Enhancement of speech signals - with a focus on voiced speech models

Nørholm, Sidsel Marie

DOI (link to publication from Publisher):
10.5278/vbn.phd.engsci.00058

Publication date:
2015

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

Link to publication from Aalborg University

Citation for published version (APA):
Nørholm, S. M. (2015). Enhancement of speech signals - with a focus on voiced speech models. Aalborg Universitetsforlag. (Ph.d.-serien for Det Teknisk-Naturvidenskabelige Fakultet, Aalborg Universitet). DOI: 10.5278/vbn.phd.engsci.00058

ENHANCEMENT OF SPEECH SIGNALS - WITH A FOCUS ON VOICED SPEECH MODELS

BY
SIDSEL MARIE NØRHOLM

DISSERTATION SUBMITTED 2015

AALBORG UNIVERSITY
DENMARK

Enhancement of Speech Signals - with a Focus on Voiced Speech Models

Ph.D. Dissertation

Sidsel Marie Nørholm

Aalborg University
Audio Analysis Lab
Department of Architecture, Design and Media Technology
Rendsburggade 14
DK-9000 Aalborg

Thesis submitted:	November, 2015
PhD supervisor:	Professor Mads Græsbøll Christensen Aalborg University
Assistant PhD supervisor:	Post Doc. Jesper Rindom Jensen Aalborg University
PhD committee:	Professor MSO Stefania Serafin (chairman) Aalborg University Copenhagen
	Professor Ioannis Stylianou University of Crete
	Reader, Dr. Wenwu Wang University of Surrey
PhD Series:	Faculty of Engineering and Science, Aalborg University

ISSN (online): 2246-1248
ISBN (online): 978-87-7112-404-0

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2015

Abstract

The topic of this thesis is speech enhancement with a focus on models of voiced speech. Speech is divided into two subcategories dependent on the characteristics of the signal. One part is the voiced speech, the other is the unvoiced. In this thesis, we primarily focus on the voiced speech parts and utilise the structure of the signal in relation to speech enhancement.

The basis for the models is the harmonic model which is a very often used model for voiced speech because it describes periodic signals perfectly. First, we consider the problem of non-stationarity in the speech signal. The speech signal changes its characteristics continuously over time whereas most speech analysis and enhancement methods assume stationarity within 20-30 ms. We propose to change the model to allow the fundamental frequency to vary linearly over time by introducing a chirp rate in the model. Filters are derived based on this model and it is shown that they perform better than filters based on the traditional harmonic model. In the filter design, estimates of the fundamental frequency and chirp rate are needed. Therefore, an iterative nonlinear least squares method to estimate the parameters jointly is suggested. The estimator reaches the Cramér-Rao bound, and the iterative approach makes the method faster than searching the original two dimensional space for the optimal combination of fundamental frequency and chirp rate. To counteract the effect of non-stationarity further, we suggest that the segment length should not be fixed but depend on the signal at the given moment. Thereby, short segments can be used when the signal characteristics vary fast, and long segments can be used when the characteristics are more stationary. We propose to choose the segment length according to the maximum a posteriori criteria and show that the segmentation based on the chirp model gives longer segments than for the harmonic model. This suggests that the chirp model fits the voiced speech signal better. Other deviations from the perfect harmonic model can occur. As it is well known from stiff-stringed musical instruments, the frequencies of the harmonics in speech may also deviate from the perfect harmonic relationship. We propose to take these deviations into account by extending the harmonic model to the inharmonic model where small perturbations at each harmonic can occur. Three different methods to estimate the inharmonicities are com-

pared, and it is shown that including the estimate in the filter design leads to better performance than a filter based on the traditional harmonic model.

We also propose to take a subspace perspective to speech enhancement by performing a joint diagonalisation of desired signal and noise. The eigenvectors generated from this operation is used to make a filter that estimates the noise, and the desired signal is estimated by subtracting the noise estimate from the observed signal. The filter is very flexible in the way that it can trade noise reduction and signal distortion based on how many eigenvectors are used in the filter design. The number of eigenvectors used in the filter in voiced speech periods can also be chosen based on the harmonic model since the number of harmonics in the speech signal is closely related to the best choice of number of eigenvectors.

The papers in this thesis show that it can be beneficial to extend the traditional harmonic model to include non-stationarity and inharmonicity of speech. The derived filters perform better than filters based on the harmonic model in terms of signal-to-noise ratio and signal distortion. The voiced speech models can also be used to make a noise covariance matrix estimate which can be used in other algorithms as, e.g., the proposed joint diagonalisation based method.

Resumé

Emnet for denne afhandling er støjreduktion i talesignaler med et fokus på modeller af stemt tale. Tale er delt i to underkategorier afhængigt af karakteristika af signalet. Den ene del er stemt tale, den anden er ustemt tale. I denne afhandling fokuserer vi primært på den stemte tale og udnytter strukturen af signalet i relation til støjreduktion.

Udgangspunktet for modellerne er den harmoniske model, som er en ofte brugt model for stemt tale. Vi adresserer problemet med ikke-stationær tale, da karakteristika for talesignalet ændrer sig kontinuært over tid. De fleste metoder til støjreduktion og analyse af talesignaler antager at signalet er stationært i analysevinduer på 20-30 ms, hvilket ikke er tilfældet. Vi foreslår at ændre den harmoniske model så den tillader fundamentalfrekvensen at ændre sig lineært indenfor et analysevindue. Dette gøres ved at introducere en chirpparameter i modellen, og vi udleder filtre baseret på denne model. I filterdesignet er der brug for at estimere både fundamentalfrekvensen og chirpparameteren. Derfor foreslår vi endvidere en iterativ metode baseret på en nonlineær mindste kvadraters metode til at estimere disse to samtidig. Den iterative tilgang gør metoden hurtigere end at gennemsøge et todimensionelt rum for den optimale kombination af fundamentalfrekvens og chirpparameter. For at kunne tage hensyn til at signalet ikke er stationært i endnu højere grad, foreslår vi at gøre længden af analysevinduet tidsvarierende og afhængigt af signalets karakteristik, og vi udleder en metode til optimal segmentering af signalet baseret på maximum a posteriori princippet. Vi viser at chirpmodellen giver længere segmenter end den traditionelle harmoniske model, hvilket antyder at chirpmodellen passer bedre til talesignalet end den traditionelle harmoniske model. Andre afvigelser fra den harmoniske model kan også forekomme. Som det er velkendt fra strengeinstrumenter, kan frekvenserne af de harmoniske også afvige fra den perfekte harmoniske relation og vi foreslår derfor at tage disse afvigelser i betragtning og udvider den harmoniske model til den inharmoniske model, hvor små afvigelser kan forekomme ved hver harmonisk frekvens. Tre forskellige metoder til at estimere inharmoniciteterne præsenteres og sammenlignes, og resultater viser at det giver bedre resultater, når inharmoniciteterne tages med i betragtning.

Vi foreslår også at tage et subspace perspektiv til støjreduktion ved at lave en fælles diagonalisering af det ønskede signal og støjen, hvor egenvektorerne fra dette bruges til at generere et filter. Filteret estimerer støjen, og det ønskede signal er fundet ved at trække dette estimat fra det observerede signal. Filteret er meget feksibelt og graden af støjreduktion i forhold til signalforvrængning kan let ændres ved at ændre antallet af egenvektorer inkluderet i filterdesignet. Antallet af egenvektorer i filterdesignet i perioder af stemt tale kan også bestemmes ud fra den harmoniske model, da antallet af harmoniske i signalet er nært relateret til det bedste antal af egenvektorer brugt i filteret.

Artiklerne i denne afhandling viser at det kan være fordelagtigt at udvide den traditionelle harmoniske model til at kunne beskrive ikke-stationær tale og inharmoniciteter. De udledte filtre virker bedre end filtre baseret på den traditionelle harmoniske model når de sammenlignes på signal-støj-forhold og signalforvrængning. Modellerne for stemt tale kan også blive brugt til at estimere støjkovariansmatricen som kan blive brugt i andre algoritmer som for eksempel den foreslåede fælles diagonaliseringsmetode.

Contents

Thesis Details

Thesis Title:	Enhancement of speech signals - with a focus on voiced speech models
Ph.D. Student:	Sidsel Marie Nørholm
Supervisors:	Prof. Mads Græsbøll Christensen, Aalborg University Postdoc Jesper Rindom Jensen, Aalborg University

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

[A] Sidsel Marie Nørholm, Jesper Rindom Jensen, Mads Græsbøll Christensen, "Enhancement of Non-Stationary Speech using Harmonic Chirp Filters," *Proc. Interspeech*, accepted for publication, 2015.

[B] Sidsel Marie Nørholm, Jesper Rindom Jensen, Mads Græsbøll Christensen, "Enhancement and Noise Statistics Estimation for Non-stationary Voiced Speech," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, submitted, 2015.

[C] Sidsel Marie Nørholm, Jesper Rindom Jensen, Mads Græsbøll Christensen, "Instantaneous Pitch Estimation with Optimal Segmentation for Non-Stationary Voiced Speech," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, submitted, 2015.

[D] Sidsel Marie Nørholm, Jesper Rindom Jensen, Mads Græsbøll Christensen, "On the Influence of Inharmonicities in Model-Based Speech Enhancement," *Proc. European Signal Processing Conf.*, pp. 1-5, 2013.

[E] Sidsel Marie Nørholm, Jesper Rindom Jensen, Mads Græsbøll Christensen, "Spatio-Temporal Audio Enhancement Based on IAA Noise Covariance Matrix Estimates," *Proc. European Signal Processing Conf.*, pp. 934-938, 2014.

[F] Sidsel Marie Nørholm, Martin Krawczyk-Becker, Timo Gerkmann, Steven van de Par, Jesper Rindom Jensen, Mads Græsbøll Christensen, "Least Squares Estimate of the Initial Phases in STFT based Speech Enhancement," *Proc. Interspeech*, accepted for publication, 2015.

[G] Sidsel Marie Nørholm, Jacob Benesty, Jesper Rindom Jensen, Mads Græsbøll Christensen, “Single-Channel Noise Reduction using Joint Diagonalization and Optimal Filtering,” *EURASIP J. on Advances in Signal Processing*, vol. 2014, no. 1, pp. 1-11, 2014.

In addition to the main papers, the following publications have also been made.

[1] Sidsel Marie Nørholm, Jacob Benesty, Jesper Rindom Jensen, Mads Græsbøll Christensen, “Noise Reduction in the Time Domain using Joint Diagonalization,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 7058-7062, 2014.

[1] Sidsel Marie Nørholm, Jesper Rindom Jensen, Mads Græsbøll Christensen, “Optimal Segmentation for Analysis of Non-Stationary Voiced Speech,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, submitted.

This thesis has been submitted for assessment in partial fulfillment of the PhD degree. The thesis is based on the submitted or published scientific papers which are listed above. Parts of the papers are used directly or indirectly in the extended summary of the thesis. As part of the assessment, co-author statements have been made available to the assessment committee and are also available at the Faculty. The thesis is not in its present form acceptable for open publication but only in limited and closed circulation as copyright may not be ensured.

Preface

This thesis is submitted to the Doctoral School of Engineering and Science at Aalborg University in partial fulfilment of the requirements for the degree of doctor of philosophy. The thesis consists of two parts: the first part is an introduction to the research area and the second part is a collection of papers that have been published or submitted to peer-reviewed conferences and journals. The work was carried out in the period from September 2012 to September 2015 in the Audio Analysis Lab at the Department of Architecture, Design and Media Technology at Aalborg University.

The project was made possible by a research grant from the Villum Foundation, and for this I am very grateful. The grant would of course never have been available in this field of study without the work of applying for the grant by my supervisor, Mads Græsbøll Christensen. Besides from this, Mads has also been passionate about discussing ideas and sharing his knowledge in the field. This I would like to thank him for. I would also like to thank my co-advisor, Jesper Rindom Jensen, for many beneficial discussions. During these three years, I had a stay at Oldenburg University which was made possible by the invitation from Timo Gerkmann and Steven van de Par so I send my thanks to Germany as well. Besides from that, I would like to thank my office mates for discussions and a nice and friendly tone in the office. In the end, I would like to thank my family and friends and not least Brian Bemman for support during the work on my PhD.

Sidsel Marie Nørholm
Aalborg University, November 18, 2015

Part I

Introduction

Introduction

This thesis is concerned with the enhancement of speech, here seen as the problem of removing noise from noise contaminated speech signals. Two different approaches to speech enhancement can be taken. One which is based on information about the desired signal and one which is based on the noise contaminating the signal. The first six contributions in this thesis are concerned with speech enhancement based on signal models. In order to make the best models possible, it is important to know something about the structure of the speech signal which is closely coupled to how the speech is produced. Therefore, this introduction will start out by explaining a bit about the speech production system in Section 1. Based on knowledge of the speech production system, models of voiced speech can be generated and used for speech enhancement. Some models of relevance for this thesis are presented in Section 2. After leaving the speaker, the speech is often corrupted by unwanted noise from the environment as, e.g., car noise or other speakers. In order to process the speech and make an effort to remove the noise, the signal has to be recorded. This can be done either with a single microphone or an array of several microphones. The signal models including the noise are presented in Section 3 in the single and multichannel case. Speech enhancement algorithms can now be applied to the recorded speech signal. A vast amount of research exists in the field of speech enhancement. In Section 4, an introduction to the different groups of speech enhancement methods is given. All methods, both the ones relying on information about the desired signal and those depending on the noise, need some extra information than what is directly given from the observed signal. Section 5 is concerned with the estimation of noise statistics and signal parameters. The last section, Section 6, gives an overview of the papers constituting the main body of this thesis and the main findings of each contribution.

1 Speech production

The production of speech is a result of pressing air from the lungs up through the trachea, larynx, pharynx and mouth or nose cavities [13]. The structures from larynx and up are in common called the vocal tract. Speech can either

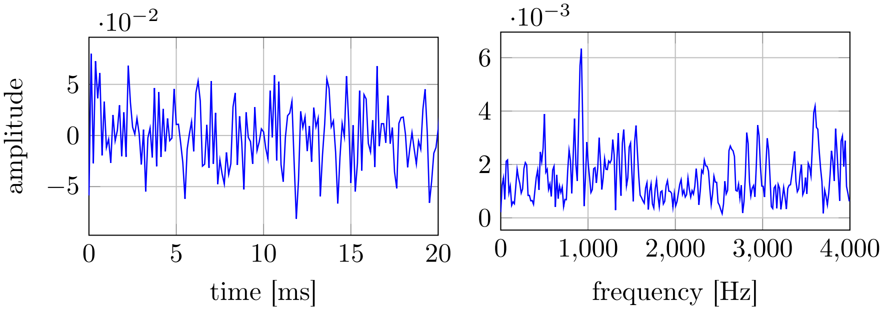


Fig. 1: A piece of unvoiced speech and its spectrum.

be generated by vibration of the two vocal folds, situated in the larynx, or by passage of air through structures in the mouth. These two speech types have very different characteristics and are, therefore, split into two groups. The speech generated by vibration of the vocal folds are called voiced speech whereas the other type is called unvoiced speech. During normal breathing and in the production of unvoiced speech, the vocal folds are relaxed and air can pass freely between them. The unvoiced speech sounds are mostly produced in the mouth cavity, e.g., by the passage of air through the teeth and lips or by a sudden opening of the lips [113]. These sounds often have a low amplitude, have a short duration of maybe 10 ms and have a close similarity to random noise [32]. An example of a piece of unvoiced speech and its spectrum is shown in Figure 1. In order to produce voiced speech sounds, the vocal folds are contracted, obstructing the air flow between the lower and upper respiratory tract. This produces an overpressure in the lungs which in the end causes the vocal folds to open and let out the air. Thereby, the pressure in the lungs decreases and the vocal folds close again due to the Bernoulli effect [13]. Hereafter, the cycle repeats which makes the vocal folds vibrate and generate a quasiperiodic signal [32]. A periodic signal, $x(t)$, is one in which the signal repeats itself with regular intervals of length T [106], i.e.,

$$x(t) = x(t+T), \quad \forall\, t. \tag{1}$$

In a quasiperiodic signal, the period changes slowly over time. Therefore, when looking at a short time interval, the signal is approximately periodic

$$x(t) \approx x(t+T), \quad t_1 < t < t_2. \tag{2}$$

The quasiperiodicity of the voiced speech can be noticed in Figure 2 where a piece of voiced speech and its spectrum is shown. In the figure a little more than three periods are seen. The three periods look very similar, but it is still easy to see that the signal is changing from one period to the next. The period of the vibration, T, or the fundamental frequency, $f_0 = 1/T$, depends

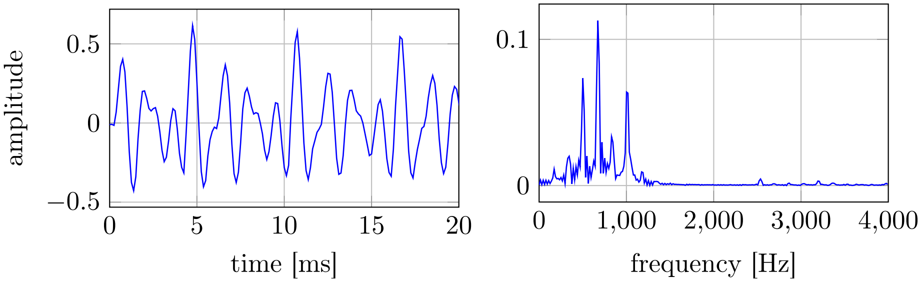


Fig. 2: A piece of voiced speech and its spectrum.

on the elasticity, tension and mass of the vocal folds [13]. Long and thick vocal folds will have longer periods of vibration than thin and short vocal folds. Therefore, men often have a lower fundamental frequency than women. The average fundamental frequency of men is around 125 Hz whereas women have an average fundamental frequency around 200 Hz [143]. However, it is possible to modify the length of the vocal folds within a speaker. A lengthening of the vocal folds increases the tension of the folds, and, thereby, the fundamental frequency also increases. The fundamental frequency is, therefore, not static but changes over time leading the generated signal to be quasiperiodic instead of periodic. The signal generated by the vocal folds contain more frequencies than the fundamental. It also contains frequencies given by a multiple integer times the fundamental, i.e., $2f_0$, $3f_0$ and so forth and falls, therefore, in the category of harmonic signals. The voiced speech segments are longer than the unvoiced with durations up to 100 ms [32].

The signal generated by the vocal folds are modified up through the vocal tract. The vocal tract can be considered a tube open in one end and will resonate accordingly. Further, the pharynx, nasal and mouth cavities are resonators [13]. These resonators shape the signal coming from the vocal folds and make it sound like different voiced sounds as, e.g., the vowels *e* and *i*. This can be seen as filtering the source signal from the vocal folds with a filter given by the characteristics of the vocal tract. The filter will to some extent be speaker dependent, but to a higher extent it will be dependent on which sound the speaker is producing. The signal generated by the vocal folds is of interest in order to build the models of the desired signal we consider later. It tells something about the frequency content of the speech signal whereas the effect of the vocal tract can be modelled by changing the amplitudes and phases of the signal.

Linear prediction (LP) [7, 30, 146] can be used for separating the source signal and the vocal tract filter. The model is made up having the source signal as input to a vocal tract filter followed by a filter representing the lip radiation which gives rise to a low pass filtering of the signal. The source signal is either random noise if the resulting speech signal is unvoiced, or a quasiperiodic pulse train followed by a glottal filter if the resulting signal is voiced [30, 63, 131]. Using linear prediction, a combined filter of the vocal tract and lip radiation is found by minimising the mean squared error between the estimated signal and the speech signal. The error obtained from this is the source signal. This causes some problems in the case of the qausiperiodic source since the least squares minimisation gives the solution with an error resembling white noise as much as possible [52]. To counteract this problem it has been suggested [33, 51, 52, 100] to set up a sparse linear prediction problem instead. This ensures that the residual is more sparse than with the traditional approach and, thereby, will resemble a voiced speech excitation more. The sparsity is obtained by changing the 2-norm to the 1-norm which is used as an approximation to the 0-norm. Linear prediction is often used for speech coding [63, 115], but it is very sensitive to noise [125] and it sounds synthetical [63]. Another way to get an insight into the source signal is to use electroglottography (EGG) which gives information about the vocal fold contact area (VFCA) [123]. The EGG signal is sometimes included in speech databases as, e.g., the Keele database [112] and since it reflects the VFCA it can also be used to estimate the fundamental frequency of the voiced speech signal.

2 Signal models

Since the primary constituent of the speech signal is voiced [32, 65], we focus on models for this type of speech. As was seen in Figure 2, the voiced speech is approximately periodic, repeating itself after a short time interval. Further, the spectrum shows a very characteristic pattern with content at a few frequencies that are spaced equally on the frequency axis. Due to this, voiced speech is often modelled using a harmonic model [3, 22, 71, 82, 83, 139]. The harmonic model is not only important in relation to voiced speech signals since it can also be used to describe a variety of other signals as, e.g., sounds from musical instruments such as guitars, pianos and violins [4, 119], sounds from other animals as birds and whales [2, 135], electrocardiograms (ECGs) [101] and astronomical data [116].

Harmonic model

The harmonic model describes the sampled voiced speech by a sum of sinusoids with their frequencies related and given by multiple integer times the fundamental frequency, f_0, [71]

$$s(n) = \sum_{l=1}^{L} A_l \cos(l\omega_0 n + \phi_l), \tag{3}$$

where $n = 0, 1, 2, ..., N$ is the discrete time index, L is the number of harmonics or the model order, A_l is the amplitude of the l'th harmonic, $\omega_0 = 2\pi f_0/f_s$ is the normalised fundamental frequency in radians per sample with f_s being the sample frequency and ϕ_l is the initial phase of the l'th harmonic. Using Eulers formula this can be rewritten as

$$s(n) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n} + \alpha_l^* e^{-jl\omega_0 n}, \tag{4}$$

where j is the imaginary unit, $\alpha_l = \frac{A_l}{2}e^{j\phi_l}$ and $(\cdot)^*$ denotes the complex conjugate. To lower the computational complexity and ease the notation, the signal can be converted into its complex counterpart by use of the Hilbert transform [22, 60, 93]

$$s(n) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n}, \tag{5}$$

where $\alpha_l = A_l e^{j\phi_l}$. Speech enhancement is often performed on real signals whereas the parameter estimation is performed using the complex model. However, speech enhancement can also be performed using the complex model since the real signal can easily be obtained from its complex counterpart [22].

The harmonic model is used for voiced speech signals under the assumption of stationarity within the analysis frame. The analysis frame is often chosen to be in the proximity of 30 ms [32]. Even though a short analysis frame is chosen it is well known that the signal is not stationary [32, 37] but has a fundamental frequency that varies continuously over time. This non-stationarity can be taken into account by using the harmonic chirp model instead of the traditional harmonic model.

Chirp model

In a chirp, the signal is not stationary but changes characteristics over time in a specific manner. In the harmonic chirp model, the instantaneous frequency varies linearly with time instead of being constant within the analysis frame. Therefore, the instantaneous frequency of the l'th harmonic, $\omega_l(n)$, is given by

$$\omega_l(n) = l(\omega_0 + kn), \tag{6}$$

where k is the fundamental chirp rate. The instantaneous phase, $\varphi_l(n)$, is given by the integral of the instantaneous frequency, i.e.,

$$\varphi_l(n) = l\left(\omega_0 n + \frac{1}{2}kn^2\right) + \phi_l, \tag{7}$$

and, thereby, the harmonic chirp model becomes

$$s(n) = \sum_{l=1}^{L} \alpha_l e^{jl\left(\omega_0 n + \frac{1}{2} k n^2\right)}. \tag{8}$$

Here, the chirp model is given in the complex framework but the models for real signals can easily be extended to include the chirp signal as well. The chirp model has earlier been considered in an alternative to the standard fast Fourier transform (FFT) for the analysis of speech signals in [79, 152]. It is shown that more sharp peaks are obtained in the positions of the harmonics compared to using the normal FFT where the spectrum is more blurred. Chirp signals have also been considered in the area of automatic speech recognition [144, 145] where more robust features are obtained when the model is extended to include chirp signals.

With a signal model matching the desired signal better, it is worth considering if longer frames than the normal ones of approximately 30 ms can be used to analyse and process the signal. The small frame sizes are set under the assumption of stationarity to ensure that this assumption at least to a reasonable degree is satisfied. If another model that takes the non-stationarity of speech into account is used, the frame size can be revised to fit the assumptions of the new model which in the case of the linear chirp model will be that the fundamental frequency changes linearly within one frame. The advantage of longer frames will be a higher resolution during frequency analysis of signals due to the time-bandwidth product [133], as well as more accurate estimates of the parameters of the model as seen in Papers C and D and in [72, 73] where the best obtainable accuracy of the estimators are seen to decrease with the frame length.

The characteristics of the signal changes all the time and some times faster than others. This means that the assumptions of stationarity or a linearly changing fundamental frequency are better fulfilled in some frames than in others. Therefore, a fixed frame length might not be an optimal choice. Instead a varying window size might be beneficial. In periods where the fundamental frequency changes with a fixed rate for a long period, a long window might as well be used, whereas at points where the characteristics of the fundamental frequency changes fast, a short window might be better, maybe even a shorter window than the original fixed window size. In [114, 115], varying segment lengths based on linear prediction and quantisation error is suggested.

