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

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

[21] A. E. Jaramillo, J. Nielsen, and M. Christensen, "Adaptive pre-whitening based on parametric NMF," in *2019 27th European Signal Processing Conference*, September 2019.

[22] P. Prandoni and M. Vetterli, "R/D optimal linear prediction," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 646–655, 2000.

[23] S. M. Nørholm, *Enhancement of speech signals-with a focus on voiced speech models*, ser. Ph.D. thesis, Aalborg Universitet, 2015.

[24] J. Benesty, M. G. Christensen, and J. R. Jensen, *Signal enhancement with variable span linear filters*. Springer, 2016, vol. 7.

[25] P. Stoica and R. L. Moses, "Spectral analysis of signals," *Pearson*, 2005.

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

[27] P. M. Djuric, "A model selection rule for sinusoids in white gaussian noise," *IEEE Transactions on Signal Processing*, vol. 44, no. 7, pp. 1744–1751, 1996.

[28] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *EUROSPEECH*, 1995.

[29] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[30] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on communications*, vol. 28, no. 1, pp. 84–95, 1980.

[31] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal processing letters*, vol. 9, no. 4, pp. 113–116, 2002.

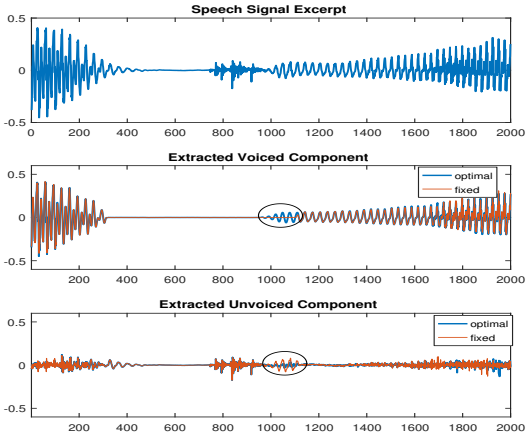

Fig. F.2: Extraction of voiced and unvoiced components from optimal and fixed segmentation on a clean signal excerpt.

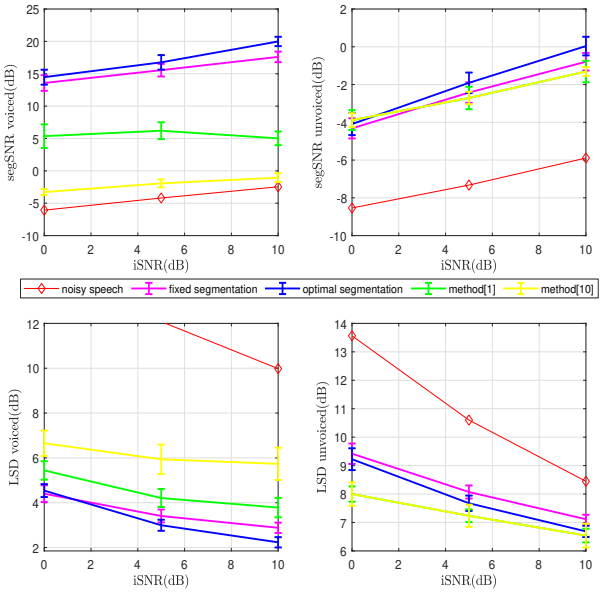

Fig. F.3: LSD and segmental SNR (segSNR) in different iSNRs averaged across four noise types.

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-984-8

AALBORG UNIVERSITY PRESS