Inharmonic model

Looking at the spectrum of a speech signal, the harmonics are not situated at exact multiples of the fundamental [110], leading to the introduction of an inharmonic signal model [22, 46]:

$$s(n) = \sum_{l=1}^{L} \alpha_l e^{j(l\omega_0 + \Delta_l)n}, \tag{9}$$

where Δ_l is a small deviation away from the harmonic frequency. However, due to the time-bandwidth product [133], it is difficult to know the exact reason for this inharmonicity. Using very short segments for the analysis gives a poor frequency resolution whereas the signal gets less stationary when the analysis frame is extended. Spectra of chirp signals can show peak splitting and other alterations of the harmonic structure when analysed using a standard FFT and it is, therefore, difficult to distinguish between the two phenomena in the analysis. No matter the reason, taking the inharmonicity into account can lead to better results, e.g., in amplitude estimation [109].

Inharmonicity is also known from musical instruments. Here the inharmonicity is more well studied and is an accepted phenomenon taken into account in piano tuning [119]. The well accepted inharmonicity in musical instruments makes it easy to believe that a similar phenomenon is present in voiced speech even though no absolute evidence has been given yet. The inharmonicity in musical instruments follows a more restricted model than the one suggested for voiced speech:

$$s(n) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0\sqrt{1+Bl^2}n} \tag{10}$$

where $B \ll 1$ is an instrument dependent stiffness parameter [22]. The model is also used for fundamental frequency estimation in musical signals [38, 56].

3 Noisy environments

Unfortunately, most often the desired signal, $s(n)$, is corrupted by noise from the environment. Therefore, we do not have access to the clean signal. Normal hearing listeners have an implicit speech enhancement system in the auditory system which make them able to extract the speech signal in varying noise conditions from street noise to picking out a single speaker out of a large company [15, 28, 58, 120]. However, for hearing impaired listeners, this property is to some extent destroyed leading to the cocktail party problem [7, 20]. This means that they need the desired signal to have a higher level relative to the noise in order to understand what is being said than do listeners with normal hearing [58], and they can, therefore, benefit from preprocessing of the signal in order to enhance the desired signal. Further, speech enhancement is important when a recording of the noisy speech is used for other purposes such as in communication systems where noise decreases the coding efficiency [80, 96, 125, 142], or in automatic speech recognition where the word error rate (WER) is increased when the signal is noisy [68, 89, 140].

The environmental noises have different characteristics dependent on their origin. Noise from machinery, including, e.g., cars, has a high content of low frequencies whereas the noise coming from other speakers at a party has a very similar frequency content to the signal which is sought isolated from the mixture. This of course makes the problem of separating one speaker from others a more difficult problem than enhancing speech contaminated by car noise. In a lot of speech enhancement and parameter estimation methods, the noise is assumed to be white and Gaussian. The assumption of the noise being Gaussian distributed is the most conservative assumption one can make and it will lead to the worst-case Cramér-Rao bound [134]. The white noise assumption means that the noise is assumed evenly distributed in the entire frequency range which is rarely the case for real life noise types. Sometimes the methods work well even though the noise does not fulfil the assumptions. Otherwise, it is possible to prewhiten the signal [61, 62] to make the noise more white, but this demands knowledge of the noise statistics, and the operation can have an influence on the desired signal too. Environmental noises are assumed independent of the desired signal, and signal and noise have an additive relation. The signal and noise can also have a convolutive relation. This is the case in rooms where the sound is reflected from the walls. The reflections are used by normal listeners to identify the room and where the sound is coming from and are often included as part of the desired signal as in [77]. However, the reflections can be a problem for hearing impaired listeners and multichannel speech enhancement methods and are, therefore, sometimes treated as noise. The problem is well studied [45, 84, 129, 130], but this thesis is focusing on the additive noise types, and convolutive noise will not be considered further.

3.1 Observed signal

The speech signal can be recorded by a single microphone leading to the observed signal:

$$x(n) = s(n) + v(n), \tag{11}$$

where $v(n)$ is the additive noise. With the harmonic signal model this leads to:

$$x(n) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n} + v(n). \tag{12}$$

The speech processing is often done on a vector of N consecutive samples, here defined for discrete time indices going forward in time from n to $n+N-1$:

$$\begin{aligned}\mathbf{x}(n) &= [x(n)\ x(n+1)\ x(n+N-1)] && (13)\\ &= \mathbf{s}(n) + \mathbf{v}(n), && (14)\end{aligned}$$

where $\mathbf{s}(n)$ and $\mathbf{v}(n)$ are defined in a similar way to $\mathbf{x}(n)$. The desired signal vector starting at $n = 0$ can be written as a combination of a matrix of Fourier vectors with frequencies corresponding to the harmonic frequencies

$$\mathbf{Z} = \left[\mathbf{z}(\omega_0)\ \mathbf{z}(2\omega_0)\ \ldots\ \mathbf{z}(L\omega_0)\right], \tag{15}$$

$$\mathbf{z}(l\omega_0) = \left[1\ e^{jl\omega_0}\ \ldots\ e^{jl\omega_0(N-1)}\right]^T, \tag{16}$$

and a vector containing the complex amplitudes $\mathbf{a} = [\alpha_1\ \alpha_2\ \ldots\ \alpha_L]^T$. However, to shift the start position of the vector to an arbitrary n, the matrix $\mathbf{D}(n)$ is included:

$$\mathbf{D}(n) = \begin{bmatrix} e^{j\omega_0 n} & & 0 \\ & \ddots & \\ 0 & & e^{jL\omega_0 n} \end{bmatrix}, \tag{17}$$

and the final model becomes

$$\mathbf{x}(n) = \mathbf{Z}\mathbf{D}(n)\mathbf{a} + \mathbf{v}(n). \tag{18}$$

The model can easily be changed to another set of time indices by multiplying the signal model with a simple delay, or it can be extended to the harmonic chirp model as we suggest in the Papers A, B and C or the inharmonic model as suggested in Paper D by changing the exponent of the exponential function.

Alternatively, the speech signal can be sampled with several microphones. An example of this can be seen in Figure 3 where a uniform linear array (ULA) is used. In the multichannel scenario the signal recorded by sensor n_s at time instant n_t is given by:

$$x_{n_s}(n_t) = s_{n_s}(n_t) + v_{n_s}(n_t). \tag{19}$$

Using the uniform linear array, assuming the harmonic model and no attenuation due to the distance to the source, the recorded signal is given by [73]:

$$x_{n_s}(n_t) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n_t} e^{-jl\omega_s n_s} + v_{n_s}(n_t), \tag{20}$$

where $\omega_s = \omega_0 f_s c^{-1} d \sin\theta$ is the spatial fundamental frequency, with c the speech of sound, d the distance between the microphones and $\theta \in [-90°; 90°]$ the direction of arrival measured relative to a line perpendicular to the ULA. The relation between ω_0 and ω_s is dependent on the array structure, and the model can be adapted to other array structures by exploiting the relation of the microphones for the array under consideration. For processing in the multichannel case, a vector model similar to the one in (18) is defined as can be seen in [73] and is used in Paper E. The model can of course also be changed to cover the chirp model or the inharmonic model.

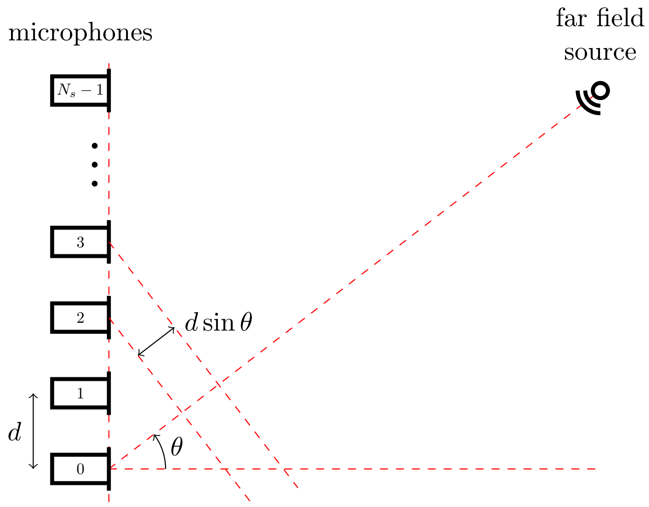


Fig. 3: Sketch of a multichannel setup with a uniform linear array.

4 Speech enhancement

The objective of speech enhancement is to recover the desired signal from the noisy observation in the best possible way. This means that the noise should be reduced while the distortion of the desired signal is kept at a minimum. There are two different viewpoints one can take to speech enhancement. One is the noise driven approach, the other is the signal driven approach. In the noise driven approach, the enhancement process is based on estimation of noise statistics whereas in the signal driven approach, the speech signal is assumed to follow a given model as, e.g., the harmonic model where the parameters of the model has to be estimated from the noisy speech signal. Speech enhancement can both take place in the time domain or some transformed domain. The choice of domain depends on the approach. The signal driven approaches often take place directly in the time domain whereas the noise driven approaches often take place in a transformed domain. The subspace methods transform the signal into a signal dependent domain as the Karhunen-Loéve domain when the eigenvalue decomposition (EVD) is used [41, 98]. A very common signal independent domain to use is the frequency domain because it is computational effective [8]. A real valued signal will be represented by complex coefficients in the frequency domain and is often separated into an amplitude and a phase term:

$$\begin{aligned} X(\omega) &= |X(\omega)|e^{\phi_x(\omega)} && (21)\\ &= S(\omega) + V(\omega) && (22)\\ &= |S(\omega)|e^{\phi_s(\omega)} + |V(\omega)|e^{\phi_v(\omega)}, && (23) \end{aligned}$$

where $X(\omega)$ is the Fourier transform of $\mathbf{x}(n)$, $|X(\omega)|$ is the amplitude, and $\phi_x(\omega)$ is the phase, and similar for $S(\omega)$ and $V(\omega)$. It is common to focus on enhancement of the amplitude and combine it with the noisy phase, but recently the focus on phase enhancement has increased as will be discussed in Section 4.3.

In the following some different speech enhancement methods are presented.

4.1 Noise driven approach

According to [91] speech enhancement methods can be divided into three groups:

- Spectral subtraction methods
- Subspace methods
- Statistical-model-based methods

The methods covered by these three groups all belong to the noise driven speech enhancement methods.

Spectral subtraction methods

The first attempt to spectral subtraction was done in 1975 [151]. The action is done on a pseudo-cepstrum by setting elements close to the origin to zero. In 1979, Boll proposed a spectral subtraction algorithm in the Fourier transform domain [12] which has been the basis of further attempts to perform spectral subtraction. The principle behind spectral subtraction is very simple and is based on the assumption that the desired signal and noise are additive. Therefore, the original spectrum can be obtained by subtracting the noise spectrum from that of the observed signal [91]:

$$\widehat{S}(\omega) = \left(|X(\omega)| - |\widehat{V}(\omega)|\right) e^{j\phi_x(\omega)}, \qquad (24)$$

where $S(\omega)$, $X(\omega)$ and $V(\omega)$ are the Fourier transformed counterparts of $s(n)$, $x(n)$ and $v(n)$, $\widehat{(\cdot)}$ denotes estimates, and $\phi_x(\omega)$ is the phase of $X(\omega)$. The spectrum of the noise is not known and has to be estimated. An incorrect estimate can, e.g., lead to negative absolute values which of course does not reflect the spectrum of the desired signal and has to be corrected for. This can be done by setting negative values to zero [91], but this leads to very abrupt changes in the spectrum which gives rise to unpleasant artefacts when listening to the enhanced signal. To lower the impact of this, it is suggested in [12] to replace the negative value with the minimum estimated value from adjacent frames, and in [11] to use the noise estimate multiplied by a factor much smaller than one. Both will give a smoother spectrum and lower the artefacts in the signal. Besides from magnitude errors, the algorithm suffers from cross-term errors and phase errors, where the magnitude errors are dominating at good noise conditions shifting towards the other two error types when the noise conditions worsen [42]. The flaws in the algorithm make it suffer from musical noise [11, 57] and much research has been put into the reduction of this kind of noise [11, 57, 90, 92, 94].

Subspace methods

The subspace methods build on the assumption that the covariance matrix of the desired signal is rank deficient which is approximately the case for voiced speech. The covariance matrix of the noise is, on the other hand, assumed to be full rank. Algorithms have been based on both the singular value decomposition (SVD) [16, 126] and the eigenvalue decomposition (EVD) [41, 127]. The difference between the two methods is that the SVD is used on a Hankel matrix composed of signal values whereas the EVD is used on the covariance matrix of the signal, and the two methods are shown to give the same end result [61]. The EVD of the $M \times M$ covariance matrix of the real observed signal is given by:

$$\mathbf{R}_x = \mathbf{U}\mathbf{\Lambda}_x\mathbf{U}^T = \mathbf{U}(\mathbf{\Lambda}_s + \mathbf{\Lambda}_v)\mathbf{U}^T, \tag{25}$$

where the last equality holds under the assumption of uncorrelated desired signal and noise. The matrix $\mathbf{U}$ is orthonormal and contains the eigenvectors, $\mathbf{\Lambda}_x$, $\mathbf{\Lambda}_s$ and $\mathbf{\Lambda}_v$ are diagonal matrices containing the non-negative eigenvalues in descending order corresponding to the covariance matrices of observed signal, desired signal and noise, and $\mathbf{R}_x$ is the covariance matrix of the observed signal:

$$\begin{align} \mathbf{R}_x &= \mathbb{E}(\mathbf{x}(n)\mathbf{x}^H(n)) \tag{26} \\ &= \mathbf{R}_s + \mathbf{R}_v. \tag{27} \end{align}$$

where the last equality again holds under the assumption of uncorrelated desired signal and noise, $\mathbb{E}(\cdot)$ denotes mathematical expectation and $\mathbf{R}_s$ and $\mathbf{R}_v$ are the covariance matrices of desired signal and noise, respectively, defined in a similar way to $\mathbf{R}_x$ in (26). When the covariance matrix of the desired signal is rank deficient with a rank of $P < M$, the last $P+1, ..., M$ eigenvalues are zero. On the other hand, for both the EVD and SVD based subspace methods, the noise is assumed white Gaussian, and all the eigenvalues will, therefore, be the same and equal to the variance of the noise. This leads to a subspace spanned by the first P eigenvectors containing desired signal plus noise, and a subspace spanned by the last $M-P$ eigenvectors containing only noise. Theoretically, the desired signal could be reconstructed by subtracting the noise variance from all eigenvalues and projecting the observed signal unto this space, however, the noise is never perfectly white and this approach is, therefore, not widely used. Alternatively, the desired signal can be estimated from projecting the observed signal on the subspace spanned by the first P eigenvectors, avoiding contributions from the subspace only containing the noise. It is a difficult task to find the correct rank P, but the choice of P is important. If the rank is chosen too small, a part of the desired signal is removed by the operation whereas if the rank is chosen too big, an unnecessary amount of noise is included. Therefore, the choice of P is a tradeoff between noise reduction and signal distortion. For voiced speech, the rank is closely related to the model order, and the rank can often be chosen relatively low. Unvoiced speech is much more distributed over the entire space, and a higher rank has to be chosen in order not to distort the desired signal too much. Some different strategies for choosing the rank of the desired signal plus noise subspace can be found in [3, 17, 31, 118].

The white noise assumption makes it necessary to prewhiten the observed signal in most practical situations. This can be done by use of the Cholesky factorisation of the noise covariance matrix [61] or, alternatively, the joint diagonalisation of signal and noise can be used [61, 66]:

$$\mathbf{V}^T \mathbf{R}_s \mathbf{V} = \mathbf{\Lambda}, \tag{28}$$

$$\mathbf{V}^T \mathbf{R}_v \mathbf{V} = \mathbf{I}_M, \tag{29}$$

where $\mathbf{V}$ and $\mathbf{\Lambda}$ are the eigenvectors and eigenvalues, respectively, of the matrix $\mathbf{R} = \mathbf{R}_v^{-1}\mathbf{R}_s$, and $\mathbf{I}_M$ is an $M \times M$ identity matrix. In [35], the Wiener filter has been derived in the framework of joint diagonalisation and the method is extended to the multichannel case. Optimal filters can also be generated based on the eigenvectors from the joint diagonalisation as we do in Paper G. Recently, a multichannel decomposition method based on frequency domain data has also been proposed in [10].

Statistical-model-based methods

The methods herein are based on a statistical model of desired signal and noise. In [97], a maximum likelihood (ML) estimator of the spectral amplitudes of the speech signal is derived based on the assumption of Gaussian models for both speech and noise. It is argued that the perception of speech is phase insensitive, and, therefore, only the amplitude is included in the estimation. The derived method is reported to work well for speech in noise, but the effect of noise suppression is not satisfactory when speech is not present. The model is extended with a probability of speech presence leading to a greater amount of noise suppression when speech is not present. Based on the same statistical assumptions, a minimum mean squared error (MMSE) spectral amplitude estimator is derived in [39]. It is shown that the MMSE estimate of the phase is given by the noisy phase, and based on this the phase of the observed signal is used unaltered. The same authors later modify the method to minimise the mean squared error of the log-spectral amplitudes instead [40]. This is motivated by the way sound is perceived by the human auditory system.

This group also contains the Wiener filter which minimises the mean squared error between the desired signal and the estimated desired signal. In the time domain the error is given by [19]

$$e_s(n) = s(n) - \widehat{s}(n) = s(n) - \mathbf{h}_w^H \mathbf{x} \tag{30}$$

where $\mathbf{h}_w = [h_0 \; h_1 \; \dots \; h_{M-1}]$ is the Wiener filter of length M. The Wiener filter is then

$$\begin{align} \mathbf{h}_w &= \arg\min_{\mathbf{h}} \|e_s(n)\|_2^2 \tag{31} \\ &= \mathbf{R}_x^{-1} \mathbf{R}_s \mathbf{i}_M, \tag{32} \end{align}$$

where $\mathbf{i}_M$ is the first column of an $M \times M$ identity matrix. The Wiener filter often introduces quite a lot of distortion. Recently, other filters have been proposed which are derived based on certain performance measures as, e.g., signal distortion [9]. This gives a new family of filters with more control over the amount of noise reduction and signal distortion. One of these filters minimise the amount of noise passing through the filter subject to the constraint that the signal distortion should stay below a given limit. This will generate a more flexible filter where it is possible to trade off noise reduction and signal distortion. The filter has a very similar appearance to the original Wiener filter [6]

$$\mathbf{h}_\lambda = \left(\mathbf{R}_s + \frac{1}{\lambda} \mathbf{R}_v \right)^{-1} \mathbf{R}_s \mathbf{i}_M. \tag{33}$$

Here $\lambda > 0$ is a tuning parameter. When $\lambda \to \infty$, $\mathbf{h}_\lambda \to \mathbf{i}_M$, which gives $\widehat{s} = x(n)$ and the observed signal is passed unaltered through the filter, when $\lambda = 1$, the filter reduces to the Wiener filter, and when $\lambda \to 0$ a large amount of signal distortion is introduced.

The Wiener filter is also often used in the frequency domain. Here, the filtered frequency coefficients are given by [86]

$$\widehat{S}(\omega) = \frac{|S(\omega)|^2}{|V(\omega)|^2 + |S(\omega)|^2} X(\omega). \tag{34}$$

The Wiener filter has also been extended to the multichannel case.

4.2 Signal driven approach

Based on the harmonic model, a set of filters for speech enhancement is derived. The principle is to pass the signal at the harmonic frequencies while suppressing the content at other frequencies. The first attempts to do this was done using the comb filter [43, 87, 102] which has peaks at the harmonic frequencies and valleys in between. The comb filter is a bit rigid and cannot adapt to different noise types. Therefore, other filters have been constructed based on the principle of the Capon beamformer [18] which passes the signal coming from one direction and suppresses other directions as much as possible. This filter has a single constraint at the angle of the desired signal whereas in [44] the algorithm is expanded to include several constraints. This could be to pass the signal from one direction while completely blocking the signal coming from another direction. This can be adapted to harmonic speech signals by setting up constraints at each of the harmonics, forcing the filter to pass them all undistorted while attenuating the signal at other frequencies. This can be expressed by the optimisation problem [22]:

$$\min_{\mathbf{h}} \mathbf{h}^H \mathbf{R}_x \mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^H \mathbf{Z} = \mathbf{1}^T, \tag{35}$$

where $\mathbf{h} = [h(0)\, h(1)\, \ldots\, h(M-1)]^H$ is the filter response of length M, and $\mathbf{1} = [1\, \ldots\, 1]^T$. The filter fulfilling this is the linearly constraint minimum variance (LCMV) filter given by:

$$\mathbf{h}_{\text{LCMV}} = \mathbf{R}_x^{-1}\mathbf{Z}(\mathbf{Z}^H \mathbf{R}_x^{-1}\mathbf{Z})^{-1}\mathbf{1}. \tag{36}$$

The estimated signal is given by:

$$\widehat{s} = \mathbf{h}^H \mathbf{x}. \tag{37}$$

Dependent on whether $\mathbf{Z}$ is generated according to the traditional harmonic model, the harmonic chirp model, or the inharmonic model, the filter will be optimised for the respective signals. In the multichannel case, the filter has a similar structure, but here both $\mathbf{R}_x$ and $\mathbf{Z}$ have to be defined slightly different than in the single channel case [73] as can be seen in Paper E.

When the speech signal fulfils the harmonic model completely, it is shown that using the covariance matrix of the observed signal, $\mathbf{R}_x$, in (35) is the same as using the noise covariance matrix, $\mathbf{R}_v$, [71]. However, when this is not the case, using the noise covariance matrix will lead to better results. Using the covariance matrix of the observed signal will make the algorithm minimise the overall output from the filter, including parts of the desired signal, whereas using the noise covariance matrix will only lead to a minimisation of the noise output from the filter. Another related filter can be derived where the noise covariance matrix is estimated as an intrinsic part of the filter design. This is the amplitude and phase estimation filter (APES) [23, 69, 70, 133]. The APES filter is derived by minimising the mean squared error between the filtered signal and the signal model:

$$J = \frac{1}{N-M+1} \sum_{n=0}^{N-M} |\mathbf{h}^H \mathbf{x}(n) - \mathbf{a}^H \mathbf{w}(n)|^2, \tag{38}$$

where $\mathbf{w}(n) = [e^{j\omega_0 1n} \ \dots \ e^{j\omega_0 Ln}]$. The error is minimised under the same constraint as the LCMV filter, $\mathbf{h}^H \mathbf{Z} = \mathbf{1}$, leading to the filter:

$$\mathbf{h}_{\text{APES}} = \mathbf{Q}^{-1}\mathbf{Z}(\mathbf{Z}^H \mathbf{Q}^{-1} \mathbf{Z})^{-1}\mathbf{1}, \tag{39}$$

where $\mathbf{Q}$ is a noise covariance matrix estimate given by:

$$\mathbf{Q} = \mathbf{R}_x - \mathbf{G}^H \mathbf{W}^{-1} \mathbf{G}, \tag{40}$$

$$\mathbf{G} = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{w}(n)\mathbf{x}^H(n), \tag{41}$$

$$\mathbf{W} = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{w}(n)\mathbf{w}^H(n). \tag{42}$$

The APES filter performs better than the LCMV due to this intrinsic noise covariance matrix estimate. However, as shown in Paper B, the APES noise covariance matrix estimate can also be beneficial to use in other filters which are not based on the signal driven approach.

4.3 Phase enhancement

In many of the enhancement methods performed in the frequency domain, only the amplitude spectrum is changed whereas the phase is left unaltered. This is the case in, e.g., the spectral subtraction methods, the ML estimator presented in [97] and the MMSE estimator in [39]. In [97], it is argued that the perception of speech is phase insensitive whereas in [39], the MMSE phase estimate is used, which happens to be the noisy phase. The use of the noisy phase is also motivated by [149] where it is shown through listening experiments that the phase is unimportant in relation to the perceived signal-to-noise ratio (SNR). However, in [105] the phase is shown to be important in the reconstruction of signals. The spectral amplitude and phase are used separately paired with a dummy phase or amplitude term, or crossed with the amplitude and phase from another signal. In the first case, the signal reconstructed from the phase, resembles the original signal more than the one reconstructed from the amplitude, and in the second case, the reconstructed signal has closest similarities with the signal delivering the phase term to the reconstructed signal. Listening tests in [107] consolidates the importance of the phase in speech enhancement.

Another motivation for modifying the phase is given in [136] where the inconsistency of phase is discussed. Only modifying the spectral amplitude, gives inconsistent spectra, meaning that the spectrum of the reconstructed signal is not the same as the original spectrum. In [59], an iterative algorithm is presented where the phase is modified to make the spectrum consistent, and in [85], an algorithm for a consistent Wiener filter is presented. In [81, 82], a method to estimate the phases independently is proposed. It is shown that the phase in voiced speech periods can be evolved using the harmonic model leading to better PESQ scores [67] for the reconstructed signals. The phase estimate can also be used to give better estimates of the spectral amplitude [47, 49]. An overview of phase enhancement and the recent advances in the topic is given in [50].

4.4 Performance measures

The performance of the speech enhancement methods is often measured relative to the ratio of signal to noise in the observed signal given by the input signal-to-noise ratio (SNR) [5]:

$$\text{iSNR} = \frac{\sigma_s^2}{\sigma_v^2}, \tag{43}$$

where σ_s^2 and σ_v^2 are the variance of the desired signal and noise, respectively, before speech enhancement.

One way to evaluate the speech enhancement method is by use of the output SNR which should be greater then the input SNR for the speech enhancement method to be successful [5]

$$\text{oSNR} = \frac{\sigma_{s,\text{nr}}^2}{\sigma_{v,\text{nr}}^2}, \tag{44}$$

where $\sigma_{s,\text{nr}}^2$ and $\sigma_{v,\text{nr}}^2$ are the variances of the desired signal and noise after noise reduction. For the time domain filtering methods where the desired signal is given as the output from the filter, the output SNR can be expressed as [5]

$$\text{oSNR} = \frac{\mathbb{E}(\mathbf{h}^H\mathbf{s}\mathbf{s}^H\mathbf{h})}{\mathbb{E}(\mathbf{h}^H\mathbf{v}\mathbf{v}^H\mathbf{h})} = \frac{\mathbf{h}^H\mathbb{E}(\mathbf{s}\mathbf{s}^H)\mathbf{h}}{\mathbf{h}^H\mathbb{E}(\mathbf{v}\mathbf{v}^H)\mathbf{h}} = \frac{\mathbf{h}^H\mathbf{R}_s\mathbf{h}}{\mathbf{h}^H\mathbf{R}_v\mathbf{h}}. \tag{45}$$

However, increasing the SNR is not the only quality a speech enhancement method should have. If the signal is distorted in the effort to increase the SNR, the quality of the signal might not be improved even though the SNR is. Therefore, the signal distortion is also an important measure in the analysis of a proposed method. The distortion can be measured in different ways as, e.g., the signal reduction factor [5]:

$$\xi_{\text{sr}} = \frac{\sigma_s^2}{\sigma_{s,\text{nr}}^2} = \frac{\sigma_s^2}{\mathbf{h}^H\mathbf{R}_s\mathbf{h}}, \tag{46}$$

or the signal distortion index [5]:

$$\nu_{sd} = \frac{\mathbb{E}\left((\widehat{s}(n) - s(n))^2\right)}{\mathbb{E}\left(s^2(n)\right)}. \tag{47}$$

If a filter is distortionless, it passes the desired signal through the filter without any modification. In such a case the signal reduction factor would be one and the signal distortion index would be zero.

It is also possible to measure the amount of noise reduction obtained by a filter using the noise reduction factor [5]:

$$\xi_{\mathrm{nr}} = \frac{\sigma_v^2}{\sigma_{v,\mathrm{nr}}^2} = \frac{\sigma_v^2}{\mathbf{h}^H \mathbf{R}_v \mathbf{h}}. \tag{48}$$

There is a close relation between the ratio of input SNR to output SNR and signal reduction to noise reduction [5]:

$$\frac{\mathrm{oSNR}}{\mathrm{iSNR}} = \frac{\xi_{\mathrm{nr}}}{\xi_{\mathrm{sr}}}. \tag{49}$$

Similar measures are derived in the frequency domain and for subspace based filtering in [7, 8, 10].

The perception of speech by human listeners is also important since there is not always a close relation between SNR, signal distortion and how the speech is perceived by listeners. For this, listening experiments can be performed. The listening tests have to be performed in a controlled environment like an acoustically damped booth made for the purpose so that no outside noise will interfere with the tests. The tests can be set up in different ways, but could, e.g., contain two different versions of the same sentence where the test subject has to choose the preferred one. This can be repeated for different test sentences to make sure that the trend is the same independent on what is being said. Further, for the experiment to be representable, a lot of test subjects need to perform the test which makes listening tests cumbersome to perform. This has motivated researchers to make objective performance measures that try to mimic how the speech would be rated by a human listener. Two of these methods are the perceptual evaluation of quality (PESQ) [121] score and short-time objective intelligibility (STOI) [138]. Both measures are performed based on the data in the frequency domain but the PESQ score includes more steps than STOI. STOI is a measure of the linear correlation between the clean speech and the processed speech. In PESQ the clean speech and processed speech are time aligned in each time frame. After this an auditory transform is performed on the data which is a psychoacoustic model mapping the signal into its perceived loudness. The difference in loudness between clean and processed speech is found and two different disturbances are obtained, one with, d_{ASYM}, and one without, d_{SYM}, an asymmetry factor. The two disturbances are combined into one performance measure as PESQ $= 4.5 - 0.1 d_{SYM} - 0.0309 d_{ASYM}$.

This mix of the two parameters is, however, optimised for speech transmitted over telephone networks and it was proposed in [67] to change the weighting to optimise the measure for speech enhancement algorithms.

5 Noise statistics and parameter estimation

For the noise driven speech enhancement approaches, information about the noise statistics are necessary to perform the speech enhancement. For time domain processing this would be a covariance matrix estimate and for frequency domain processing, the power spectral density (PSD) of the noise is estimated. Using a signal driven approach on the other hand, no information about the noise is needed, but dependent on the signal model some different signal parameters have to be estimated. For the harmonic model this is, e.g., the fundamental frequency and model order. Further, an estimate of the observed signal statistics is needed, either to use directly as in the spectral subtraction in (24), or as a means to estimate the statistics of the desired signal as is needed in, e.g., the Wiener filter in (32) and (34). If the desired signal and noise are not correlated, the statistics of the signals are additive:

$$\mathbf{R}_x = \mathbf{R}_s + \mathbf{R}_v \Leftrightarrow \tag{50}$$

$$\mathbf{R}_s = \mathbf{R}_x - \mathbf{R}_v, \tag{51}$$

and

$$|X(\omega)| = |S(\omega)| + |V(\omega)| \Leftrightarrow \tag{52}$$

$$|S(\omega)| = |X(\omega)| - |V(\omega)|. \tag{53}$$

In the time domain, it is common to estimate the covariance matrix of the observed signal by use of the sample covariance matrix estimate [22]:

$$\mathbf{R}_x = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{x}(n)\mathbf{x}^H(n), \tag{54}$$

where $M < N/2 + 1$ to ensure a full rank matrix which is important if it has to be inverted. This will produce an $M \times M$ matrix representing the covariance of M consecutive samples. However, the estimate is based on $N > 2(M-1)$ samples. This is no problem for stationary signals, but as speech is non-stationary this estimate can cause problems, especially in periods where the signal is fast varying, and it becomes an even bigger issue in the multichannel case. An alternative to the sample covariance matrix estimate is the covariance matrix generated by use of the iterative adaptive approach (IAA) [72, 122, 155]. This approach minimises a weighted least squares cost function in the attempt to estimate the spectral amplitudes of the signal. As an intrinsic part of the process, the covariance matrix of the signal is also estimated. The final estimates are obtained by iterating between estimating the covariance matrix based on a prior estimate of the amplitudes and estimating the amplitudes based on the obtained covariance matrix estimate. The IAA approach does not have the same restrictions on the relationship between M and N to make the covariance matrix full rank. Therefore, we may choose $M = N$ which makes the estimate more suited for fast varying signals. The drawback of the IAA approach is that it has a high computational complexity and is, therefore, too slow for some applications. However, work has been done to make the method more computationally effective by exploiting the structure of the covariance matrix and rewrite it in a framework where the discrete Fourier transform can be utilised [53–55].

5.1 Noise estimation

Much work has been invested in estimating the PSD of the noise. This is not a trivial problem and is still far from being solved. The initial way of solving this problem was to estimate the PSD in periods without speech found by use of voice activity detectors [117, 132, 141] and use this estimate in periods of speech as well. However, this is not optimal in the case of non-stationary noise since the estimated noise PSD will not be representative for the noise at a different point in time. This lead to the suggestion of methods to estimate the noise in presence of speech.

In [94, 95], the power is estimated in each frequency bin separately using minimum statistics. It is based on the observation that the power in a frequency bin at frequent points in time drops to the level of the noise. Since the spectrum fluctuates rapidly over time, it is smoothed and the minimum is found. A bias compensation is introduced to obtain the average noise level from the minimum. One of the challenges in this method is to find a good value for the smoothing parameter and the time window. Insufficient smoothing will give a high variance of the noise estimate whereas too much smoothing will smooth out the minima and give a wrong estimation. The window size is a trade-off between having speech free periods where the level decreases to the noise level and having the noise characteristics change too much within one time window. A typical time span is 400 ms to one second [91]. In [34], it was suggested to modify the method to track the noise continuously instead of within fixed time frames in order to react better to non-stationary noise.

Another suggested method is to update the noise estimate in a given frequency bin based on an estimate of the SNR [88]. The estimate of the power in the bin will then be a combination of the estimate prior in time and the present estimate, weighted according to the SNR. When the SNR is high, the estimate is primarily given by the former estimate whereas when the SNR is low, the estimate is primarily given by the present noise estimate. Alternatively, the weighting can be done according to the probability of speech being present [48].

In the multichannel case, it is not enough to estimate the PSD of the noise. Here, knowledge about the correlation between the noise in the different microphones is also needed. This makes the multichannel noise estimation much more difficult than in the single channel case [64].

5.2 Signal parameters

For the harmonic signal model presented earlier, the parameters to be estimated includes the fundamental frequency, the model order, and often times also the amplitudes and phases. For the two modified harmonic models also the chirp rate or the inharmonicities are needed.

The methods for parameter estimation can be divided into three groups [22]:

- Statistical methods
- Subspace methods
- Filtering methods

Statistical methods

As for the enhancement methods, these methods are based on a statistical model of the signal. A very common method is maximum likelihood (ML) [14, 21, 22, 36, 111] where a probability density function (PDF) of the observed data given the parameters are set up and maximised with respect to the wanted parameter(s). It is normal to assume that the data follows a Gaussian distribution. A further assumption is often that the noise is white Gaussian in which case the ML estimator of the fundamental frequency turns into the non-linear least squares (NLS) estimator. This estimator minimises the two norm between the observed signal and the signal model. It is also possible to estimate several parameters jointly using these methods as in [73] where the fundamental frequency and direction of arrival (DOA) are found jointly in a multi-channel setup. ML methods to jointly estimate fundamental frequency and chirp rate are presented in [26, 36]. Since the computational load of doing a combined search of two parameters instead of one increases dramatically, in [36] it is proposed to minimise a two step approximate cost function instead of the original one, that makes it possible to take advantage of the DFT whereas in [26] it is proposed to estimate the two parameters by iterating between optimising with respect to first the chirp rate and then the fundamental frequency until convergence of the original cost function. The fundamental frequency and the structured inharmonicity in pianos would also be possible to estimate jointly using these methods whereas the unstructured inharmonicities in human speech would lead to a search space with a too high dimensionality for the problem to be reasonable. In [108], a model containing both a chirp parameter and inharmonicities is presented. The model is approximated with a Taylor polynomial which makes the estimation easier, but the resulting signal will deviate from the true model, with the deviation getting bigger for higher harmonics. Another way to estimate the inharmonicities would be to estimate the fundamental frequency first since the inharmonicities in speech are localised around the harmonic frequencies, and afterwards estimate the inharmonicities iteratively. However, this might be the best way to solve the problem since the inharmonicities are not taken into account in the estimation of the fundamental frequency. For the multi-pitch estimation scenario where more than one source is present, it is proposed in [1] to exploit block sparsity. No prior information about the number of sources or the number of harmonics for each source is assumed which means that a model order estimate for each source can be obtained with the algorithm at the same time. The method minimises the two norm as was the case for the NLS method, but introduces a penalty to ensure that the result is sparse. The block sparsity approach has been extended to also estimate the fundamental frequency in the presence of inharmonicities [104] and in linear chirp signals [137].

Another very useful method is the maximum a posteriori (MAP). Whereas ML is maximising the probability of the data given the parameters, MAP maximises the probability of the parameters given the data. The MAP approach can be used for fundamental frequency and model order estimation, but can also be used to choose between different models such as the harmonic model, the harmonic chirp model and a noise-only model. Thereby, it can be used as a voiced/unvoiced detector. The MAP estimator is part of the Bayesian framework where also other parameter estimators are derived [29, 56, 103].

Subspace methods

The most well-known subspace parameter estimation methods are the multiple signal classification (MUSIC) method and estimation of signal parameters by rotational invariance technique (ESPRIT). Both methods can be used to estimate fundamental frequency and model order [22]. The methods build on the EVD and are both first derived for sensor arrays and DOA estimation but can easily be adapted to parameter estimation of harmonic signals [24, 25]. The MUSIC method [128] is based on the fact that the signal and noise subspaces are orthogonal. The $\mathbf{Z}$ matrix in the signal model can also be shown to be orthogonal to the noise subspace given the correct model order and fundamental frequency, and these parameters can, therefore, be found as those who maximise the orthogonality between $\mathbf{Z}$ and the noise subspace. In [150], a polynomial rooting method based on MUSIC is suggested for combined estimation of direction of arrival and range in order to avoid a two dimensional grid search. The first parameter is found by solving for the roots on a polynomial whereas the second parameter are found by a grid search. The rooting in MUSIC is used in connection with fundamental frequency estimation in [74], and could possibly be extended to the two parameter case as an alternative to the ML iterative method described in Section 5.2 to decrease computational complexity. The ESPRIT method [124] takes advantage of the shift invariance of the matrix $\mathbf{Z}$. Two submatrices of the signal subspace matrix are generated, one being a shifted version of the other. The difference between these two matrices, one multiplied with a shifting matrix, is minimised in order to find the desired parameter estimate. MUSIC and ESPRIT have also been extended to joint DOA and pitch estimation in the multichannel case [153, 154]. Another possibility is to use weighted least squares fitting used for direction of arrival estimation in [147, 148]. Here, the difference between a weighted version of the eigenvectors and the steering vector multiplied with a matrix, that is solved for during the derivation, is minimised. As with the subspace enhancement methods, one of the drawbacks of the methods is that the noise has to be white.

Filtering methods

The filtering methods are based on the signal driven filters introduced in Section 4.2. The filter is constructed to minimise the output power of the filter. After the generation of the filter, the parameters are found as the candidates maximising the output power since at the correct fundamental frequency a set of harmonics will pass through the filter and increase the output power [27, 75]. As with the enhancement, the first filter introduced in this group was the comb filter [99, 102]. The LCMV and APES filters described in the enhancement section can also be used for finding signal parameters. The LCMV filter gives a better spectral resolution than the APES filter [70]. The APES filter has a worse spectral resolution and the fundamental frequency estimate is biased. However, better estimates of the amplitudes and phases are obtained with the APES filter compared to the LCMV filter. Therefore, the CAPES method is proposed in [70] where a fundamental frequency estimate is first obtained using the LCMV filter, and then, based on this estimate, the amplitudes and phases of the sinusoids at the harmonic frequencies are found using APES. The filtering methods have also been extended to the multichannel case [76].

5.3 Cramér-Rao bound

The best possible estimation accuracy of unbiased estimators is set by the Cramér-Rao bound (CRB). This gives a measure of the average minimum variance of the estimates obtainable by any estimator. Therefore, when the estimates of an estimator on average hit the CRB, it is not possible for the estimator to perform any better. The CRB is dependent on the SNR and the segment length. For high SNR and long segment lengths, it is possible to obtain better estimates than at low SNR and short segment lengths.

The Cramér-Rao bound sets a lower limit to the variance of an unbiased estimator as [78]:

$$\mathrm{var}(\widehat{\theta}_g) \geq [\mathcal{I}^{-1}(\boldsymbol{\theta})]_{gg}, \tag{55}$$

where θ_g is the g'th parameter of the parameter vector $\boldsymbol{\theta}$ of length G, $[\cdot]_{gg}$ is the matrix element of row g and column g and $\mathcal{I}(\boldsymbol{\theta})$ is the Fisher information matrix (FIM):

$$\mathcal{I}(\boldsymbol{\theta}) = -\mathbb{E}\left(\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_g \partial\theta_h}\right), \tag{56}$$

where $p(\mathbf{x}|\boldsymbol{\theta})$ is the probability density function of the vector $\mathbf{x}$ given the parameters in the vector $\boldsymbol{\theta}$. The CRB is given under the assumption of the regularity condition

$$\mathbb{E}\left(\frac{\partial \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) = 0 \quad \forall\ \theta. \tag{57}$$

For a harmonic signal, the parameter vector is given as

$$\boldsymbol{\theta} = [\omega_0\ A_1\ \phi_1\ \ldots\ A_L\ \phi_L]. \tag{58}$$

If it is assumed that the covariance matrix does not depend on the parameters and the noise is white Gaussian, the FIM can be written as:

$$\mathcal{I}(\boldsymbol{\theta}) = \frac{2}{\sigma_v^2}\mathfrak{Re}\left(\frac{\partial \mathbf{s}(n,\boldsymbol{\theta})^H}{\partial \boldsymbol{\theta}}\frac{\partial \mathbf{s}(n,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right), \tag{59}$$

where $\mathfrak{Re}(\cdot)$ denotes the real part of the argument. This can be rewritten as:

$$\mathcal{I}(\boldsymbol{\theta}) = \frac{2}{\sigma_v^2}\mathfrak{Re}\left(\mathbf{D}^H(\boldsymbol{\theta})\mathbf{D}(\boldsymbol{\theta})\right), \tag{60}$$

where

$$\mathbf{D}(\boldsymbol{\theta}) = [\mathbf{d}(\omega_0)\ \mathbf{d}(A_1)\ \mathbf{d}(\phi_1)\ \ldots\ \mathbf{d}(A_L)\ \mathbf{d}(\phi_L)], \tag{61}$$

$$\mathbf{d}(\theta) = \frac{\partial \mathbf{s}(n,\boldsymbol{\theta})}{\partial \theta}. \tag{62}$$

Differentiating the signal vector $\mathbf{s}(n,\boldsymbol{\theta}) = \sum_{l=1}^{L} A_l e^{j\phi_l} e^{jl\omega_0 n}$ with respect to the parameters gives

$$[\mathbf{d}(\omega_0)]_n = \sum_{l=1}^{L} jlnA_l e^{j\phi_l} e^{jl\omega_0 n} \tag{63}$$

$$[\mathbf{d}(A_l)]_n = e^{j\phi_l} e^{jl\omega_0 n} \tag{64}$$

$$[\mathbf{d}(\phi_l)]_n = jA_l e^{j\phi_l} e^{jl\omega_0 n}. \tag{65}$$

The CRB can then easily be found by use of (61) and (55). This will give the exact bound, but finding the asymptotic bound can also be beneficial which gives an overview of the dependency of the bound on the different parameters and the segment length. Asymptotic bounds for the harmonic model can be found in [22]. The bounds for the harmonic chirp model and the inharmonic model can easily be found in a similar way by replacing the signal vector and making differentiations with respect to all parameters in the given model.

6 Contributions

The topic of this thesis is speech enhancement with a focus on voiced speech models. We have primarily exploited ways to use or extend the harmonic model in order to take deviations from the model into account. The deviations over frequency are modelled by the inharmonic model introduced in Paper D. The deviations over time come from the assumption of stationarity within the analysis frame. This problem we try to address by extending the harmonic model to a harmonic chirp model in Papers A, B and C, where we in Paper A and B explore the model in relation to speech enhancement and in Paper A look at parameter estimation and changing segment lengths. Paper E uses the traditional harmonic model but in combination with the IAA covariance matrix estimate where much shorter segments are necessary for the estimation and the stationarity assumption, therefore, has to be fulfilled in much shorter time frames. In Paper F, we exploit the harmonic model to make an estimate of the phase in the STFT domain and in Paper G, we make speech enhancement based on joint diagonalisation of desired signal and noise. The necessary statistics for desired signal and noise may be obtained by methods proposed in the preceding papers.

Paper A The first paper in this thesis is introducing the harmonic chirp model in relation to speech enhancement. The LCMV and APES filters are derived based on the harmonic chirp model and compared to the traditional harmonic model on synthetic signals and speech signals. It is shown that including the chirp parameter in the signal model gives better output SNR and PESQ score and less signal distortion.

Paper B This paper is exploring the harmonic chirp model further in relation to speech enhancement. The derivations of the LCMV and APES filters are done more in depth and further experimental analysis of the filters are performed in order to investigate their performance as a function of the input SNR, the segment length and the filter length. It is shown that including the chirp rate in the model gives a better output SNR and less distortion. As mentioned, the APES filter gives a noise covariance matrix estimate as a part of the filter design. The performance of the filter with this covariance matrix estimate is compared to the performance when a noise covariance matrix estimate based on the PSD is used. These two covariance matrices are also compared in a Winer filtering setup, and it is shown that the APES noise covariance matrix estimate also performs well in combination with the Wiener filter.

Paper C In this paper, we focus on the estimation of the fundamental frequency and chirp rate of the harmonic chirp model. We propose to use the maximum likelihood (ML) estimator because it reaches the CRB. In the paper we assume that the noise is white and Gaussian and, thereby, the ML estimator turns into the nonlinear least squares (NLS) estimator. Based on this estimator a two dimensional grid search is needed in order to find the minimum of the cost function. This is a computational heavy task, so we suggest to make an iterative search instead, initialised at the harmonic fundamental frequency and a chirp rate of zero. Further, since the change in fundamental frequency is not constant, we suggest to vary the segment lengths dependent on the characteristics of the signal. The MAP estimator is used to choose the best segment length at each instant, and the combination of segment lengths is found by backtracking. The results show that the harmonic chirp model gives rise to longer segments than the traditional harmonic model. Both NLS and MAP estimators are derived under the assumption of white Gaussian noise and, therefore, we also suggest two filters to prewhiten the signal.

Paper D In this paper, we investigate the deviations from the harmonic model over frequency by extending the harmonic model to take inharmonicities into account. The inharmonicity at each harmonic is estimated using NLS, the Capon (LCMV) filter and the APES filter, and the performance of the three estimation methods is compared to the CRB. The estimated perturbations are used to generate an APES filter which is compared to an APES filter based on the traditional harmonic assumption. The NLS and Capon filter gives good estimates of the inharmonicities and using the obtained estimates for enhancement in an APES filter increases the output SNR and lowers the signal distortion relative to the harmonic model.

Paper E The focus of this paper is multichannel speech enhancement using IAA covariance matrix estimates. In order to make the sample covariance matrix estimate full rank there is a restriction on the relation between the segment length and the size of the covariance matrix. Often there will also be a restriction on the number of microphones used which means that the time segments have to be very long. This can be a problem, especially if the fundamental frequency is fast varying. We suggest to make an estimate of the noise covariance matrix by subtracting the contribution at the harmonic frequencies from the IAA covariance matrix estimate of the observed signal and compare the result to the APES filter. For an equal number of samples, enhancement based on the IAA estimate outperforms the APES estimate.

Paper F In this paper, we move from the time domain to the frequency domain and suggest a least squares method to estimate the initial phases at the harmonic frequencies. We base the estimate on the method described in [81] where the phase evolution over time is estimated based on the harmonic model. To avoid an offset between the clean phase and the estimated phase, we suggest to estimate the initial phases by minimising the squared error between the noisy phase and the phase estimated in [81]. The error on the phase, and the mean squared error and PESQ score after reconstructing the signal using the clean amplitude all show better performance than the method in [81].

Paper G This paper looks at speech enhancement based on joint diagonalisation. The covariance matrices of desired signal and noise are jointly diagonalised and an estimate of the noise is obtained which is subtracted from the observed signal to give an estimate of the desired signal. The first filter is derived under the assumption of a rank deficient desired signal covariance matrix. The filter is generated to estimate the noise based on the eigenvectors corresponding to the least significant eigenvalues. For a rank deficient desired signal covariance matrix we get noise reduction without any signal distortion, but if some signal distortion is allowed, the filter works for other desired signals as well. The amount of noise reduction to signal distortion can be changed by choosing the number of eigenvectors included in the filter. It is shown that the gain in SNR and signal distortion are dependent on the number of eigenvectors used in the filter but independent of the input SNR.

The main part of the papers in this thesis are concerned with voiced speech signals that do not fulfil the traditional harmonic model. The problem is handled in different ways, either by modifying the model to make it fit the signal better or by changing the segment length to accommodate the underlying stationarity assumption of the harmonic model. In cases of an extended model, methods to estimate the extra parameters are proposed. The papers show that making the harmonic model more flexible or changing the segment length give better speech enhancement results in terms of output SNR and signal distortion.

Changing the model makes it fit better to the desired signal and, thereby, it gives rise to the reported increase in performance. It should however be investigated further how large the coupling between the non-stationarity and the inharmonicity is. Analysing non-stationary signals under the assumption of stationarity can give rise to similar spectral phenomena as inharmonicities, and a combined study of inharmonicities and non-stationarity would, therefore, be interesting. Further, the proposed iterative estimator of fundamental frequency and chirp rate shows that it is possible to expand the harmonic model to the harmonic chirp model without having to make a computational demanding two dimensional grid search for the parameters and still reach the CRB. This makes it more appealing to include an extra parameter in the model. Expanding the model to take non-stationarities into account raises the question of how non-stationary the signal is. This of course changes over time so a fixed segment length is not necessarily the best choice. The proposed segmentation method gives an alternative to fixed segment lengths which makes the model fit better within each segment and takes advantage of long segments, and the benefits it gives, whenever possible.

The speech enhancement methods based on the harmonic model have the drawback that they only work in periods of voiced speech. The voiced speech periods cover the largest part of the speech signal, but it would still be beneficial to have models covering the entire signal. It is very difficult to find a speech model covering both voiced and unvoiced speech since they are so different in the structure. Instead, more research in the combination of models for voiced and unvoiced speech models would be very relevant in the future.

Alternatively, speech enhancement can be performed as we did with the filters based on joint diagonalisation. These exploit that the covariance matrix of voiced speech has a rank corresponding to the number of harmonics, which is lower than the rank of the noise covariance matrix. One great advantage of these filters is that it is possible to trade off noise reduction and signal distortion without having to derive completely new filters. In the future, the performance of the filters can be increased by looking into the choice of eigenvectors included in the filter and make this choice signal dependent and changing from frame to frame. In periods of voiced speech, the optimal number of eigenvectors is closely coupled to the model order, and the performance of the filter might be increased by including an estimation of this parameter. We also showed that even though a signal driven approach is not taken in the design of the filter, it can still be beneficial to estimate the noise covariance matrix based on the signal statistics in voiced speech periods.

References

[1] S. I. Adalbjornsson, A. Jakobsson, and M. G. Christensen, “Estimating multiple pitches using block sparsity,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 6220–6224.

[2] W. W. Au, A. A. Pack, M. O. Lammers, L. M. Herman, M. H. Deakos, and K. Andrews, “Acoustic properties of humpback whale songs,” *J. Acoust. Soc. Am.*, vol. 120, no. 2, pp. 1103–1110, 2006.

[3] S. Bakamidis, M. Dendrinos, and G. Carayannis, “SVD analysis by synthesis of harmonic signals,” *IEEE Trans. Signal Process.*, vol. 39, no. 2, pp. 472–477, Feb. 1991.

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

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

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

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

[8] J. Benesty, J. Chen, and E. A. Habets, *Speech enhancement in the STFT domain.* Springer, 2012.

[9] J. Benesty, M. G. Christensen, J. R. Jensen, and J. Chen, “A brief overview of speech enhancement with linear filtering,” *EURASIP J. on Advances in Signal Processing*, vol. 2014, no. 1, pp. 1–10, 2014.

[10] J. Benesty, J. R. Jensen, M. G. Christensen, and J. Chen, *Speech Enhancement: A Signal Subspace Perspective.* Elsevier, 2014.

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

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

[13] G. Borden and K. Harris, *Speech science primer.* Williams & Wilkins, 1980.

[14] N. M. Botros and R. S. Adamjee, “Speech-pitch detection using maximum likelihood algorithm,” in *BMES/EMBS Conference*, vol. 2, Oct. 1999, p. 882.

[15] G. J. Brown and M. Cooke, “Computational auditory scene analysis,” *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, 1994.

[16] J. A. Cadzow, “Signal enhancement - a composite property mapping algorithm,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 1, pp. 49–62, Jan. 1988.

[17] ——, “SVD representation of unitarily invariant matrices,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 3, pp. 512–516, Jun. 1984.

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

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

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

[21] M. G. Christensen, “Multi-channel maximum likelihood pitch estimation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 409–412.

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

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

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

[25] ——, “Sinusoidal order estimation using angles between subspaces,” *EURASIP J. on Advances in Signal Processing*, vol. 2009, p. 62, 2009.

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

[27] M. G. Christensen, J. H. Jensen, A. Jakobsson, and S. H. Jensen, “Joint fundamental frequency and order estimation using optimal filtering.” *EURASIP J. on Advances in Signal Processing*, vol. 2011, p. 13, 2011.

[28] M. Cooke and D. P. Ellis, “The auditory organization of speech and other sources in listeners and computational models,” *Speech Communication*, vol. 35, no. 3, pp. 141–177, 2001.

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

[30] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals.* New York: Wiley, 2000.

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

[32] L. Deng and D. O’Shaughnessy, *Speech processing: a dynamic and optimization-oriented approach.* CRC Press, 2003.

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

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

[35] S. Doclo and M. Moonen, “GSVD-based optimal filtering for single and multi-microphone speech enhancement,” *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[36] Y. Doweck, A. Amar, and I. Cohen, “Joint model order selection and parameter estimation of chirps with harmonic components,” *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1765–1778, Apr. 2015.

[37] F. R. Drepper, “A two-level drive-response model of non-stationary speech signals,” *Nonlinear Analyses and Algorithms for Speech Processing*, vol. 1, pp. 125–138, Apr. 2005.

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

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

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

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

[42] N. W. Evans, J. S. Mason, W. M. Liu, and B. Fauve, “On the fundamental limitations of spectral subtraction: an assessment by automatic speech recognition,” in *Proc. European Signal Processing Conf.*, 2005, pp. 1–4.

[43] R. Frazier, S. Samsam, L. Braida, and A. V. Oppenheim, “Enhancement of speech by adaptive filtering,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 1976, pp. 251–253.

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

[45] N. D. Gaubitch and P. A. Naylor, *Speech Dereverberation.* Springer, 2010.

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

[47] T. Gerkmann, “Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase,” *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4199–4208, Aug. 2014.

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

[49] T. Gerkmann and M. Krawczyk, “MMSE-optimal spectral amplitude estimation given the STFT-phase,” *IEEE Signal Process. Lett.*, vol. 20, no. 2, pp. 129–132, 2013.

[50] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, “Phase processing for single-channel speech enhancement: History and recent advances,” *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, Mar. 2015.

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

[52] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, “Sparse linear prediction and its applications to speech processing,” *IEEE Trans. Audio, Speech and Language Process.*, vol. 20, no. 5, pp. 1644–1657, 2012.

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

[54] ——, “Efficient implementation of iterative adaptive approach spectral estimation techniques,” *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4154–4167, 2011.

[55] ——, “Superfast approximative implementation of the iaa spectral estimate,” *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 472–478, 2012.

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

[57] Z. Goh and K.-C. Tan, “Postprocessing method for suppressing musical noise generated by spectral subtraction,” *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp. 287–292, May 1998.

[58] S. Greenberg, W. A. Ainsworth, and R. R. Fay, Eds., *Speech Processing in the Auditory System.* Springer, 2004.

[59] D. Griffin and J. S. Lim, “Signal estimation from modified short-time fourier transform,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr 1984.

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

[61] P. C. Hansen and S. H. Jensen, “Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis,” *EURASIP J. on Advances in Signal Processing*, vol. 2007, no. 1, p. 24, Jun. 2007.

[62] ——, “Prewhitening for rank-deficient noise in subspace methods for noise reduction,” *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3718–3726, 2005.

[63] P. Hedelin, “A glottal LPC-vocoder,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 9, Mar. 1984, pp. 21–24.

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

[65] G. Hu and D. Wang, “Segregation of unvoiced speech from nonspeech interference,” *J. Acoust. Soc. Am.*, vol. 124, no. 2, pp. 1306–1319, 2008.

[66] Y. Hu and P. Loizou, “A generalized subspace approach for enhancing speech corrupted by colored noise,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 334–341, Jul. 2003.

[67] ——, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[68] J. M. Huerta and R. M. Stern, "Distortion-class weighted acoustic modeling for robust speech recognition under GSM RP-LTP coding," in *Proc. of the robust methods for speech recognition in adverse conditions*, 1999.

[69] A. Jakobsson, T. Ekman, and P. Stoica, "Capon and APES spectrum estimation for real-valued signals," *Eighth IEEE Digital Signal Processing Workshop*, 1998.

[70] A. Jakobsson and P. Stoica, "Combining Capon and APES for estimation of spectral lines," *Circuits, Systems and Signal Processing*, vol. 19, pp. 159–169, Mar. 2000.

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

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

[73] ——, "Nonlinear least squares methods for joint DOA and pitch estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 5, pp. 923–933, 2013.

[74] ——, "Fundamental frequency estimation using polynomial rooting of a subspace based method," *Proc. European Signal Processing Conf.*, pp. 502–506, 2010.

[75] ——, "Joint DOA and fundamental frequency estimation methods based on 2-d filtering," *Proc. European Signal Processing Conf.*, pp. 2091–2095, 2010.

[76] J. R. Jensen, M. G. Christensen, J. Benesty, and S. H. Jensen, "Joint spatio-temporal filtering methods for DOA and fundamental frequency estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 174–185, 2015.

[77] J. R. Jensen, M. G. Christensen, and J. Benesty, "Multichannel signal enhancement using non-causal, time-domain filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2013, pp. 7274–7278.

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

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

[80] B. Koo, J. D. Gibson, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 1989, pp. 349–352.

[81] M. Krawczyk and T. Gerkmann, "STFT phase improvement for single channel speech enhancement," in *International Workshop on Acoustic Signal Enhancement; Proceedings of IWAENC 2012;*, Sept 2012, pp. 1–4.

[82] ——, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1931–1940, 2014.

[83] T. Kronvall, S. I. Adalbjornsson, and A. Jakobsson, "Joint doa and multi-pitch estimation using block sparsity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2014, pp. 3958–3962.

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

[85] J. Le Roux, E. Vincent, Y. Mizuno, H. Kameoka, N. Ono, and S. Sagayama, "Consistent wiener filtering: Generalized time-frequency masking respecting spectrogram consistency," in *Latent Variable Analysis and Signal Separation.* Springer Berlin Heidelberg, 2010, pp. 89–96.

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

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

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

[89] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1 – 15, 1997.

[90] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (nss), hidden markov models and the projection, for robust speech recognition in cars," *Speech Communication*, vol. 11, no. 2, pp. 215 – 228, 1992.

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

[92] Y. Lu and P. Loizou, "A geometric approach to spectral subtraction," *Speech Communication*, vol. 50, no. 6, pp. 453–466, 2008.

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

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

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

[96] R. Martin, I. Wittke, and P. Jax, "Optimized estimation of spectral parameters for the coding of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, Jun. 2000, pp. 1479–1482.

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

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

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

[100] M. N. Murthi and B. D. Rao, “Towards a synergistic multistage speech coder,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 1998, pp. 369–372.

[101] V. K. Murthy, L. J. Haywood, J. Richardson, R. Kalaba, S. Salzberg, G. Harvey, and D. Vereeke, “Analysis of power spectral densities of electrocardiograms,” *Mathematical Biosciences*, vol. 12, pp. 41 – 51, 1971.

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

[103] J. K. Nielsen, M. G. Christensen, and S. H. Jensen, “Default bayesian estimation of the fundamental frequency,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 3, pp. 598–610, 2013.

[104] T. Nilsson, S. I. Adalbjornsson, N. R. Butt, and A. Jakobsson, “Multi-pitch estimation of inharmonic signals,” in *Proc. European Signal Processing Conf.*, 2013, pp. 1–5.

[105] A. V. Oppenheim and J. S. Lim, “The importance of phase in signals,” *Proc. IEEE*, vol. 69, no. 5, pp. 529–541, May 1981.

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

[107] K. Paliwal, K. Wójcicki, and B. Shannon, “The importance of phase in speech enhancement,” *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011.

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

[109] ——, “Iterative estimation of sinusoidal signal parameters,” *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 461–464, Feb. 2010.

[110] ——, “On the properties of a time-varying quasi-harmonic model of speech.” in *Proc. Interspeech*, 2008, pp. 1044–1047.

[111] T. W. Parks and J. D. Wise, “Maximum likelihood pitch estimation,” in *IEEE Conference on Decision and Control*, 1977.

[112] F. Plante, G. F. Meyer, and W. A. Ainsworth, “A pitch extraction reference database,” in *Proc. Eurospeech*, Sep. 1995, pp. 837–840.

[113] T. Poulsen, “Acoustic communication, hearing and speech,” DTU Electrical Engineering - Acoustic Technology, 2008.

[114] P. Prandoni, M. M. Goodwin, and M. Vetterli, “Optimal time segmentation for signal modeling and compression,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1997, pp. 2029–2032.

[115] P. Prandoni and M. Vetterli, "R/D optimal linear prediction," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 646–655, 2000.

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

[117] L. Rabiner and M. Sambur, "Voiced-unvoiced-silence detection using the itakura lpc distance measure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 1977, pp. 323–326.

[118] S. S. Rao and D. C. Gnanaprakasam, "A criterion for identifying dominant singular values in the SVD based method of harmonic retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, Apr. 1988, pp. 2460–2463.

[119] R. A. Rasch and V. Heetvelt, "String inharmonicity and piano tuning," *Music Perception: An Interdisciplinary Journal*, vol. 3, no. 2, pp. 171–189, Winter 1985.

[120] L. Riecke, M. Vanbussel, L. Hausfeld, D. Baskent, E. Formisano, and F. Esposito, "Hearing an illusory vowel in noise: Suppression of auditory cortical activity," *The Journal of Neuroscience*, 2012.

[121] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2001, pp. 49–752.

[122] W. Roberts, P. Stoica, J. Li, T. Yardibi, and F. A. Sadjadi, "Iterative adaptive approaches to mimo radar imaging," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 1, pp. 5–20, Feb. 2010.

[123] M. Rothenberg, "A multichannel electroglottograph," *Journal of Voice*, vol. 6, no. 1, pp. 36–43, 1992.

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

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

[126] L. L. Scharf, "The SVD and reduced rank signal processing," *Signal Processing*, vol. 25, no. 2, pp. 113 – 133, 1991.

[127] L. L. Scharf and D. W. Tufts, "Rank reduction for modeling stationary signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 3, pp. 350–355, 1987.

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

[129] B. Schwartz, S. Gannot, E. Habets *et al.*, "Online speech dereverberation using kalman filter and em algorithm," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 2, pp. 394–406, 2015.

[130] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions,"

IEEE/ACM Trans. Audio, Speech, Language Process., vol. 23, no. 2, pp. 240–251, Feb. 2015.

[131] J. Skoglund, “Analysis and quantization of glottal pulse shapes,” *Speech Communication*, vol. 24, no. 2, pp. 133–152, 1998.

[132] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Process. Lett.*, vol. 6, no. 1, Jan. 1999.

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

[134] P. Stoica and P. Babu, “The gaussian data assumption leads to the largest cramér-rao bound [lecture notes],” *IEEE Signal Process. Mag.*, vol. 28, no. 3, pp. 132–133, 2011.

[135] D. Stowell and M. D. Plumbley, “Framewise heterodyne chirp analysis of birdsong,” *Proc. European Signal Processing Conf.*, pp. 694–2698, Aug 2012.

[136] N. Sturmel and L. Daudet, “Signal reconstruction from STFT magnitude: a state of the art,” *International Conference on Digital Audio Effects (DAFx)*, pp. 375–386, 2011.

[137] J. Sward, J. Brynolfsson, A. Jakobsson, and M. Hansson-Sandsten, “Sparse semi-parametric chirp estimation,” in *Rec. Asilomar Conf. Signals, Systems, and Computers*, Nov. 2014, pp. 1236–1240.

[138] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Trans. Audio, Speech and Language Process.*, vol. 9, no. 17, pp. 2125–2136, 2011.

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

[140] T. Takiguchi and Y. Ariki, “PCA-based speech enhancement for distorted speech recognition,” *Journal of Multimedia*, vol. 2, no. 5, pp. 13–18, Sep 2007.

[141] S. G. Tanyer and H. Ozer, “Voice activity detection in nonstationary noise,” *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 478–482, Jul. 2000.

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

[143] I. R. Titze, *Principles of Voice Production.* Prentice Hall, Inc., 1994.

[144] Z. Tüske, F. R. Drepper, and R. Schlüter, “Non-stationary signal processing and its application in speech recognition,” *Workshop on statistical and perceptual audition*, Sep. 2012.

[145] Z. Tüske, P. Golik, R. Schlüter, and F. R. Drepper, “Non-stationary feature extraction for automatic speech recognition,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 5204–5207, May 2011.

[146] P. P. Vaidyanathan, *The Theory of Linear Prediction.* Morgan & Claypool Publishers, 2008.

[147] M. Viberg and B. Ottersten, “Sensor array processing based on subspace fitting,” *IEEE Trans. Signal Process.*, vol. 39, no. 5, pp. 1110–1121, 1991.

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

[149] D. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 4, pp. 679–681, 1982.

[150] A. J. Weiss and B. Friedlander, "Range and bearing estimation using polynomial rooting," *IEEE J. Ocean. Eng.*, vol. 18, no. 2, pp. 130–137, 1993.

[151] M. Weiss, E. Aschkenasy, and T. Parsons, "Study and development of the INTEL technique for improving speech intelligibility," Nicolet Scientific Corporation, Northvale, NJ, Tech. Rep., Apr. 1975.

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

[153] Y. Wu, L. Amir, J. R. Jensen, and G. Liao, "Joint pitch and doa estimation using the esprit method," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 32–45, 2015.

[154] J. Xi Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, "Joint DOA and multi-pitch estimation based on subspace techniques," *EURASIP J. on Advances in Signal Processing*, vol. 2012, no. 1, 2012.

[155] T. Yardibi, J. Li, P. Stoica, M. Xue, and A. B. Baggeroer, "Source localization and sensing: A nonparametric iterative adaptive approach based on weighted least squares," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 1, pp. 425–443, Jan. 2010.

Part II

Papers

Paper A

Enhancement of Non-Stationary Speech using Harmonic Chirp Filters

Sidsel Marie Nørholm, Jesper Rindom Jensen and Mads Græsbøll Christensen

The paper has been accepted for publication in the
Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2015.

Abstract

In this paper, the issue of single channel speech enhancement of non-stationary voiced speech is addressed. The non-stationarity of speech is well known, but state of the art speech enhancement methods assume stationarity within frames of 20–30 ms. We derive optimal distortionless filters that take the non-stationarity nature of voiced speech into account via linear constraints. This is facilitated by imposing a harmonic chirp model on the speech signal. As an implicit part of the filter design, the noise statistics are also estimated based on the observed signal and parameters of the harmonic chirp model. Simulations on real speech show that the chirp based filters perform better than their harmonic counterparts. Further, it is seen that the gain of using the chirp model increases when the estimated chirp parameter is big corresponding to periods in the signal where the instantaneous fundamental frequency changes fast.

Index Terms: speech enhancement, single-channel, non-stationary signals, harmonic chirp model.

1 Introduction

Speech enhancement is important in many systems such as mobile phones, hearing aids and teleconferencing systems where the desired signal is corrupted by noise. Speech enhancement can be approached in different ways, common ones being spectral subtraction [1, 2] performed in the frequency domain or Wiener filtering performed in the frequency or time domain [2, 3]. These, and most other speech enhancement methods, assume that the signal is stationary within an analysis window, for speech this window is often assumed to be 20–30 ms.

Often, a noise driven approach is taken to speech enhancement where the power spectral density is estimated after transformation to the frequency domain. This can be done in speech free periods using a voice activity detector (VAD) [4] and extrapolating to periods with speech. In [5], this is expanded to also include new calculations in short speech pauses and brief breaks in between words, and in [6] the VAD is substituted with a speech probability, but, still, the noise estimation relies primarily on speech pauses. Therefore, the noise has to be stationary for longer periods than 20–30 ms in order for these methods to work properly. Alternatively, a signal driven approach can be taken where a model for the desired signal is assumed. An often used model is the harmonic model. Here, the signals, speech and noise, are assumed stationary within the window of 20–30 ms. However, this assumption is not fulfilled [7] since the speech signal is non-stationary and varies continuously over time.

Speech enhancement of non-stationary speech is not well covered in the literature, but the issue of non-stationary speech is introduced in related fields.

In [8, 9] a fan-chirp transform is suggested as an alternative to the traditional Fourier transform to analyse harmonic signals. The frequency is here allowed to vary linearly over time, leading to more sharp peaks in the spectrum when applied to a speech signal. Also in the field of speech recognition, non-stationary speech is taken into consideration by using gammachirp filters instead of traditional gammatone filters [10, 11], making the methods more robust to noise. In parameter estimation, a harmonic model extended to take non-stationarity into account has been considered in [12, 13]. In [12], the basis is a very flexible model including both a chirp parameter to take changes over time into account and a detuning parameter which can account for individual variations away from the harmonic frequencies. The model is then approximated with a Taylor polynomial which leads to bigger and bigger deviations from the original model as the harmonic number increases, as is also mentioned in the paper. In [13], a harmonic chirp model is used to describe the voiced speech signal. This model has a harmonic structure, but the instantaneous fundamental frequency is allowed to change linearly within each segment, making the model capable of coping with non-stationary speech. The focus in these papers is, however, not on speech enhancement.

In this paper, we investigate the harmonic chirp model used in [13] in relation to speech enhancement. The model is compared to the traditional harmonic model [14], a common model used to describe voiced speech (see, e.g., [15–17]) which is the major component of speech signals. Voiced/unvoiced detectors [18] make it possible to discriminate voiced and unvoiced parts and only use the model on the relevant parts. The unvoiced parts can then be filtered by, e.g., a Wiener filter. The harmonic model assumes that the desired signal is composed of a set of sinusoids having frequencies given by an integer multiple of a fundamental frequency. In the traditional harmonic model, the fundamental frequency is assumed constant in segments of 20–30 ms, whereas the harmonic chirp model allows the fundamental frequency to vary linearly within each segment by introducing a chirp parameter in the model. In the harmonic framework, signals are often filtered by use of the Linearly Constrained Minimum Variance (LCMV) filter or the Amplitude and Phase EStimation (APES) based filter [14, 17, 19]. The principle in these filters is to pass the desired signal undistorted while the noise is reduced as much as possible. We derive the LCMV and APES based filters using the harmonic chirp model and compare their performance on synthetic and real speech signals to similar filters based on the traditional harmonic model. As a part of the derivation of the APES based filter, a noise covariance matrix estimate is obtained which takes the non-stationarity of speech into account.

In Section 2, the harmonic chirp model is introduced, in Section 3, the LCMV and APES based filters are derived according to the harmonic chirp model, and, in Section 5, their performance is compared to similar filters based on the harmonic model. The paper is concluded in Section 6.

2 Harmonic Chirp Model

Often it is assumed that the desired signal is stationary within blocks of 20-30 ms. In such a framework a normally used model for voiced speech is the harmonic signal model. However, the assumption of stationarity does not hold since the frequencies of the harmonics are changing continuously over time. Therefore, we here suggest to use a model which does not assume stationarity but instead assumes that the harmonic frequencies change linearly within one of these short segments. This can be done by using a linear chirp model and the instantaneous frequency of the l'th harmonic, ω_l, can then be expressed as:

$$\omega_l(n) = l(\omega_0 + kn), \tag{A.1}$$

for time indices $n = 0, \cdots, N-1$ where ω_0 is the normalised fundamental frequency and k is the fundamental chirp rate. The instantaneous phase, θ_l, of the harmonic components of the speech signal is given by the integral of the instantaneous frequency:

$$\theta_l(n) = l\left(\omega_0 n + \frac{1}{2}kn^2\right) + \phi_l \tag{A.2}$$

where ϕ_l is the initial phase of the l'th harmonic. Thereby, the harmonic chirp model can be expressed by:

$$s(n) = \sum_{l=1}^{L} \alpha_l e^{jl(\omega_0 n + k/2n^2)}, \tag{A.3}$$

where L is the number of harmonics, and the initial phase is included in the amplitude term to give the complex amplitude of the l'th harmonic, $\alpha_l = A_l e^{j\phi_l}$, with $A_l > 0$ being the real amplitude. We choose to work in the complex domain since this leads to simpler expressions. A real signal can be transformed to a complex signal by use of the Hilbert transform, and back again by only considering the real part of the complex signal.

We are looking at the case where the desired signal, $s(n)$, is corrupted by noise, $v(n)$, to give the observed signal, $x(n)$,

$$x(n) = s(n) + v(n). \tag{A.4}$$

The signal and noise are assumed uncorrelated and, therefore, we have that the variance of the observed signal is the sum of the variances of desired signal and noise, $\sigma_x^2 = \sigma_s^2 + \sigma_v^2$.

The enhancement problem considered in this paper is then to get a good estimate of the desired signal, $\widehat{s}(n)$, based on filtering of the observed signal

$$\widehat{s}(n) = \mathbf{h}^H \mathbf{x}(n) = \mathbf{h}^H \mathbf{s}(n) + \mathbf{h}^H \mathbf{v}(n), \tag{A.5}$$

where $\mathbf{h} = [h(0)\, h(1)\, \cdots\, h(M-1)]^H$ is the filter with length M, $\mathbf{x}(n) = [x(n)\; x(n+1)\; \cdots\; x(n+M-1)]^T$, $\mathbf{v}(n)$ and $\mathbf{s}(n)$ are defined in a similar way to $\mathbf{x}(n)$ and $\{\cdot\}^T$ ($\{\cdot\}^H$) denotes the (Hermitian) transpose. Again, under the assumption of uncorrelated signals, we have that $\sigma^2_{\widehat{s}} = \sigma^2_{x,\mathrm{nr}} = \sigma^2_{s,\mathrm{nr}} + \sigma^2_{v,\mathrm{nr}}$, where $\sigma^2_{x,\mathrm{nr}} = \mathbf{h}^H \mathbf{R}_x \mathbf{h}$ is the variance of the observed signal after noise reduction, and similar for $\sigma^2_{s,\mathrm{nr}}$ and $\sigma^2_{v,\mathrm{nr}}$.

3 Filters

One filter that can be used for extracting harmonic signals is the LCMV filter [14] which is minimising the output power of the filter while passing the desired signal according to the signal model undistorted. This filter can be modified to fit harmonic chirp signals instead and is then the solution to the optimisation problem:

$$\min_{\mathbf{h}} \mathbf{h}^H \mathbf{R}_x \mathbf{h}, \quad \text{s.t.} \quad \mathbf{h}^H \mathbf{Z} = \mathbf{1}^T, \tag{A.6}$$

where $\mathbf{1} = [1 \cdots 1]^T$, $\mathbf{R}_x$ is the covariance matrix of the observed signal defined as:

$$\mathbf{R}_x = \mathbf{E}\{\mathbf{x}(n)\mathbf{x}^H(n)\}, \tag{A.7}$$

with $\mathbf{E}\{\cdot\}$ denoting statistical expectation, and $\mathbf{Z}$ is constructed from a set of modified Fourier vectors:

$$\mathbf{Z} = [\mathbf{z}(\omega_0, k)\, \mathbf{z}(2\omega_0, 2k)\, \cdots\, \mathbf{z}(L\omega_0, Lk)], \tag{A.8}$$

with

$$\mathbf{z}(l\omega_0, lk) = \begin{bmatrix} 1 \\ e^{jl(\omega_0 + k/2)} \\ \vdots \\ e^{jl(\omega_0(M-1) + k/2(M-1)^2)} \end{bmatrix}. \tag{A.9}$$

The solution to the minimisation problem is:

$$\mathbf{h} = \mathbf{R}_x^{-1}\mathbf{Z}(\mathbf{Z}^H \mathbf{R}_x^{-1}\mathbf{Z})^{-1}\mathbf{1}. \tag{A.10}$$

The harmonic LCMV filter is a special case of this filter for $k = 0$, and in this case the problem reduces to the one in [14].

In practice the covariance matrix is not known but has to be estimated. This is often done by use of the sample covariance estimate

$$\widehat{\mathbf{R}}_x = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{x}(n)\mathbf{x}^H(n). \tag{A.11}$$

However, in this estimate it is assumed that the signal is stationary over the set of N samples. This is not the case when non-stationary speech is considered. Therefore, we also suggest a modification of the APES based filter [17]. As a part of the design of this filter, an estimate of the noise covariance matrix is generated. This is done by subtracting the part coming from the desired signal from the covariance matrix of the observed signal. By modifying this filter it will be possible to obtain a noise covariance matrix which is independent of the part of the desired signal aligning with the chirp signal model.

The APES based filter is the solution to the mean squared error (MSE) between the filtered signal and the signal model:

$$\text{MSE} = \frac{1}{N-M+1} \sum_{n=0}^{N-M} |\mathbf{h}^H \mathbf{x}(n) - \mathbf{a}^H \mathbf{w}(n)|^2, \tag{A.12}$$

where $\mathbf{a} = [\alpha_1 \, \alpha_2 \, \cdots \, \alpha_L]^H$ and

$$\mathbf{w}(n) = \begin{bmatrix} e^{j(\omega_0 n + k/2n^2)} \\ e^{j2(\omega_0 n + k/2n^2)} \\ \vdots \\ e^{jL(\omega_0 n + k/2n^2)} \end{bmatrix}. \tag{A.13}$$

The solution to this minimisation, under the same constraint as in (A.6), is given by:

$$\mathbf{h} = \mathbf{Q}^{-1}\mathbf{Z}(\mathbf{Z}^H\mathbf{Q}^{-1}\mathbf{Z})^{-1}\mathbf{1} \tag{A.14}$$

with

$$\mathbf{Q} = \widehat{\mathbf{R}}_x - \mathbf{G}^H\mathbf{W}^{-1}\mathbf{G}, \tag{A.15}$$

$$\mathbf{G} = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{w}(n)\mathbf{x}^H(n), \tag{A.16}$$

and

$$\mathbf{W} = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{w}(n)\mathbf{w}^H(n). \tag{A.17}$$

The LCMV filter in (A.10) and the APES based filter in (A.14) look very similar. The difference between the two filters is that the LCMV filter uses the covariance matrix of the observed signal, $\mathbf{R}_x$, whereas the covariance matrix used in the APES based filter, $\mathbf{Q}$, can be seen as an estimate of the noise covariance matrix.

4 Simulations

The two new harmonic chirp filters are compared to the harmonic LCMV [14] and APES based [17] filters. These filters are special cases of the harmonic chirp filters and are obtained by setting $k = 0$. The performance is measured by means of the output signal-to-noise ratio (oSNR),

$$\text{oSNR}(\mathbf{h}) = \frac{\sigma^2_{s,\text{nr}}}{\sigma^2_{v,\text{nr}}} = \frac{\mathbf{h}^H \mathbf{R}_s \mathbf{h}}{\mathbf{h}^H \mathbf{R}_v \mathbf{h}}, \tag{A.18}$$

where $\mathbf{R}_s$ and $\mathbf{R}_v$ are the covariance matrices of desired signal and noise, and the signal reduction factor,

$$\xi_{\text{sr}}(\mathbf{h}) = \frac{\sigma^2_s}{\sigma^2_{s,\text{nr}}} = \frac{\sigma^2_s}{\mathbf{h}^H \mathbf{R}_s \mathbf{h}}. \tag{A.19}$$

The output SNR should be as high as possible whereas the signal reduction factor should be as close to one as possible to avoid signal distortion.

The filters were first tested on synthetic harmonic chirp signals made according to (A.3) through Monte Carlo simulations (MCS) [20]. The signals were generated with $L = 10$, $A_l = 1 \,\forall\, l$, random phases, fundamental frequency and fundamental chirp rate in the intervals: $\phi_l \in [0, 2\pi]$, $f_0 \in [150, 250]$ Hz, $k \in [0, 200]$ Hz^2. The signals were added white Gaussian noise with a variance calculated to fit the desired input SNR,

$$\text{iSNR} = \frac{\sigma^2_s}{\sigma^2_v}. \tag{A.20}$$

The signal and segment length were set to $N = 200$ and the filter length $M = 50$. The output SNR and signal reduction factor of the filter were calculated for each realisation of the chirp signal and averaged over 500 MCSs.

In Figs. B.1a and B.1b the output SNR and signal reduction factor are shown as a function of the input SNR. Five filters are compared in the figures. LCMV$_{\text{opt}}$ is a chirp LCMV filter with the covariance matrix estimated directly from the noise signal, and, therefore, it sets an upper limit for the performance of the filters but cannot be used in practice where there is no access to the clean noise signal. The other two LCMV filters are the chirp LCMV (LCMV$_{\text{c}}$) and the harmonic LCMV (LCMV$_{\text{h}}$) and likewise with the two APES based filters, APES$_{\text{c}}$ and APES$_{\text{h}}$. The two APES based filters perform better than the corresponding LCMV filters, and the two chirp based filters perform better than their harmonic counterparts. At low SNRs all filters perform almost equally, but when the input SNR is increased, the output SNR of the optimal LCMV filter and the chirp APES based filter increases almost linearly whereas the output SNR of the other three filters falls off. The signal reduction factor

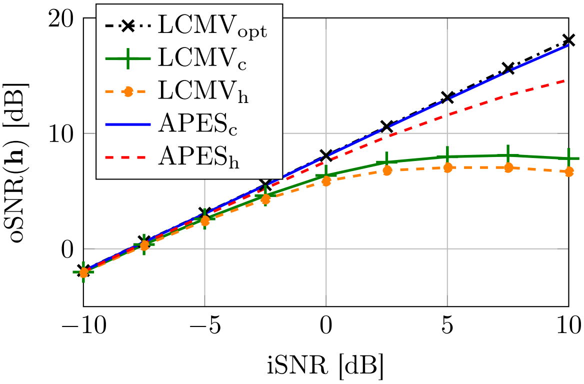


Fig. A.1: Output SNR as a function of the input SNR for a synthetic chirp signal.

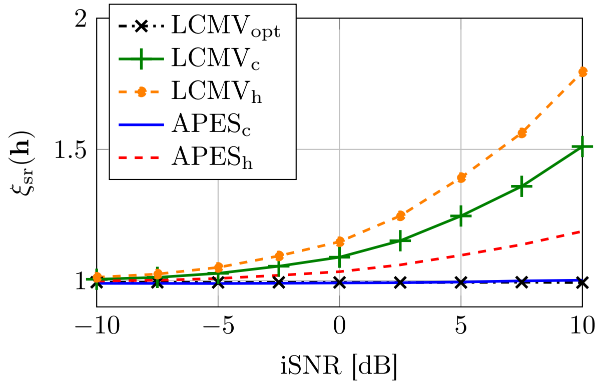


Fig. A.2: Signal reduction factor, $\xi_{\text{sr}}(\mathbf{h})$, as a function of the input SNR for a synthetic chirp signal.

for the optimal LCMV and the chirp APES based filter is very close to one for all input SNRs whereas it increases with input SNR for the other filters.

The filters are next evaluated on a speech signal. The signal is a female speaker uttering the sentence "Why were you away a year, Roy?" sampled at $f_s = 8000$ Hz. To evaluate the potential of the methods, and since the focus is here on enhancement and not parameter estimation, the fundamental frequency, fundamental chirp rate and number of harmonics are estimated on the clean speech signal using nonlinear least squares (NLS) estimators [13, 14]. Again the noise is white Gaussian and added to give the desired input SNR.

The output SNR over time is shown in Fig. A.3 for an input SNR of 10 dB. Except for very few points in time, the chirp APES based filter is seen to set an upper limit to the performance of the four filters. The same tendency as for the synthetic signal is seen, with the APES based filters giving a higher output SNR than the LCMV filters and the chirp versions performing better

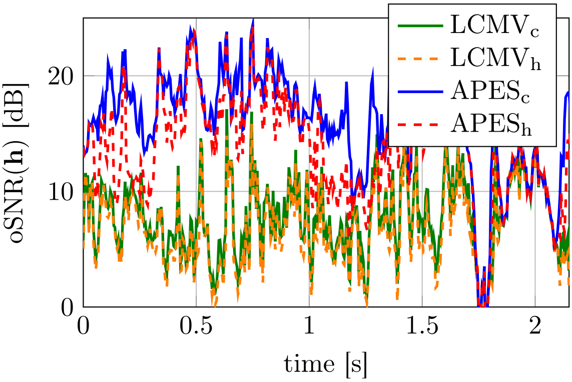


Fig. A.3: Output SNR over time for a speech signal with input SNR = 10 dB.

than the harmonic ones. The difference in output SNR for the two APES based filters, $\text{oSNR}_\Delta = \text{oSNR}(\text{APES}_\text{c}) - \text{oSNR}(\text{APES}_\text{h})$ is compared to the absolute value of the fundamental chirp rate in Fig. A.4. Here, it is again seen that, except for a few places with small negative differences, the difference is positive, meaning that the chirp APES based filter gives a higher output SNR than the harmonic APES based filter. In the figure it is also seen that the gain obtained by using the chirp APES based filter instead of the harmonic APES based filter is closely related to the estimated chirp parameter. When the absolute value of the chirp parameter is big, a gain in the oSNR is obtained whereas the gain is close to zero when the chirp parameter is close to zero. This makes sense if the harmonic chirp model describes the speech signal better than the harmonic model. If the fundamental frequency de- or increases a lot in one segment of the signal, the chirp parameter will have a large absolute value, and the difference between the harmonic and harmonic chirp model will be large, and, thereby, there will be an advantage in using the harmonic chirp model. If the fundamental frequency is almost constant in a segment, the chirp parameter will be close to zero and the chirp harmonic model reduces to the harmonic model, leading to similar output SNRs for the two models.

In Figs. A.5-A.7 the output SNR, signal distortion and Perceptual Evaluation of Speech Quality (PESQ) score [21] are shown as a function of the input SNR. The results are averaged over 50 Monte Carlo simulations. Here it is seen that the speech signal follows the same tendencies as the synthetic signal. The output SNRs of the filters are very similar to the output SNRs in the synthetic case, however, the signal distortion is increased for all filters, but the chirp APES based filter still has the lowest distortion. Also in terms of PESQ score the same conclusions can be drawn. The chirp filters perform better than their harmonic counterparts.

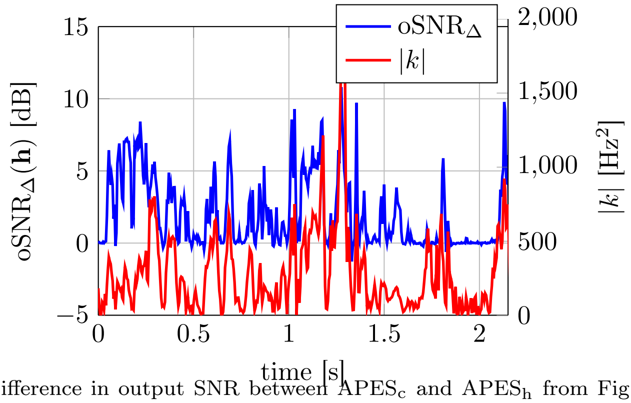


Fig. A.4: Difference in output SNR between APES$_\mathrm{c}$ and APES$_\mathrm{h}$ from Fig. A.3, oSNR$_\Delta$, and the estimated chirp parameter, $|k|$.

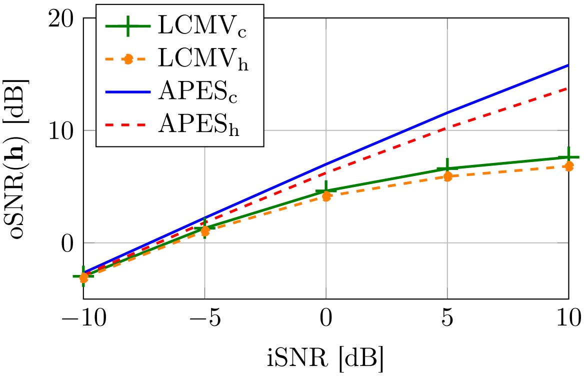


Fig. A.5: Output SNR as a function of the input SNR for a speech signal.

5 Conclusions

In this paper, the non-stationarity of speech is taken into account to increase the performance of enhancement filters. The voiced speech was described with a harmonic chirp model and two filters based on the Linearly Constrained Minimum Variance (LCMV) filter and Amplitude and Phase EStimation (APES) based filter were presented and compared to their harmonic counterparts. It was shown that the chirp based filters perform better in terms of output SNR, signal distortion and PESQ score. As part of the derivation of the chirp APES based filter, a noise covariance matrix estimate is generated which can be used in other filters as, e.g., the Wiener filter.

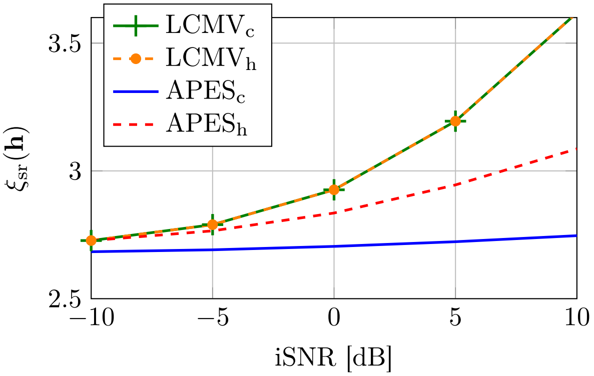


Fig. A.6: Signal reduction factor, $\xi_{sr}(\mathbf{h})$, as a function of the input SNR for a speech signal.

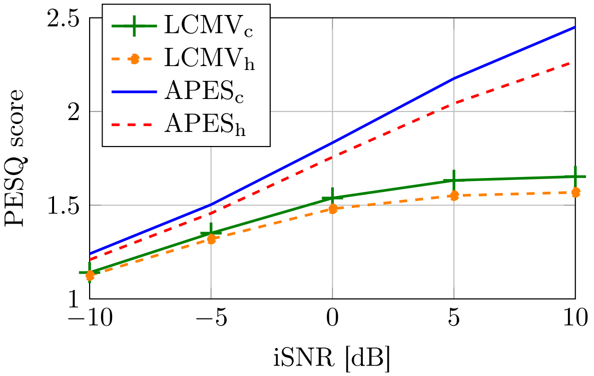


Fig. A.7: PESQ score as a function of the input SNR for a speech signal.

References

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

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

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

[4] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, Jan. 1999.

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

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

[7] F. R. Drepper, “A two-level drive-response model of non-stationary speech signals,” *Nonlinear Analyses and Algorithms for Speech Processing*, vol. 1, pp. 125–138, Apr. 2005.

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

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

[10] Z. Tüske, P. Golik, R. Schlüter, and F. R. Drepper, “Non-stationary feature extraction for automatic speech recognition,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 5204–5207, May 2011.

[11] Z. Tüske, F. R. Drepper, and R. Schlüter, “Non-stationary signal processing and its application in speech recognition,” *Workshop on statistical and perceptual audition*, Sep. 2012.

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

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

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

[15] D. Ealey, H. Kelleher, and D. Pearce, “Harmonic tunnelling: tracking non-stationary noises during speech,” in *Proc. Eurospeech*, Sep. 2001, pp. 437–440.

[16] M. Krawczyk and T. Gerkmann, “STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1931–1940, 2014.

[17] M. G. Christensen and A. Jakobsson, “Optimal filter designs for separating and enhancing periodic signals,” *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

[18] K. I. Molla, K. Hirose, N. Minematsu, and K. Hasan, “Voiced/unvoiced detection of speech signals using empirical mode decomposition model,” in *Int. Conf. Information and Communication Technology*, Mar. 2007, pp. 311–314.

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

[20] N. Metropolis, “The beginning of the monte carlo method,” *Los Alamos Science*, no. 15, pp. 125–130, 1987.

[21] Y. Hu and P. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, 2008.

Paper B

Enhancement and Noise Statistics Estimation for Non-Stationary Voiced Speech

Sidsel Marie Nørholm, Jesper Rindom Jensen and Mads Græsbøll Christensen

The paper has been submitted to the
IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2015.

in peer-review

The layout has been revised.

Abstract

In this paper, single channel speech enhancement in the time domain is considered. We address the problem of handling non-stationary speech by describing the voiced speech parts by a harmonic linear chirp model instead of using the traditional harmonic model. This means that the speech signal is not assumed stationary, instead the fundamental frequency can vary linearly within each frame. The linearly constrained minimum variance (LCMV) filter and the amplitude and phase estimation (APES) filter are derived in this framework and compared to the harmonic versions of the same filters. It is shown through simulations on synthetic and speech signals, that the chirp versions of the filters perform better than their harmonic counterparts in terms of output signal-to-noise ratio (SNR) and signal reduction factor. For synthetic signals, the output SNR for the harmonic chirp APES based filter is increased 3 dB compared to the harmonic APES based filter at an input SNR of 10 dB, and at the same time the signal reduction factor is decreased. For speech signals, the increase is 1.5 dB along with a decrease in the signal reduction factor of 0.7. As an implicit part of the APES filter, a noise covariance matrix estimate is obtained. We suggest using this estimate in combination with other filters such as the Wiener filter. The performance of the Wiener filter and LCMV filter are compared using the APES noise covariance matrix estimate and a power spectral density (PSD) based noise covariance matrix estimate. It is shown that the APES covariance matrix works well in combination with the Wiener filter, and the PSD based covariance matrix works well in combination with the LCMV filter.

Index Terms: Speech enhancement, chirp model, harmonic signal model, non-stationary speech.

1 Introduction

Speech enhancement has many applications as in, e.g., mobile phones and hearing aids. Often, the speech enhancement is carried out in a transformed domain, a common one being the frequency domain. Here, the methods based on computational auditory scene analysis (CASA) [1], spectral subtraction [2] and Wiener filtering [3] are well known methods. The CASA methods are based on feature extraction of the speech signal whereas spectral subtraction and Wiener filtering require an estimate of the power spectral density (PSD) of the noise. This can be estimated in different ways [4–6], but common to these methods is that they primarily rely on periods without speech to update the noise statistics. In periods of speech, the PSD is mostly given by the previous estimate of the PSD. This update pattern makes the PSD estimates very vulnerable to non-stationary noise. Furthermore, in order to make enhancement in the frequency domain, the data needs to be transformed by use of the Fourier transform. This transform assumes that the signals are stationary within the analysis window which for speech signals is often between 20 ms and 30 ms. It is, however, well known that this assumption of stationary speech does not hold [7, 8], as, e.g., the fundamental frequency and formants vary continuously over time in periods of voiced speech, making the speech signal non-stationary. In [9, 10], it is suggested replacing the standard Fourier transform with a fan-chirp transform in the analysis of non-stationary harmonic signals. The voiced speech parts of a speech signal are often described by a harmonic model, and since voiced speech is the main constituent of speech, it makes good sense to use this transform on speech signals. The voiced speech can also easily be separated from the unvoiced speech by use of voiced/unvoiced detectors [11, 12]. The assumption behind the fan-chirp transform is that the harmonic frequencies of the signal vary linearly over time, and it is shown that spectra obtained using the fan-chirp transform have much more distinct peaks at the positions of the harmonic frequencies. Alternatively, the enhancement can be done directly in the time domain where, e.g., the Wiener filter has also been defined [13]. Most time domain filters also depend on noise statistics in the form of a covariance matrix. These are often obtained by averaging over a small frame of the observed signal, and, therefore, the signal in these frames is also assumed stationary. Also, a common way to filter speech in the time domain is by describing the voiced speech parts by a harmonic model [14–16]. The signal based on this model is composed of a set of sinusoids where the frequency of each sinusoid is given by an integer multiple of a fundamental frequency. The fundamental frequency in this model is constant within a frame, and so the voiced speech is assumed stationary. In [15], it is proposed making a noise estimate by subtracting an estimate of the desired signal based on the harmonic model, and, from this, make a noise covariance matrix estimate. In doing so, the observed signal only needs to be stationary within the frame of 20 to 30 ms when the noise statistics are estimated and not from one speech free period to the next, as was mostly the case for the PSD. The non-stationarity of speech is considered in [17–19] in relation to modelling and parameter estimation. In these papers, a modified version of the harmonic model is used where a chirp parameter is introduced to allow the frequency of the harmonics to change linearly within each frame. In [17], the first model introduced to describe the speech signal is very flexible, but it is approximated with a Taylor expansion that leads to bigger and bigger deviations from the original model when the harmonic number increases, as mentioned in the paper. In [18, 19], a harmonic chirp model is used to describe the voiced speech, and the parameters of the model are estimated based on maximum likelihood estimation, but using different ways to avoid making a two dimensional search for the fundamental frequency and chirp rate.

We investigate the harmonic chirp model further in relation to speech enhancement. The linearly constrained minimum variance (LCMV) and the amplitude and phase estimation (APES) filters have previously been derived under the harmonic framework [16, 20, 21]. One objective of this work is to increase the performance of these filters by deriving them according to the harmonic chirp model. Both LCMV and APES filter have the goal of minimising the output noise power from the filter under the constraint that the desired signal is passed undistorted, or equivalently, when the constraint is fulfilled, to maximise the output signal-to-noise ratio (SNR). Therefore, we evaluate the performance of the filters by use of the output SNR and the signal reduction factor which measures the distortion of the desired signal introduced by the filters. Another objective is to investigate the noise covariance matrix that is obtained implicitly when the APES based filter is made in relation to other filters as, e.g., the Wiener filter. The noise covariance matrix estimate is made under the assumption of non-stationary speech when the harmonic chirp model is used. It is generated from the covariance matrix of the observed signal by subtracting the part that conforms to the harmonic chirp model. We propose using this estimate in combination with other filters as well and compare the performance of the Wiener filter using the APES noise covariance matrix to the chirp APES based filter. Alternatively, we suggest making a noise covariance matrix estimate based on the earlier mentioned state of the art PSD estimates [5, 6] since more work has been put into noise PSD estimates than estimation of time domain noise statistics. The PSD is related through the Fourier transform to the autocorrelation and, thereby, to the covariance matrix as well.

In Section 2, the harmonic chirp model is introduced. In Section 3, the LCMV and APES based filters for harmonic chirp signals are derived. The Wiener filter and a family of trade-off filters are then introduced. In Section 4, the estimation of covariance matrices are discussed and suggestions on how to do it is given. In Section 5, the performance of the LCMV and APES filters are considered through derivations of the used performance measures. In Section 6, experimental results on synthetic and real speech signals are shown and discussed, and the presented work is concluded in Section 6.

2 Framework

We are here considering the problem of recovering a desired signal, $s(n)$, from an observed signal, $x(n)$, with the desired signal buried in additive noise, i.e.,

$$x(n) = s(n) + v(n), \tag{B.1}$$

for discrete time indices $n = 0, ..., N-1$. The desired signal and noise are assumed to be zero mean signals and mutually uncorrelated. Further, we assume that the desired signal is quasi periodic which is a reasonable assumption for voiced speech. Often, voiced speech is described by a harmonic model [16, 22, 23], but here we are using a harmonic chirp model which makes the model capable of handling non-stationary speech.

The signal is built up by a set of harmonically related sinusoids as in the normal harmonic model where the sinusoid with the lowest frequency is the fundamental and the other sinusoids have frequencies given by an integer multiple of the fundamental. In the harmonic model, the speech signal is assumed stationary in short segments which is rarely the case. Instead the fundamental frequency is varying slowly over time which can be modelled by using a harmonic linear chirp model. In a linear chirp signal the instantaneous frequency of the l'th harmonic, $\omega_l(n)$, is not stationary but varies linearly with time [24],

$$\omega_l(n) = l(\omega_0 + kn), \tag{B.2}$$

where $\omega_0 = f_0/f_s 2\pi$, with f_s the sampling frequency, is the normalised fundamental frequency and k is the fundamental chirp rate. The instantaneous phase, $\theta_l(n)$, of the sinusoids are given by the integral of the instantaneous frequency as

$$\theta_l(n) = l\left(\omega_0 n + \frac{1}{2}kn^2\right) + \phi_l, \tag{B.3}$$

and, thereby, this leads to the harmonic chirp model for a voiced speech signal, $s(n)$:

$$\begin{aligned} s(n) &= \sum_{l=1}^{L} A_l \cos\left(\theta_l(n)\right) && \text{(B.4)} \\ &= \sum_{l=1}^{L} A_l \cos\left(l\left(\omega_0 n + \frac{k}{2}n^2\right) + \phi_l\right). && \text{(B.5)} \end{aligned}$$

where L is the number of harmonics, $A_l > 0$ is the amplitude and ϕ_l is the initial phase of the l'th harmonic, respectively. A special case of the harmonic chirp model for $k = 0$ is then the traditional harmonic model:

$$s(n) = \sum_{l=1}^{L} A_l \cos\left(l\omega_0 n + \phi_l\right) \tag{B.6}$$

In the speech enhancement process later, it is instructive to make the relationship between the time dependent part of the instantaneous phase, $l(\omega_0 n + k/2n^2)$, and the initial phase, ϕ_l multiplicative instead of additive. This either leads to the real signal model [14]:

$$\begin{aligned} s(n) = \sum_{l=1}^{L} & a \cos\left(l\left(\omega_0 n + \frac{k}{2}n^2\right)\right) \\ & + b \sin\left(l\left(\omega_0 n + \frac{k}{2}n^2\right)\right), \end{aligned} \tag{B.7}$$

where $a = A_l \sin(\phi_l)$ and $b = A_l \cos(\phi_l)$, or, by using Eulers formula, to the complex signal model:

$$\begin{aligned} s(n) &= \sum_{l=1}^{L} \alpha_l e^{jl\left(\omega_0 n + k/2n^2\right)} + \alpha_l^* e^{-jl\left(\omega_0 n + k/2n^2\right)} \\ &= \sum_{l=1}^{L} \alpha_l z^l(n) + \alpha_l^* z^{-l}(n), \end{aligned} \tag{B.8}$$

where

$$z(n) = e^{j(\omega_0 n + k/2n^2)} \tag{B.9}$$

and $\alpha_l = \frac{A_l}{2} e^{j\phi}$. Since (B.7) and (B.8) are two ways of describing the same signal, it is possible to design optimal filters based on both, but the complex model in (B.8) gives a more intuitive and simple notation, and, therefore, we will use this model in the following instead of the real model in (B.7) [14].

Defining a subvector of samples

$$\mathbf{s}(n) = [s(n)\ s(n-1)\ \ldots\ s(n-M+1)]^T \tag{B.10}$$

where $M \leq N$ and $(\cdot)^T$ denotes the transpose, the signal model can be written as

$$\mathbf{s}(n) = \mathbf{Z}\mathbf{D}(n)\mathbf{a}, \tag{B.11}$$

where $\mathbf{Z}$ is a matrix with Vandermonde structure constructed from a set of L modified Fourier vectors matching the harmonics of the signal,

$$\mathbf{Z} = [\mathbf{z}(1)\ \ \mathbf{z}(-1)\ \ \mathbf{z}(2)\ \ \mathbf{z}(-2)\ \ \ldots\ \ \mathbf{z}(L)\ \ \mathbf{z}(-L)], \tag{B.12}$$

with

$$\mathbf{z}(l) = \begin{bmatrix} 1 \\ e^{jl(\omega_0 + k/2)} \\ \vdots \\ e^{jl(\omega_0(M-1) + k/2(M-1)^2)} \end{bmatrix} = \begin{bmatrix} z(0)^l \\ z(1)^l \\ \vdots \\ z(M-1)^l \end{bmatrix}. \tag{B.13}$$

The $\mathbf{Z}$ matrix is made with reference to $n = 0$, and, therefore, the diagonal matrix $\mathbf{D}(n)$ is included to take care of the delay from $n = 0$ to the actual start of the subvector, $\mathbf{s}(n)$, i. e.,

$$\mathbf{D}(n) = \begin{bmatrix} z(n)^1 & & & & \\ & z(n)^{-1} & & \mathbf{0} & \\ & & \ddots & & \\ & \mathbf{0} & & z(n)^L & \\ & & & & z(n)^{-L} \end{bmatrix}. \tag{B.14}$$

The vector $\mathbf{a}$ contains the complex amplitudes of the harmonics, $\mathbf{a} = [\alpha_1\ \alpha_1^*\ \alpha_2\ \alpha_2^*\ \ldots\ \alpha_L\ \alpha_L^*]^T$, where $\{\cdot\}^*$ denotes the complex conjugate.

The observed signal vector, $\mathbf{x}(n)$, is then given by

$$\mathbf{x}(n) = \mathbf{s}(n) + \mathbf{v}(n), \tag{B.15}$$

where $\mathbf{x}(n)$ and $\mathbf{v}(n)$ are defined in a similar way to $\mathbf{s}(n)$ in (B.10). Due to the assumption of zero mean uncorrelated signals, the variance of the observed signal is given by the sum of the variances of the desired signal and noise, $\sigma_x^2 = \sigma_s^2 + \sigma_v^2$, where the variance of a signal $g(n)$ is defined by $\sigma_g^2 = \mathbb{E}\{g^2(n)\}$ with $\mathbb{E}\{\cdot\}$ denoting statistical expectation. The level of the desired signal relative to the noise in the observed signal is described by the input signal-to-noise ratio (SNR):

$$\text{iSNR} = \frac{\sigma_s^2}{\sigma_v^2}. \tag{B.16}$$

The objective is then to recover the desired signal in the best possible way from the observed signal. This can be done by filtering $\mathbf{x}(n)$ with a filter $\mathbf{h} = [h(0)\ h(1)\ \ldots\ h(M-1)]^T$, where $M \leq N$ is the filter length and $\{\cdot\}^T$ denotes the transpose. However, because both the observed signal and the filter are real, multiplying the observed signal with the Hermitian transposed, $\{\cdot\}^H$, filter gives the same result as multiplying with the transposed filter. Due to the choice of a complex representation of the real signal, the Hermitian notation is used throughout the paper since this gives more intuitive interpretations of some intermediate variables such as covariance matrices. That is,

$$\hat{s}(n) = \mathbf{h}^H\mathbf{x}(n) = \mathbf{h}^H\mathbf{s}(n) + \mathbf{h}^H\mathbf{v}(n), \tag{B.17}$$

gives an estimate, $\hat{s}(n)$, of the desired signal, $s(n)$. The variance of the estimate is then $\sigma_{\hat{s}}^2 = \sigma_{x,\text{nr}}^2 = \sigma_{s,\text{nr}}^2 + \sigma_{v,\text{nr}}^2$, where $\sigma_{x,\text{nr}}^2$ is the variance of the observed signal after noise reduction, i.e.,

$$\sigma_{x,\text{nr}}^2 = \mathbb{E}\{(\mathbf{h}^H\mathbf{x}(n))^2\} = \mathbf{h}^H\mathbf{R}_x\mathbf{h}, \tag{B.18}$$

with $\mathbf{R}_x$ being the covariance matrix of the observed signal defined as:

$$\mathbf{R}_x = \mathbb{E}\{\mathbf{x}(n)\mathbf{x}^H(n)\}. \tag{B.19}$$

Similar definitions of the variance after noise reduction and the covariance matrix hold for the desired signal and the noise signal. Further, using the signal model in (C.7), the covariance matrix of the desired signal can be expressed as

$$\begin{aligned}
\mathbf{R}_s &= \mathbb{E}\{\mathbf{s}(n)\mathbf{s}^H(n)\} && \text{(B.20)}\\
&= \mathbb{E}\left\{(\mathbf{Z}\mathbf{D}(n)\mathbf{a})\,(\mathbf{Z}\mathbf{D}(n)\mathbf{a})^H\right\} && \text{(B.21)}\\
&= \mathbf{Z}\mathbf{P}\mathbf{Z}^H, && \text{(B.22)}
\end{aligned}$$

where

$$\mathbf{P} = \mathbb{E}\{\mathbf{D}(n)\mathbf{a}\mathbf{a}^H\mathbf{D}(n)^H\} = \mathbb{E}\{\mathbf{a}\mathbf{a}^H\}. \tag{B.23}$$

Here, $\mathbf{P}$ is the covariance matrix of the amplitudes. If the phases are independent and uniformly distributed, it reduces to a diagonal matrix with the powers of the harmonics on the diagonal.

If $s(n)$ and $v(n)$ are uncorrelated, $\mathbf{R}_x$ is given by the sum of the covariance matrix of the desired signal, $\mathbf{R}_s$, and the covariance matrix of the noise, $\mathbf{R}_v$,

$$\mathbf{R}_x = \mathbf{R}_s + \mathbf{R}_v. \tag{B.24}$$

Like the input SNR, the output SNR is the ratio of the desired signal to noise but now after noise reduction

$$\begin{align} \text{oSNR}(\mathbf{h}) &= \frac{\sigma^2_{s,\text{nr}}}{\sigma^2_{v,\text{nr}}} \tag{B.25} \\ &= \frac{\mathbf{h}^H\mathbf{R}_s\mathbf{h}}{\mathbf{h}^H\mathbf{R}_v\mathbf{h}}. \tag{B.26} \end{align}$$

It is desirable to have as high an output SNR as possible, but if the filter distorts the desired signal along with removing the noise, it might be more beneficial to make a compromise between noise reduction and signal distortion. The signal distortion can be described by the signal reduction factor which is the ratio between the variance of the desired signal before and after noise reduction:

$$\begin{align} \xi_{\text{sr}}(\mathbf{h}) &= \frac{\sigma^2_s}{\sigma^2_{s,\text{nr}}} \tag{B.27} \\ &= \frac{\sigma^2_s}{\mathbf{h}^H\mathbf{R}_s\mathbf{h}}. \tag{B.28} \end{align}$$

A distortionless filter will give a signal reduction factor of one, even though a filter can introduce distortion in sub-bands and still have a signal signal reduction factor of one.

3 Filters

3.1 Traditional filters

A set of different filters can be defined by looking at the error, $e(n)$, between the desired signal, $s(n)$, and the estimate of the desired signal, $\hat{s}(n)$,

$$\begin{align} e(n) &= s(n) - \hat{s}(n) = s(n) - \mathbf{h}^H\mathbf{x}(n) \\ &= s(n) - \mathbf{h}^H\mathbf{s}(n) - \mathbf{h}^H\mathbf{v}(n). \tag{B.29} \end{align}$$

From this, the minimum mean squared error (MSE) criterion can be defined

$$J(\mathbf{h}) = \mathbb{E}\{e(n)^2\} = \mathbb{E}\{(s(n) - \mathbf{h}^H\mathbf{x}(n))^2\} \tag{B.30}$$

$$= \mathbb{E}\{\left(s(n) - \mathbf{h}^H\mathbf{s}(n) - \mathbf{h}^H\mathbf{v}(n)\right)^2\} \tag{B.31}$$

Minimisation of $J(\mathbf{h})$ leads to the classical Wiener filter [13]:

$$\mathbf{h}_w = \mathbf{R}_x^{-1}\mathbf{R}_s\mathbf{i}_M, \tag{B.32}$$

where $\mathbf{i}_M$ is the first column of the $M \times M$ identity matrix. Using (B.24), the Wiener filter can be rewritten as

$$\mathbf{h}_w = \mathbf{R}_x^{-1}(\mathbf{R}_x - \mathbf{R}_v)\mathbf{i}_M, \tag{B.33}$$

which is often convenient when the covariance matrices are to be estimated.

More flexible filters can be obtained if the error signal, $e(n)$, is seen as composed of two parts, one expressing the signal distortion, $e_s(n)$, the other the amount of residual noise, $e_v(n)$,

$$e_s(n) = s(n) - \mathbf{h}^H\mathbf{s}(n), \tag{B.34}$$

$$e_v(n) = \mathbf{h}^H\mathbf{v}(n), \tag{B.35}$$

with the corresponding MSEs being

$$J_s(\mathbf{h}) = \mathbb{E}\{e_s(n)^2\} = \mathbb{E}\{(s(n) - \mathbf{h}^H\mathbf{s}(n))^2\} \tag{B.36}$$

$$J_v(\mathbf{h}) = \mathbb{E}\{e_v(n)^2\} = \mathbb{E}\{(\mathbf{h}^H\mathbf{v}(n))^2\}. \tag{B.37}$$

These error measures make it possible to, e.g., minimise the noise power output of the filter while constraining the amount of signal distortion the filter introduces [25], i.e.,

$$\min_{\mathbf{h}} J_v(\mathbf{h}) \quad \text{s.t.} \quad J_s(\mathbf{h}) = \beta\sigma_s^2, \tag{B.38}$$

where β is a tuning parameter. Solving for the filter by use of the Lagrange multiplier λ gives:

$$\mathbf{h}_\lambda = \left(\mathbf{R}_s + \frac{1}{\lambda}\mathbf{R}_v\right)^{-1}\mathbf{R}_s\mathbf{i}_M, \tag{B.39}$$

where $\lambda > 0$ satisfies $J_s(\mathbf{h}) = \beta\sigma_s^2$. When $\lambda \to \infty$, $\mathbf{h} \to \mathbf{i}_M$ which gives $\beta \to 0$ and $\hat{s}(n) = x(n)$. When $\lambda = 1$ the filter reduces to the Wiener filter and $\lambda \to 0 \Rightarrow \beta \to 1$ which means that the difference in variance between the desired signal and the estimated signal is equal to the variance of the desired signal and so a large amount of signal distortion is introduced.

3.2 Parametric filters

The filter in (B.39) has no control over the distortion of the single harmonics in a voiced speech signal. This is, however, possible by minimisation of $J_v(\mathbf{h})$ under the constraint that the desired signal is passed undistorted, i.e.,

$$\min_{\mathbf{h}} J_v(\mathbf{h}) \quad \text{s.t.} \quad s(n) - \mathbf{h}^H \mathbf{s}(n) = 0. \tag{B.40}$$

Expressing the signal using the harmonic chirp model in (C.7), the restriction can be rewritten as

$$\begin{aligned} s(n) - \mathbf{h}^H \mathbf{s}(n) = 0 \Leftrightarrow \quad & \text{(B.41)} \\ \mathbf{i}_M^H \mathbf{Z}\mathbf{D}(n)\mathbf{a} - \mathbf{h}^H \mathbf{Z}\mathbf{D}(n)\mathbf{a} = 0 \Leftrightarrow \quad & \text{(B.42)} \\ \mathbf{1}^H = \mathbf{h}^H \mathbf{Z}, \quad & \text{(B.43)} \end{aligned}$$

where $\mathbf{1} = [1 \ \ldots \ 1]^T$, and using the relation in (B.18), (B.40) can be rewritten as

$$\min_{\mathbf{h}} \mathbf{h}^H \mathbf{R}_v \mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^H \mathbf{Z} = \mathbf{1}^T, \tag{B.44}$$

where the filter should be longer than the number of constraints, i.e., $M > 2L$ to ensure a nontrivial solution. If the signal is passed through the filter undistorted, the variance of the signal before and after filtering is the same, and the output SNR reduces to

$$\text{oSNR}(\mathbf{h}) = \frac{\sigma_s^2}{\mathbf{h}^H \mathbf{R}_v \mathbf{h}}. \tag{B.45}$$

Minimising $\mathbf{h}^H \mathbf{R}_v \mathbf{h}$ under the constraint of an undistorted signal will, therefore, lead to a filter that maximises the output SNR under the same constraint.

The solution to (B.44) is the linearly constrained minimum variance (LCMV) filter and is given by [20]:

$$\mathbf{h}_{\text{LCMV}} = \mathbf{R}_v^{-1} \mathbf{Z} (\mathbf{Z}^H \mathbf{R}_v^{-1} \mathbf{Z})^{-1} \mathbf{1}. \tag{B.46}$$

The filter reduces to the LCMV filter for harmonic signals when $k = 0$. The covariance matrix of the noise signal is not known and has to be estimated. This is not trivial, but in an optimal situation where the signal model fits perfect, the noise covariance matrix can be replaced by the covariance matrix of the observed signal, $\mathbf{R}_x$, [15], which is easier to estimate, i.e.,

$$\mathbf{h}_{\text{LCMV}} = \mathbf{R}_x^{-1} \mathbf{Z} (\mathbf{Z}^H \mathbf{R}_x^{-1} \mathbf{Z})^{-1} \mathbf{1}. \tag{B.47}$$

Another more empirical approach taking its starting point in the MSE is the amplitude and phase estimation (APES) filter [16]. Here, the harmonic chirp model is also assumed and the expectation is approximated by an average over time, leading to the estimated MSE:

$$J_a(\mathbf{h}) = \frac{1}{N-M+1}\sum_{n=0}^{N-M} |s(n) - \mathbf{h}^H\mathbf{x}(n)|^2, \tag{B.48}$$

$$= \frac{1}{N-M+1}\sum_{n=0}^{N-M} |\mathbf{a}^H\mathbf{w}(n) - \mathbf{h}^H\mathbf{x}(n)|^2, \tag{B.49}$$

where

$$\mathbf{w}(n) = \mathbf{D}(n)^H\mathbf{Z}^H\mathbf{i}_M = \mathbf{D}(n)^H\mathbf{1}. \tag{B.50}$$

Writing out the terms in the quadratic expression and solving for the amplitudes [16] gives $\widehat{\mathbf{a}} = \mathbf{W}^{-1}\mathbf{G}\mathbf{h}$, and, thereby,

$$J_a(\mathbf{h}) = \mathbf{h}^H\mathbf{R}_x\mathbf{h} - \mathbf{h}^H\mathbf{G}^H\mathbf{W}^{-1}\mathbf{G}\mathbf{h} \tag{B.51}$$

$$= \mathbf{h}^H(\mathbf{R}_x - \mathbf{G}^H\mathbf{W}^{-1}\mathbf{G})\mathbf{h}, \tag{B.52}$$

$$= \mathbf{h}^H\mathbf{Q}\mathbf{h} \tag{B.53}$$

with

$$\mathbf{G} = \frac{1}{N-M+1}\sum_{n=0}^{N-M} \mathbf{w}(n)\mathbf{x}^H(n), \tag{B.54}$$

$$\mathbf{W} = \frac{1}{N-M+1}\sum_{n=0}^{N-M} \mathbf{w}(n)\mathbf{w}^H(n). \tag{B.55}$$

and

$$\mathbf{Q} = \mathbf{R}_x - \mathbf{G}^H\mathbf{W}^{-1}\mathbf{G}. \tag{B.56}$$

As with the LCMV filter, the MSE is minimised with a constraint that the desired signal should be passed undistorted, leading to a similar filter [16]:

$$\mathbf{h}_{\text{APES}} = \mathbf{Q}^{-1}\mathbf{Z}(\mathbf{Z}^H\mathbf{Q}^{-1}\mathbf{Z})^{-1}\mathbf{1}. \tag{B.57}$$

4 Covariance matrix estimates

The covariance matrices used in the derived filters are not known but have to be estimated. The covariance matrix of the observed signal can, e.g., be estimated by use of the sample covariance matrix estimate [20]:

$$\widehat{\mathbf{R}}_x = \frac{1}{N-M+1}\sum_{n=0}^{N-M} \mathbf{x}(n)\mathbf{x}^H(n). \tag{B.58}$$

In order to make the estimate nonsingular, it is required that $2M+1 \leq N$. For this to give a good estimate, the signal should be nearly stationary not only in the set of the filtered M samples, but for all N samples. Otherwise, the N samples are not a good representation of the signal within the M samples, and the sample covariance matrix will not be a good estimate of the observed signal covariance matrix. In such a case, the filters in (B.46) and (B.47) are not identical, and it is, therefore, necessary to find an estimate of the noise covariance matrix.

Exchanging $\mathbf{x}(n)$ in (B.54) with $\mathbf{ZD}(n)\mathbf{a} + \mathbf{v}(n)$, it can be shown that the term $\mathbf{G}^H\mathbf{W}^{-1}\mathbf{G}$ in (B.56) reduces to $\mathbf{ZPZ}^H$ for large sample sizes. This means that $\mathbf{G}^H\mathbf{W}^{-1}\mathbf{G}$ can be seen as an estimate of the covariance matrix of the desired signal, and, therefore, $\mathbf{Q}$ is an estimate of the noise covariance matrix. The APES filter is, therefore, an estimate of the optimal LCMV filter. These covariance matrix estimates are an implicit feature of the APES minimisation.

The APES based noise covariance matrix estimate is obtained using a signal driven approach. Alternatively, we suggest taking a noise driven approach and estimate the noise covariance matrix based on noise PSDs. This can be advantageous since several methods exist for estimating the noise power spectral density in the frequency domain, e.g., based on minimum statistics [5] or minimum mean square error (MMSE) [6]. The power spectral density of a signal $g(n)$, $S_g(\omega)$, is related to the autocorrelation, $R_g(\tau)$, and, thereby, also to the covariance matrix of a signal through the Fourier transform [26]

$$R_g(\tau) = \int_{-\infty}^{\infty} S_g(\omega)e^{j\omega\tau}d\omega, \tag{B.59}$$

where τ denotes a time lag. The autocorrelation is also defined as

$$R_g(\tau) = \mathbb{E}\{g(n)g(n-\tau)\}. \tag{B.60}$$

In order to get a good approximation to the expectation by taking the mean over the samples and to make the covariance matrix full rank, the same restriction on M relative to N applies here, $2M+1 \leq N$.

The noise covariance matrix is then estimated as:

$$\mathbf{R}_v(p,q) = \begin{cases} R_v(q-p) & \text{for } q \geq p \\ R_v(N+q-p) & \text{for } q < p \end{cases} \tag{B.61}$$

for p and $q \in [1{,}M]$.

5 Performance of parametric filters

The theoretical performance of the LCMV filter in (B.46) can be found by inserting the expression for the filter in (B.26) and (B.28). Moreover, the expression for the covariance matrix of the desired signal introduced in (B.22) is used. The output power of the desired signal and noise can be expressed as:

$$\begin{aligned}&\mathbf{h}^H\mathbf{R}_s\mathbf{h} = \\ &\mathbf{1}^T(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z}\mathbf{P}\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z}(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})^{-1}\mathbf{1} \\ &\quad = \mathbf{1}^T\mathbf{P}\mathbf{1} = \sigma_s^2\end{aligned} \tag{B.62}$$

and

$$\begin{aligned}&\mathbf{h}^H\mathbf{R}_v\mathbf{h} = \\ &\mathbf{1}^T(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{R}_v\mathbf{R}_v^{-1}\mathbf{Z}(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})^{-1}\mathbf{1} \\ &\quad = \mathbf{1}^T(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})^{-1}\mathbf{1}.\end{aligned} \tag{B.63}$$

The output SNR and signal reduction factor then becomes:

$$\text{oSNR}(\mathbf{h}) = \frac{\sigma_s^2}{\mathbf{1}^T(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})^{-1}\mathbf{1}}, \tag{B.64}$$

and

$$\xi_{\text{sr}}(\mathbf{h}) = 1. \tag{B.65}$$

These expressions for output SNR and signal reduction are made under the assumption that the noise statistics and the parameters of the signal are known, and that the model fits the desired signal perfectly. Looking at the expression for the output power of the desired signal from the filter in (B.62), it is seen that a distortionless response is dependent on the model of the signal. In order to let the signal pass undistorted through the filter, the model has to fit the signal, and a good estimation of the parameters is needed. The amount of distortion is independent of the noise covariance matrix. The output power of the noise from the filter is, on the other hand, not dependent on the parameters of the model, it is only dependent on a good noise covariance matrix estimate. Using the harmonic chirp model instead of the traditional harmonic model, should for all parametric filters decrease the amount of signal reduction since the model fits the signal better. For the APES filter, a better signal model will also lead to a better noise covariance matrix estimate, and, thereby, influencing both the power output of the desired signal and the noise.

6 Experiments

The simulations are separated in three parts. In the first part, the filters based on the harmonic chirp model are tested on synthetic signals. This is done to verify that the derived filters work in an expected manner and to compare their performance to filters based on the traditional harmonic model under controlled conditions. After that, we turn to simulations on real speech signals to confirm that the harmonic chirp model describes voiced speech better than the traditional harmonic model, and that the harmonic chirp filters perform better than their harmonic counterparts. In the end, the LCMV and APES filters are compared to the Wiener filter where covariance matrix estimates based on the APES principle and PSD are used in both filters.

6.1 Synthetic signal

Setup

The LCMV and APES filters based on the harmonic chirp model were tested on a synthetic chirp signal made according to (C.4) with the same length as the segment length, N. The signal was generated with $L = 10$, $A_l = 1 \, \forall \, l$, random phase, fundamental frequency, and fundamental chirp rate, in the intervals $\phi_l \in [0, 2\pi]$, $f_0 \in [150, 250]$ Hz, $k \in [0, 200]$ Hz^2. The signal is sampled at 8 kHz and added to white Gaussian noise with a variance calculated to fit the desired input SNR.

The filters are evaluated as a function of the input SNR, the segment length, N, and the filter length, M. When the parameters are not varied they are set to: iSNR $= 10$ dB, $N = 230$ and $M = 50$. In the simulations varying M, one covariance matrix is made according to the longest filter length, and the covariance matrix for the shorter filters are taken as submatrices of this. This is done to make the conditions as similar as possible for all filter lengths, with the same segment length N and the same number of elements in the sum in (B.58). The fundamental frequency and fundamental chirp rate are assumed known when designing the filters for the synthetic signals. The results are averaged over 1000 Monte Carlo simulations (MCS). The filters are compared by means of the output SNR in (B.26) and the signal reduction factor in (B.28).

Compared filters

The performance of the chirp based filters is compared to equivalent filters based on the harmonic model. A set of six filters are compared in the simulations:

- $\mathbf{LCMV_{opt}}$: chirp LCMV filter made according to (B.46) with $\mathbf{R}_v$ estimated from the clean noise signal. This filter will have the best possible performance a harmonic chirp LCMV filter can have, but can not be made in practice since there is no access to the clean noise signal.
- $\mathbf{LCMV_h}$: harmonic LCMV filter made according to (B.47) with $k = 0$.
- $\mathbf{LCMV_c}$: chirp LCMV filter made according to (B.47).

- $\mathbf{APES_h}$: harmonic APES filter made according to (B.57) with $k = 0$.
- $\mathbf{APES_c}$: chirp APES filter made according to (B.57).
- $\mathbf{APES_{hc}}$: APES filter made as a combination of the chirp and normal harmonic model with $\mathbf{Z}$ based on the chirp model whereas the estimation of $\mathbf{Q}$ is based on the normal harmonic model. This filter is included to separate the contribution from the modified $\mathbf{Z}$ vector and the modified $\mathbf{Q}$ matrix.

Evaluation

The output SNR and signal reduction factor as a function of the input SNR are shown in Fig. B.1. At an input SNR of -10 dB all filters perform equally well, but as the input SNR is increased the difference in performance between the filters is increased. As expected, the LCMV_opt sets an upper bound for the performance with a similar gain in SNR at all considered levels of input SNR and no distortion of the desired signal. The harmonic chirp APES based filter, APES_c, has similar performance to the optimal LCMV filter. The difference between the two filters, APES_h and APES_hc, is only minor. They deviate from the LCMV_opt around 0 dB input SNR and at an input SNR of 10 dB the gain in SNR is around 3 dB less than for the optimal LCMV filter. They also introduce some distorion of the desired signal, with APES_h distorting the desired signal slightly more than APES_hc. These two filters have the same noise covariance matrix estimate but different versions of the $\mathbf{Z}$ matrix, as is also the case for the two LCMV filters, LCMV_h and LCMV_c, based on the covariance matrix of the observed signal. LCMV_h and LCMV_c have the worst performance of the compared filters, but show the same tendencies as APES_h and APES_hc. The difference between the two filters is mainly a smaller signal distortion for the chirp based filter, but here also with a slight difference in the output SNRs of the two filters. This shows, at least for relatively short filter lengths of $M = 50$, that the major change in performance comes from changing the covariance matrix, from the covariance matrix of the observed signal to the harmonic APES covariance matrix and further again to the harmonic chirp APES covariance matrix. Changing $\mathbf{Z}$ has a minor role but still has an influence, primarily with respect to the distortion of the desired signal.

The same relationships between the filters can be seen in Fig. B.2 where the segment length, N, is varied. The LCMV_opt has the best performance, LCMV_c almost as good, LCMV_h and LCMV_c have the worst performances and APES_h and APES_hc have performances in between. The filters being most influenced by the change in segment length are APES_h and APES_hc. They have a drop in output SNR of around 6 dB when the segment length is increased from 150 to 400 whereas the LCMV filters and the chirp APES based filter only give rise to a decrease in output SNR of 1 to 2 dB. Looking at the signal reduction factor, again the chirp APES based filter and the optimal LCMV filter have more or less no distortion of the desired signal whereas the other filters distort the signal more and more when N is increased.

The filter length, M, is varied in Fig. B.3. Also here, the difference between the filters increases with increasing filter length. Again, the optimal LCMV filter and the harmonic chirp APES based filter perform best whereas the other filters have a lower output SNR and more signal distortion. However, here the output SNR for APES_c starts to deviate from LCMV_opt for filter lengths above approximately 60.

As an example of the filtering, a signal with a length of 500 samples is generated. The fundamental frequency is set to $f_0 = 200$ Hz, the chirp rate to $k = 200\,\text{Hz}^2$, the initial phases are again random and the sampling rate is $f_s = 8$ kHz. The covariance matrices are based on $N = 230$ samples and the filter length is $M = 50$. The fundamental frequency and chirp rate are also here assumed known. The signal is added to white Gaussian noise to give an input SNR of 10 dB. The used filters are the APES_h giving the estimated signal $\widehat{s}_\text{h}$ and APES_c giving the signal $\widehat{s}_\text{c}$ since these two filters showed the best performance in the previous experiments. The estimates are compared to the clean signal and the noisy signal in Fig. B.4. It is seen in the figure that the chirp filter gives a better estimate of the clean signal than the traditional harmonic filter, and the estimate is also closer to the clean signal than the noisy one is.

6.2 Speech signals

Setup

The speech signals used are the 30 sentences included in the NOIZEUS database [27]. Three male and three female speakers produced the 30 Harvard sentences contained in the database. The signals are sampled at 8 kHz and corrupted with noise from the AURORA database [28]. In the first part of this evaluation of speech signals, where the chirp model is compared to the harmonic model, the parameters of the speech signals are estimated from the clean signals. This is done to be able to compare the results for speech signals with the simulations on synthetic data where the parameters were assumed known. In the second part, where the LCMV and Wiener filters are compared, results based on parameters estimated from the noisy signals are shown. The model order and a preliminary fundamental frequency are estimated for every 50 samples using a nonlinear least squares (NLS) estimator [20] with the lower and upper limit for the fundamental frequency given by 80 Hz and 400 Hz, respectively. This is followed by a smoothing [29] and joint estimation of the fundamental frequency and chirp parameter for each sample using the iterative NLS estimator described in [18]. The filter length is increased to $M = 70$. This is done because the real speech signals in many frames have more harmonics than the 10 used to create the synthetic signals, and, therefore, a filter with more degrees of freedom is preferred. A good compromise between filter length and segment length for the LCMV and APES filters would according to [30] be $N = 4M$, but this would lead to quite long segments with the given filter length and, as a compromise, the segment length is again set to $N = 230$. The voiced periods are picked out using a generalised likelihood ratio test [31, 32]. Alternatively, the MAP criteria [20] or other voiced/unvoiced detectors can be used [11, 12]. In some cases where unvoiced speech is mistakenly assigned as voiced, the filters become numerically unstable, and these samples are, therefore, excluded from the evaluation. If the filter is not unstable, the unvoiced speech assigned as voiced is processed as if it was voiced speech. This is expected to give a slight decrease in the performance since it is not possible to obtain noise reduction without signal distortion when using the harmonic model in periods of unvoiced speech. In the first part, where the LCMV filters are compared, white Gaussian noise is used and the output SNR and signal reduction factor are calculated using (B.26) and (B.28) to facilitate the comparison with the results for the synthetic signal. When the LCMV and APES filters are compared to the Wiener filter, babble noise is used, where the noisy signals are taken from the NOIZEUS speech corpus. The noise levels in the NOIZEUS speech corpus range from 0 dB to 15 dB. The babble noise is chosen because it is one of the most difficult noise types to remove. Results are shown both when the parameters are estimated from the clean signal and when the parameters are estimated from the noisy signals. The filters are compared in terms of the output SNR in (B.25) and the signal reduction factor in (B.27) after the voiced speech parts have been concatenated.

Compared filters

In the first part of the simulations with real speech, the same filters used for the synthetic signals are compared. In the second part, the LCMV and APES based filters are compared to the Wiener filter. This is done for two different choices of covariance matrices, the first one using the APES derivation, the other using (B.61) based on the MMSE criterion [6] for finding the PSD. Filters based on the PSD using MMSE and minimum statistics perform almost equally well, and, therefore, only one type of these filters is shown. Further, flexible Wiener filters with two different values of λ are included in the comparisons, leading to six filters:

- **APES$_\mathbf{c}$**: chirp APES filter made according to (B.57).
- **LCMV [6]**: chirp LCMV filter made according to (B.46) with $\mathbf{R}_v$ estimated from (B.61) using MMSE.

- $\mathbf{W_c}$: Wiener filter made according to (B.32) with $\mathbf{R}_s$ estimated using the APES principle as $\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G}$.
- **W [6]**: Wiener filter made according to (B.33) with $\mathbf{R}_v$ estimated from (B.61) using MMSE.
- $\mathbf{W}_{\lambda=0.2}$: Trade-off Wiener filter made according to (B.39) with $\lambda = 0.2$ and $\mathbf{R}_s$ estimated using the APES principle as $\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G}$.
- $\mathbf{W}_{\lambda=5}$: Trade-off Wiener filter made according to (B.39) with $\lambda = 5$ and $\mathbf{R}_s$ estimated using the APES principle as $\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G}$.

Evaluation

In Fig. B.5, the output SNR and signal reduction factor are shown as a function of the input SNR. The output SNR and signal reduction factor are calculated using (B.26) and (B.28) as was also the case for the synthetic signals. It is seen that the tendencies are the same as for the synthetic signal. APES_c does not follow the optimal LCMV filter as closely as it did for the synthetic signal, but this is not surprising since the synthetic signals were made according to the harmonic chirp model, and the parameters were assumed known. For the speech signals, the parameters are estimated, and the model does not fit perfectly since the fundamental frequency will not be completely linear in any considered piece within a speech signal. Even though the performance of the APES_c filter deviates more from the optimal LCMV filter than it did considering synthetic signals, it still has a better performance than the other considered filters. This means that the harmonic chirp model is better at describing the voiced parts of a speech signal and increased performance can be obtained by replacing the traditional harmonic filters with chirp filters.

As an example, the speech signal 'Why were you away a year, Roy?' uttered by a female speaker is filtered. The signal has the advantage that it only contains voiced speech, and the entire signal can, therefore, be filtered by the proposed methods. The signal is sampled at 8 kHz, the segment length is 230, the filter length is 70, and the parameters are estimated in the same way as the previous speech signals. The noise is white Gaussian and the input SNR is 10 dB. The spectrograms of the filtered speech signal using APES_h and APES_c are shown in Fig. B.6 together with the output SNR over time. It is seen that the output SNR of the chirp filter is larger or equal to the output SNR of the harmonic filter. The difference is most pronounced in the first 0.25 seconds and between 1 and 1.25 seconds where the fundamental frequency is changing the most. Here, it is also seen in the spectrograms that the harmonics look slightly cleaner when the chirp filter is used. The Perceptual Evaluation of Speech Quality (PESQ) score [33] for the speech filtered with the harmonic filter is 2.21 whereas the chirp filter gives a PESQ score of 2.32 and the noisy signal gives a PESQ score of 1.57. The speech signals related to this comparison and the comparison in Fig. B.10 can be found at http://www.create.aau.dk/smn.

The increased performance of the harmonic chirp filters relative to the harmonic filters should of course be viewed in light of an increased computational complexity since the joint estimation of the fundamental frequency and chirp rate is based on a search in a two-dimensional space. However, [18] describes how to find the parameters iteratively which will decrease the complexity relative to a two-dimensional grid search, and the initial fundamental frequency estimate used in the algorithm is only estimated for every 50 samples in this work which seems to be sufficient for giving good estimates.

Now we turn to alternative combinations of filters and covariance matrices. Here, the output SNR and signal reduction factor are calculated according to (B.25) and (B.27). This ensures that no filter is favoured in the way the performance is calculated since the covariance matrices based on the sample covariance principle and the PSD are made in two fundamentally different ways. In Fig. B.7a it is seen that five of the six filters work very similar. The Wiener filter in combination with the PSD noise covariance matrix perform significantly worse than the rest when it comes to output SNR. However, the PSD covariance matrix works quite well in combination with the LCMV filter. This filter is one of the better filters at higher input SNRs with respect to output SNR, and it has a low level of distortion at all input SNRs as is seen in Fig. B.7b. This can probably be explained by looking at the filters in (B.32) and (B.46). The Wiener filter is dependent on two covariance matrices, and the relative levels of these two matrices are, therefore, important for the look of the filter. The LCMV based filters are only dependent on one covariance matrix, and in some way the denominator of the LCMV can be seen as a normalisation which makes the filter independent of the absolute size of the covariance matrix used. The trade-off Wiener filter with $\lambda = 0.2$ gives a higher output SNR than the Wiener filter but at the same time it also gives rise to a higher signal distortion. The flexible Wiener filter with $\lambda = 5.0$ works in the opposite way. It gives a lower output SNR, but also a lower degree of signal distortion. In Fig. B.8, the parameters are estimated from the noisy signals whereas the voiced/unvoiced detection is based on the clean signal. The output SNR for the signal dependent filters is decreased a few dBs at low input SNRs whereas it is very similar at high input SNRs. This makes sense since the estimation of parameters is more difficult at low SNRs than at high SNRs. The Wiener filter dependent on the PSD has the same performance in the two situations. In Fig. B.9, also the voiced/unvoiced detection is made based on the noisy signal. The overall performance of all filters is slightly decreased compared to making the detection based on the clean signal, but the tendency between the filters is very similar. This suggests that more unvoiced periods are assigned as voiced speech where the voiced signal model will not apply, and thus the performance will decrease slightly.

As an example, the speech signal 'Why were you away a year, Roy?' is again filtered, now in the presence of babble noise at an input SNR of 10 dB. The filters used for this comparison are the APES_c, LCMV [6], W_c and W [6] and the spectrograms of the resulting signals are shown in Fig. B.10 along with spectrograms of the clean and the noisy signal. From this figure, it seems like the Wiener filter in combination with the APES covariance matrix removes the most noise between the harmonics whereas the APES filter and the LCMV filter remove the noise slightly less, both between the harmonics and outside the range of the speech signal. The Wiener filter in combination with the PSD noise covariance matrix seems to perform no noise reduction and the harmonics are even more difficult to distinguish than in the noisy signal. These observations are in line with the curves of output SNR when looking at an input SNR of 10 dB where the W [6] performs worse than the noisy signal, the APES_c and LCMV [6] perform almost equally well and the W_c performs the best. The PESQ scores for the four filtered signals are, APES_c: 2.09, LCMV [6]: 2.25, W_c: 2.18 and W [6]: 1.54. It is interesting to see that the LCMV [6] gives rise to the highest PESQ score since this was not clear from the spectrograms, but this filter gives a lower signal reduction factor than the APES_c and W_c filters, and, therefore, it makes good sense. The noisy signal has a PESQ score of 2.06. Comparing to the signals in white Gaussian noise in Fig. B.6, the PESQ score of the filtered signals decreased whereas the PESQ score of the noisy signal increased. This difference is mainly due to the different noise types while the fact that the parameters in Fig. B.6 were estimated from the clean signal only contributes slightly. Since babble noise is noise made up from several speakers speaking at the same time, it is distributed in the same frequency range as the speech signal. This makes it more difficult to estimate the relevant parameters and also more difficult to filter out the noise afterwards. However, prewhitening of the noisy signal can help mediate this problem [34] with the noise statistics found using one of the methods in [35].

7 Conclusion

In this paper, the non-stationarity of voiced speech is taken into account in speech enhancement. This is done by describing the speech by a harmonic chirp model instead of the traditional harmonic model. The chirp used is a linear chirp which allows the fundamental frequency to vary linearly within each segment, and, therefore, the speech signal is not assumed stationary within a segment. Versions of the linearly constraint minimum variance (LCMV) filter and amplitude and phase estimation (APES) filter are derived in the framework of harmonic chirp signals. As an implicit part of the APES filter, a noise covariance matrix estimate is derived. This makes the APES filter an estimate of the optimal LCMV filter which maximises the output SNR under the constraint that the desired signal is passed undistorted. APES gives a noise covariance matrix estimate which only assumes the noise signal to be stationary in frames of 20-30 ms as opposed to methods based on power spectral densities (PSDs) which primarily update the noise statistics in periods of unvoiced speech. It is shown through simulations on synthetic and speech signals that the chirp filters give rise to a higher output SNR and a lower signal distortion than their harmonic counterparts, and, therefore, the chirp model describes voiced speech better than the traditional harmonic model. We suggest also using the APES noise covariance matrix estimate in other filters as, e.g., the Wiener filter, and we compare it to a noise covariance matrix estimate based on the PSD. The APES noise covariance matrix estimate is shown to work well in combination with the Wiener and trade-off Wiener filters, whereas the PSD based noise covariance matrix estimate works well in combination with the LCMV filter. All chirp based Wiener and LCMV filters outperform the Wiener filter in combination with the PSD noise covariance matrix estimate.

References

[1] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, algorithms, and applications.* Wiley-IEEE Press, 2006.

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

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

[4] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Process. Lett.*, vol. 6, no. 1, Jan. 1999.

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

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

[7] F. R. Drepper, “A two-level drive-response model of non-stationary speech signals,” *Nonlinear Analyses and Algorithms for Speech Processing*, vol. 1, pp. 125–138, Apr. 2005.

[8] L. Deng and D. O'Shaughnessy, *Speech processing: a dynamic and optimization-oriented approach.* CRC Press, 2003.

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

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

[11] K. I. Molla, K. Hirose, N. Minematsu, and K. Hasan, "Voiced/unvoiced detection of speech signals using empirical mode decomposition model," in *Int. Conf. Information and Communication Technology*, Mar. 2007, pp. 311–314.

[12] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 2, pp. 250–255, 1993.

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

[14] A. Jakobsson, T. Ekman, and P. Stoica, "Capon and APES spectrum estimation for real-valued signals," *Eighth IEEE Digital Signal Processing Workshop*, 1998.

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

[16] M. G. Christensen and A. Jakobsson, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

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

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

[19] Y. Doweck, A. Amar, and I. Cohen, "Joint model order selection and parameter estimation of chirps with harmonic components," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1765–1778, Apr. 2015.

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

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

[22] D. Ealey, H. Kelleher, and D. Pearce, "Harmonic tunnelling: tracking non-stationary noises during speech," in *Proc. Eurospeech*, Sep. 2001, pp. 437–440.

[23] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1931–1940, 2014.

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

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

[26] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications.* Prentice Hall, Inc., 1996.

[27] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7–8, pp. 588 – 601, 2007.

[28] D. Pearce and H. G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. Int. Conf. Spoken Language Process.*, Oct 2000.

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

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

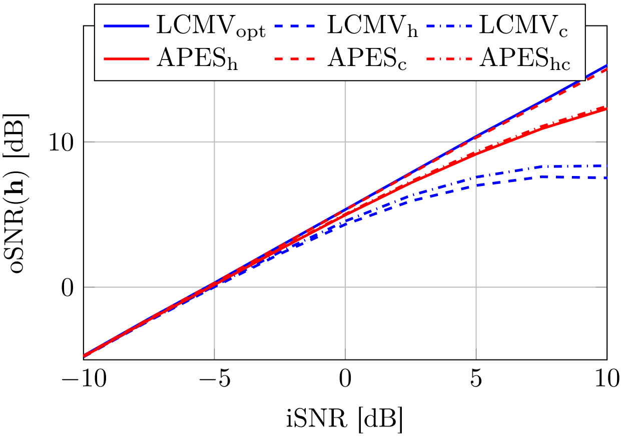
[31] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory.* Prentice Hall, Inc., 1998.

[32] E. Fisher, J. Tabrikian, and S. Dubnov, "Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 2, pp. 502–510, 2006.

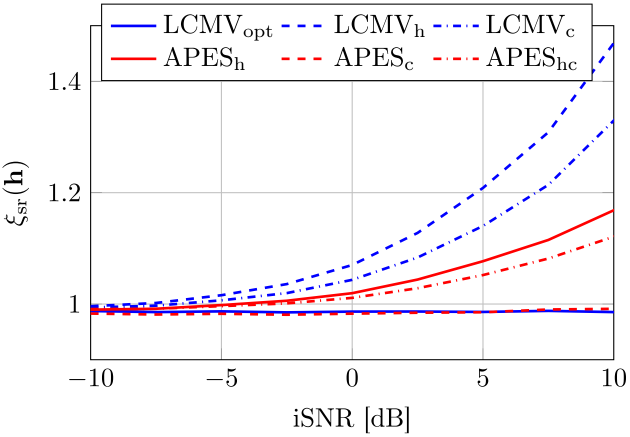
[33] Y. Hu and P. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[34] P. C. Hansen and S. H. Jensen, “Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis,” *EURASIP J. on Advances in Signal Processing*, vol. 2007, no. 1, p. 24, Jun. 2007.

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

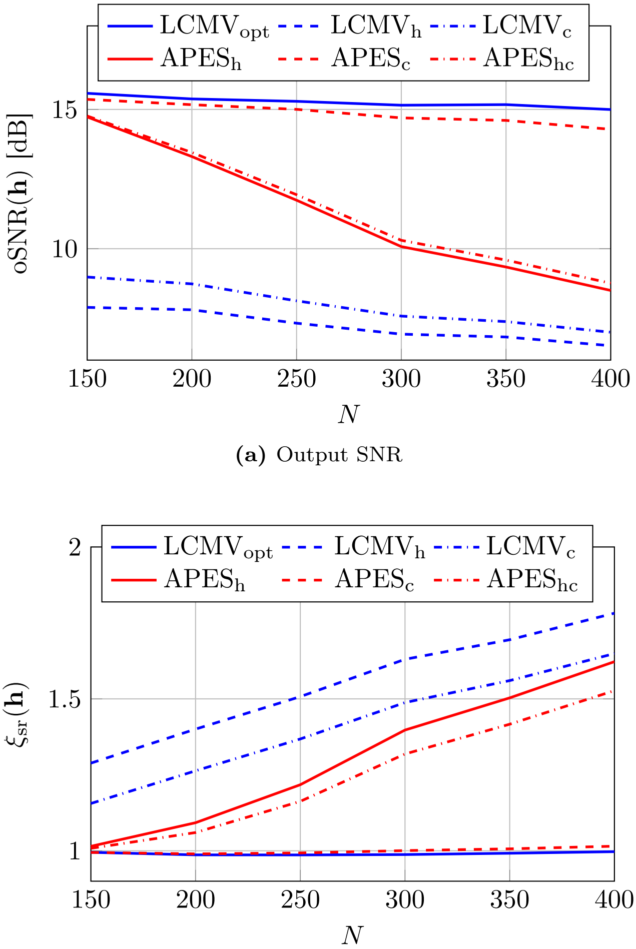


(a) Output SNR



(b) Signal reduction factor

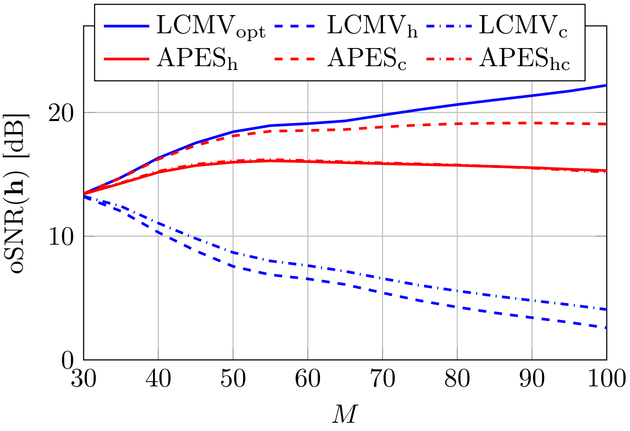
Fig. B.1: Performance as a function of the input SNR for synthetic chirp signals.



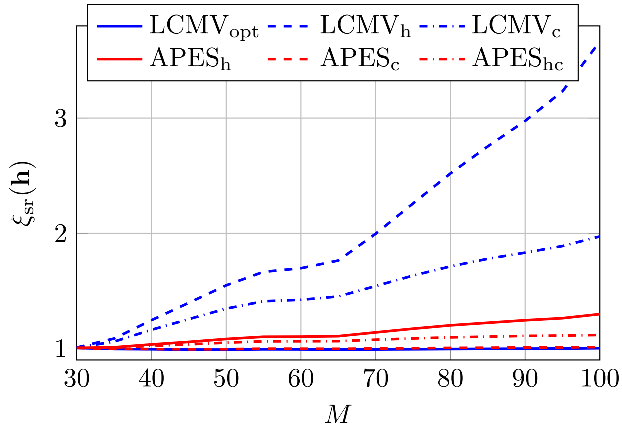
(a) Output SNR

(b) Signal reduction factor

Fig. B.2: Performance as a function of the number of samples N for synthetic chirp signals.



(a) Output SNR



(b) Signal reduction factor

Fig. B.3: Performance as a function of the filter length M for synthetic chirp signals.

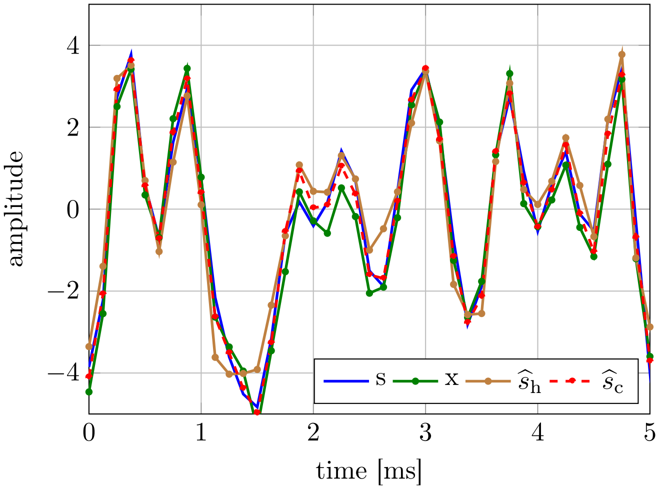
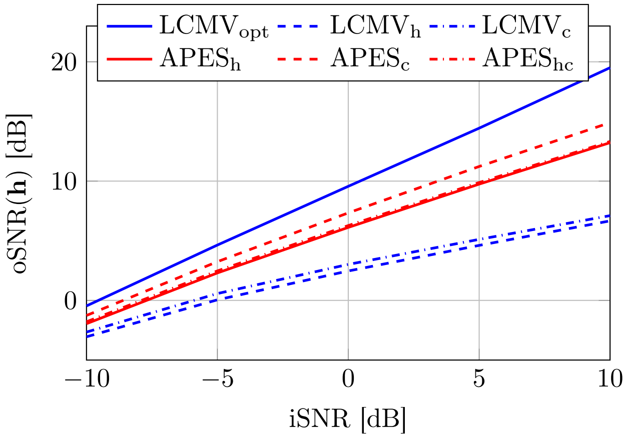
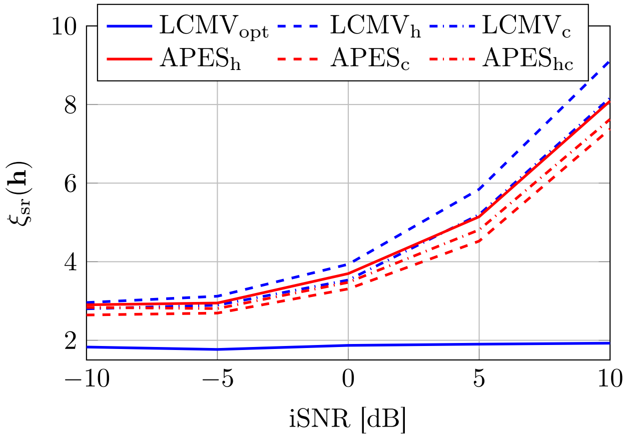


Fig. B.4: Reconstructed signal using $\mathrm{APES_h}$ and $\mathrm{APES_c}$ filters compared to the clean and noisy signals. The noise is white Gaussian and the input SNR is 10 dB.

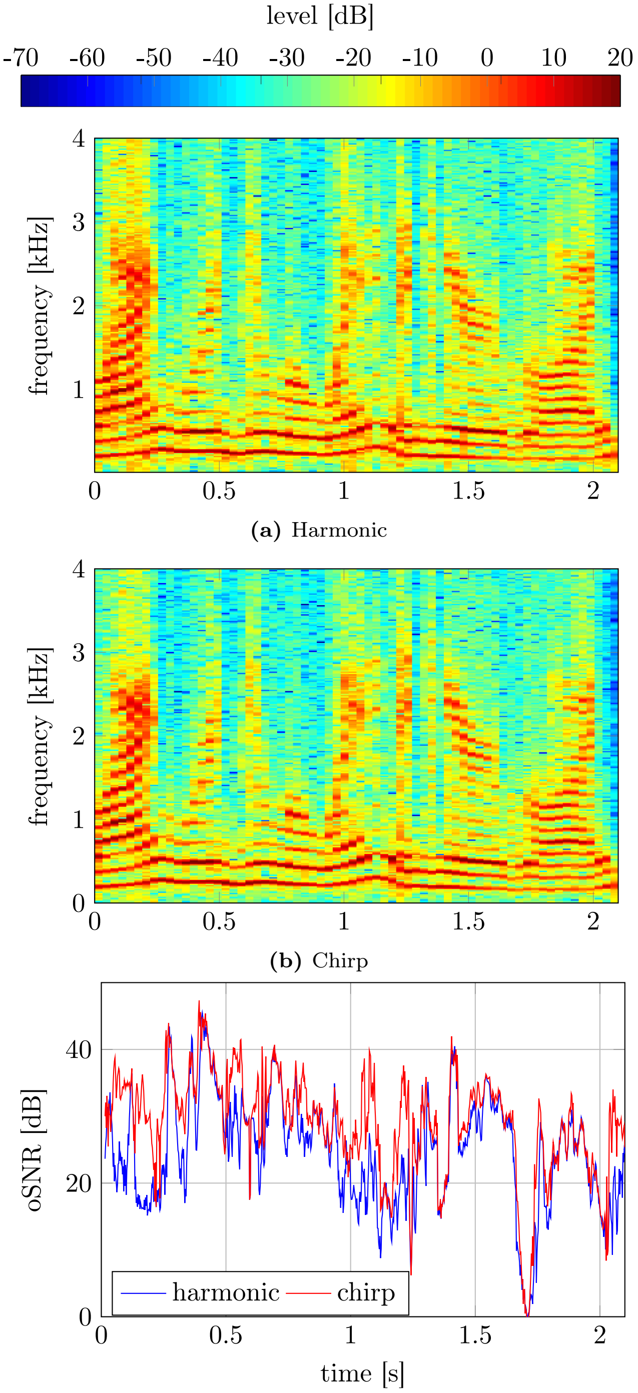


(a) Output SNR



(b) Signal reduction factor

Fig. B.5: Performance as a function of the input SNR, average over NOIZEUS speech corpus added white noise. Parameters estimated from clean speech signals.

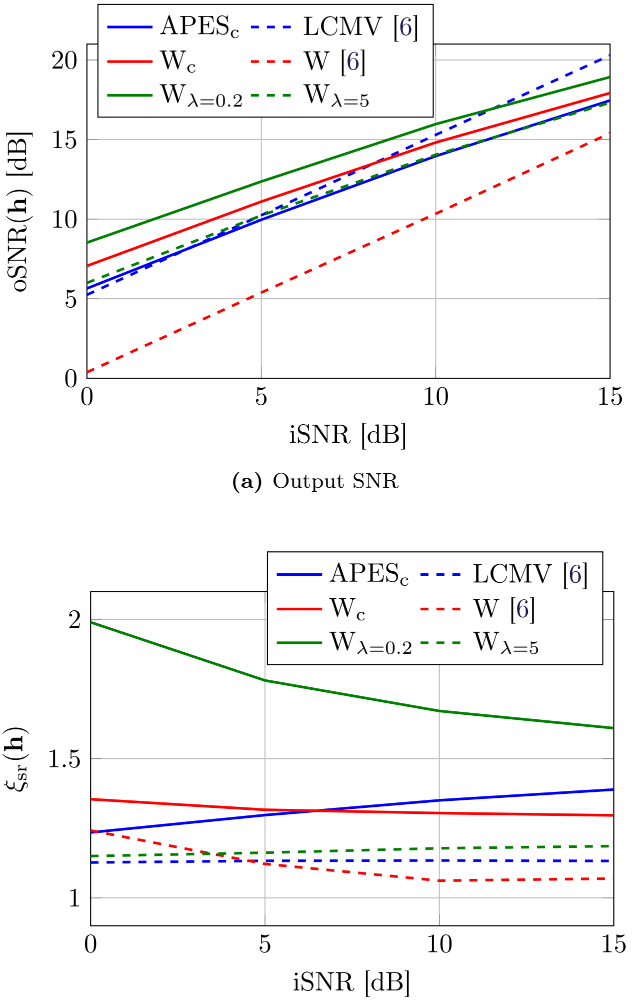


(a) Harmonic

(b) Chirp

(c) Output SNR over time

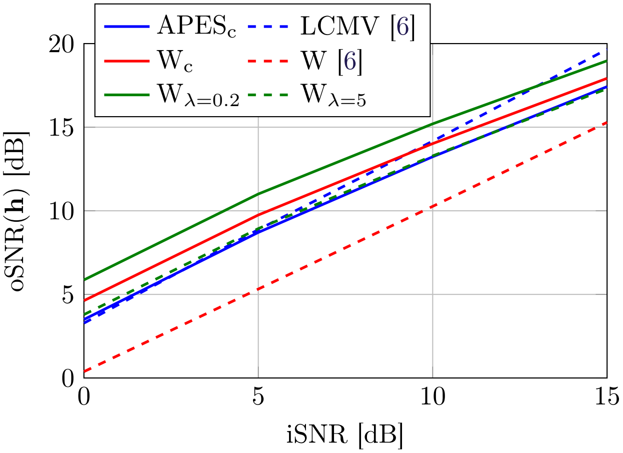
Fig. B.6: Spectrograms of speech signal after filtering with (a) traditional harmonic filter and (b) harmonic chirp filter. In (c) the output SNR over time is shown. The input SNR is 10 dB and the noise is white Gaussian. The clean signal can be seen in Fig. B.10.



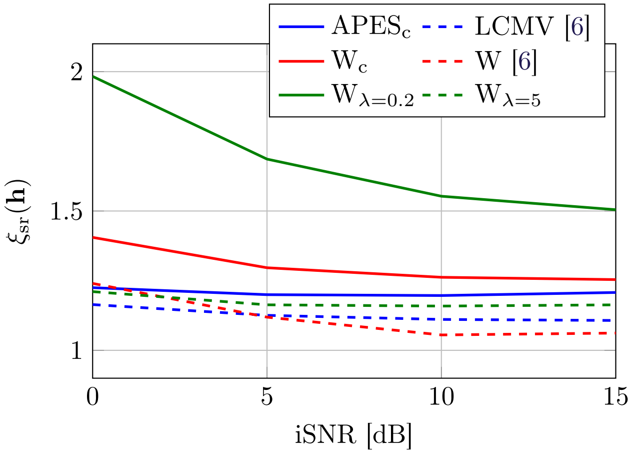
(a) Output SNR

(b) Signal reduction factor

Fig. B.7: Performance as a function of the input SNR, averaged over NOIZEUS corpus with babble noise. Parameters estimated from clean speech signals. Voiced/unvoiced detection from clean signal.

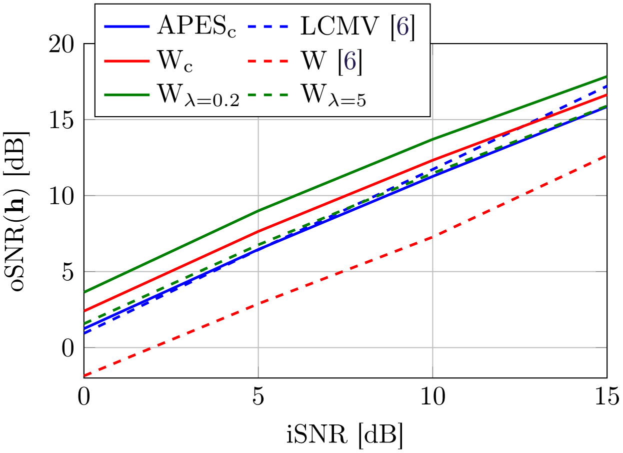


(a) Output SNR

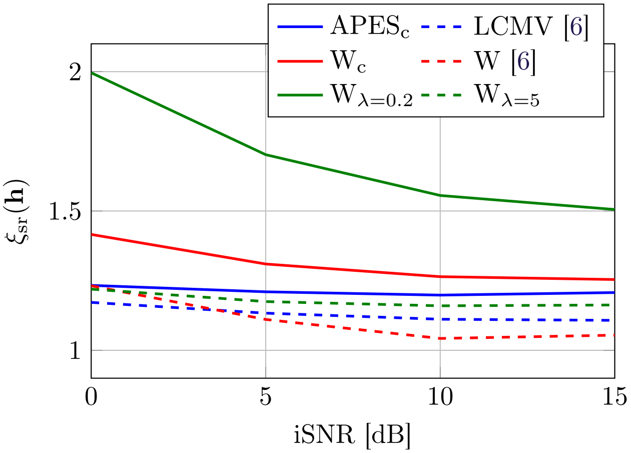


(b) Signal reduction factor

Fig. B.8: Performance as a function of the input SNR, averaged over NOIZEUS corpus with babble noise. Parameters estimated from noisy speech signals. Voiced/unvoiced detection from clean signal.

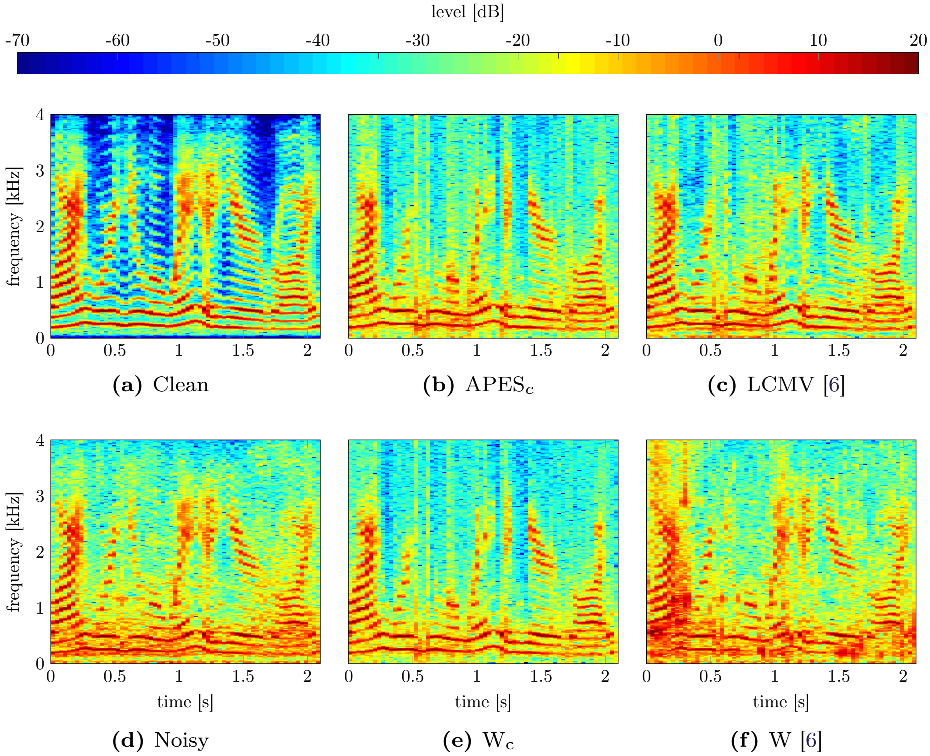


(a) Output SNR



(b) Signal reduction factor

Fig. B.9: Performance as a function of the input SNR, averaged over NOIZEUS corpus with babble noise. Parameters estimated from noisy speech signals. Voiced/unvoiced detection from noisy signal.



(a) Clean **(b)** APES$_c$ **(c)** LCMV [6]

(d) Noisy **(e)** W$_{\mathrm{c}}$ **(f)** W [6]

Fig. B.10: Spectrograms of clean, noisy and filtered speech. Babble noise giving an input SNR of 10 dB is used.

Paper C

Instantaneous Pitch Estimation with Optimal Segmentation for Non-Stationary Voiced Speech

Sidsel Marie Nørholm, Jesper Rindom Jensen and Mads Græsbøll Christensen

The paper has been submitted to the
IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2015.

in peer-review
The layout has been revised.

Abstract

In speech processing, the speech is often considered stationary within segments of 20–30 ms even though it is well known not to be true. In this paper, we take the non-stationarity of voiced speech into account by using a linear chirp model to describe the speech signal. We propose a maximum likelihood estimator of the fundamental frequency and chirp rate of this model, and show that it reaches the Cramer-Rao bound. Since the speech varies over time, a fixed segment length is not optimal, and we propose to make a segmentation of the signal based on the maximum a posteriori (MAP) criterion. Using this segmentation method, the segments are on average seen to be longer for the chirp model compared to the traditional harmonic model. For the signal under test, the average segment length is 24.4 ms and 17.1 ms for the chirp model and traditional harmonic model, respectively. This suggests a better fit of the chirp model than the harmonic model to the speech signal. The methods are based on an assumption of white Gaussian noise, and, therefore, two prewhitening filters are also proposed.

Index Terms: Harmonic chirp model, parameter estimation, segmentation, prewhitening.

1 Introduction

Parameter estimation of harmonic signals is relevant in fields such as speech processing and communication. In speech models, the speech signal is often split up into two parts, a voiced part and an unvoiced part. The voiced part of the speech signal is produced by the vibration of the vocal cords, and, therefore, has a structure with a fundamental frequency and a set of overtones given by integer multiples of the fundamental. Due to this, the voiced speech is often modelled by the harmonic model [1–4]. To estimate the parameters of the this model, it is normal to split the signal into segments of 20–30 ms [5] and perform parameter estimation of each segment separately. In most models, including the traditional harmonic model, the signal is assumed to be stationary within each frame, even though it is well known that this assumption of stationarity does not hold [5, 6]. To take the non-stationarity of speech into account, the harmonic model can be extended to a harmonic chirp model which has also been suggested in [7–9]. Here, the harmonic structure is still the foundation of the model, but the fundamental frequency is allowed to change linearly within each segment. This introduces an extra parameter to estimate, but it also introduces some benefits based on the fact that the model fits the speech signal better. Using the harmonic chirp model instead of the traditional harmonic model can, therefore, lead to better speech enhancement [10], but with a better fit of the model it is also possible to work with longer segments. Longer segments lead to better performance of the estimators, and, thereby, a smaller error on the estimated parameters can be obtained. However, the optimal segment length is very dependent on the features of the signal which are varying over time in the case of speech signals. At some time instances, the parameters are almost constant, and, in such periods, long segments can be used whereas at other points in time, the parameters will change fast and shorter segments should be used. Instead of using a fixed segment length it is, therefore, better to have a varying segment length dependent on the signal characteristics at the given point in time. In [11, 12], the signal is modelled based on linear prediction (LP), and the segment length is chosen according to a trade off between bit rate and distortion. The principle can, however, be used in connection with other criteria for choosing the segment length dependent on what is most relevant in the given situation. The noise characteristics also have an impact on the performance of parameter estimators and optimal segmentation. Most methods make an assumption of white Gaussian noise which is rarely experienced in real life scenarios. One way to address this problem is to preprocess the signal in a way that makes the noise resemble white Gaussian noise as is, e.g., done by use of the Cholesky factorisation [13].

In this paper, we propose to estimate the fundamental frequency and fundamental chirp rate by maximising the likelihood. Since this maximisation leads to a search in a two dimensional space, we suggest an iterative procedure where first a one dimensional optimisation of the chirp parameter is performed followed by a one dimensional optimisation of the fundamental frequency based on the newly found estimate of the chirp rate. The estimation process is ended by convergence of the two dimensional cost function. The parameter estimator is a continuation of [14]. The iterative procedure presents some benefits over the method suggested in [9] where an approximate cost function is introduced in order to decrease the computational load. This approximate cost function is evaluated over a two dimensional grid whereas in this paper, the original cost function is evaluated iteratively which makes the procedure suggested in this paper faster. Based on the estimated parameters, we further suggest to make an optimal segmentation based on the principle suggested in [11, 12] by adopting it to the harmonic chirp model by using the maximum a posteriori (MAP) criterion for choosing the segment length. Both the maximum likelihood estimator of the fundamental frequency and chirp rate and the MAP criterion are based on an assumption of white Gaussian noise. Therefore, we further suggest two different methods to prewhiten the signal based on noise power spectral density (PSD) estimation [15–18], generating a filter to counteract the spectral shape of the noise. The filter is either based directly on the estimated spectrum of the noise or linear prediction (LP) of the noise.

The paper is organised as follows. In Section 2, the harmonic chirp model is introduced. In Section 3, the maximum likelihood estimator of the fundamental frequency and fundamental chirp rate is derived. In Section 4, the general MAP criterion is introduced for the harmonic chirp model along with the MAP model selection criterion between the traditional harmonic model, the harmonic chirp model and the noise only model. This is followed by the segmentation principle based on the MAP criterion in Section 5. In Section 6, the two prewhitening methods are described. In Section 7, the proposed methods are tested through simulations on synthetic chirp signals and speech, and the paper is concluded in Section 6.

2 Harmonic chirp model

The harmonic chirp model is an extension of the traditional harmonic model. Therefore, the harmonics still have the same relationship, but the fundamental frequency changes linearly within a segment, and, thereby, the frequency of the l'th harmonic, $\omega_l(n)$, varies with the time index $n = n_0, ..., n_0 + N - 1$ and can be expressed as

$$\omega_l(n) = l(\omega_0 + kn), \tag{C.1}$$

where $\omega_0 = f_0/f_s 2\pi$, with f_s the sampling frequency, is the normalised fundamental frequency and k is the normalised fundamental chirp rate. The instantaneous phase, $\varphi_l(n)$, of the sinusoids are given by the integral of the instantaneous frequency as

$$\varphi_l(n) = l\left(\omega_0 n + \frac{1}{2}kn^2\right) + \phi_l, \tag{C.2}$$

where $\phi_l \in [0, 2\pi]$ is the initial phase of the l'th harmonic. This leads to the complex harmonic chirp model (HCM) for a voiced speech signal, $s(n)$:

$$s(n) = \sum_{l=1}^{L} A_l e^{j\varphi_l(n)} \tag{C.3}$$

$$= \sum_{l=1}^{L} \alpha_l e^{jl\left(\omega_0 n + k/2n^2\right)}, \tag{C.4}$$

where L is the number of harmonics and $\alpha_l = A_l e^{j\phi_l}$, $A_l > 0$ is the complex amplitude of the l'th harmonic. For speech signals the model order has to be estimated which can be done, e.g., by use of the MAP criterion introduced in Section 4 (see also [19]). The complex signal model is used instead of the real because it can ease both notation and computation. A real signal can easily be converted to a complex signal by use of the Hilbert transform [20] and without loss of information be downsampled by a factor of two.

A special case of the harmonic chirp model for $k = 0$ is then the traditional harmonic model (HM):

$$s(n) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n}. \tag{C.5}$$

Defining a vector of samples

$$\mathbf{s} = [s(n_0)\ s(n_0+1)\ \ldots\ s(n_0+N-1)]^T, \tag{C.6}$$

where $(\cdot)^T$ denotes the transpose. Note that the dependency on the index n_0 is left out for ease of notation. The signal model is then written as

$$\mathbf{s} = \mathbf{Z}\mathbf{a}, \tag{C.7}$$

where $\mathbf{Z}$ is a matrix constructed from a set of L modified Fourier vectors matching the harmonics of the signal,

$$\mathbf{Z} = [\mathbf{z}(\omega_0, k)\ \ \mathbf{z}(2\omega_0, 2k)\ \ \ldots\ \ \mathbf{z}(L\omega_0, Lk)], \tag{C.8}$$

with

$$\mathbf{z}(l\omega_0, lk) = \begin{bmatrix} e^{jl(\omega_0 n_0 + k/2 n_0^2)} \\ e^{j2l(\omega_0(n_0+1)+k/2(n_0+1)^2)} \\ \vdots \\ e^{jl(\omega_0(n_0+N-1)+k/2(n_0+N-1)^2)} \end{bmatrix}. \tag{C.9}$$

The vector $\mathbf{a}$ contains the complex amplitudes of the harmonics, $\mathbf{a} = [\alpha_1\ \alpha_2\ \ldots\ \alpha_L]^T$.

Often, the signal, we want to make parameter estimation on, is buried in noise, $v(n)$, to give the observed signal, $x(n)$,

$$x(n) = s(n) + v(n), \tag{C.10}$$

which can also be put into a vector of observed samples

$$\mathbf{x} = \mathbf{s} + \mathbf{v}, \tag{C.11}$$

where $\mathbf{x}$ and $\mathbf{v}$ are defined similarly to $\mathbf{s}$ in (C.6). For real signals as speech, the signal model will not fit the desired signal perfectly, and $\mathbf{v}$ will, therefore, also cover the part of the speech signal that does not align with the given model as, e.g., unvoiced speech during mixed excitations.

3 Estimation of frequency and chirp rate

The fundamental frequency and chirp rate are estimated by maximising the likelihood. The maximum likelihood estimates are the parameters of the model that describe the observed signal the best, i.e., the parameters that maximises the probability of the observed data, $\mathbf{x}$, given the parameters:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta}), \tag{C.12}$$

where $\boldsymbol{\theta}$ is a vector containing the parameters of the model. Under the assumption of circularly symmetric Gaussian noise, the likelihood function can be written as [19]:

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\theta}) &= \frac{1}{\pi^N \det(\mathbf{R}_v)} e^{-(\mathbf{x}-\mathbf{s})^H \mathbf{R}_v^{-1}(\mathbf{x}-\mathbf{s})} && \text{(C.13)} \\ &= \frac{1}{\pi^N \det(\mathbf{R}_v)} e^{-\mathbf{v}^H \mathbf{R}_v^{-1}\mathbf{v}}, && \text{(C.14)} \end{aligned}$$

where $\det\{\cdot\}$ denotes the determinant of the argument, $(\cdot)^H$ the Hermitian transpose and $\mathbf{R}_v = \mathbb{E}[\mathbf{v}\mathbf{v}^H]$ the noise covariance matrix, with $\mathbb{E}(\cdot)$ the mathematical expectation. Often the log likelihood is maximised instead of the likelihood

$$\ln \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = -N \ln \pi - \ln \det(\mathbf{R}_v) - \mathbf{v}^H \mathbf{R}_v^{-1}\mathbf{v}. \tag{C.15}$$

In the case of white noise, the noise covariance matrix reduces to a diagonal matrix, $\mathbf{R}_v = \sigma_v^2 \mathbf{I}_N$, where σ_v^2 is the variance of the noise signal and $\mathbf{I}_N$ is an $N \times N$ identity matrix. Thereby, the log likelihood can be reduced to

$$\ln \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = -N \ln \pi - N \ln \sigma_v^2 - \frac{1}{\sigma_v^2}||\mathbf{v}||_2^2. \tag{C.16}$$

The noise and its variance can be found using the signal model in (C.7)

$$\begin{aligned} \mathbf{v} = \mathbf{x} - \mathbf{s} &= \mathbf{x} - \mathbf{Z}\mathbf{a} \Rightarrow && \text{(C.17)} \\ ||\mathbf{v}||_2^2 &= ||\mathbf{x} - \mathbf{Z}\mathbf{a}||_2^2, && \text{(C.18)} \\ \sigma_v^2 &= \frac{1}{N}||\mathbf{x} - \mathbf{Z}\mathbf{a}||_2^2, && \text{(C.19)} \end{aligned}$$

which turns the log likelihood into

$$\ln \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = -N \ln \pi - N \ln \frac{1}{N}||\mathbf{x} - \mathbf{Z}\mathbf{a}||_2^2 - N. \tag{C.20}$$

In the estimation of the fundamental frequency and chirp rate, it is only necessary to consider terms dependent on these two parameters, and the log likelihood function can be reduced to the nonlinear least squares (NLS) estimator that minimises the error between the observed signal and the signal model:

$$\begin{aligned} \{\widehat{\mathbf{a}}, \widehat{\omega}_0, \widehat{k}\} &= \arg\min_{\mathbf{a},\omega_0,k} ||\mathbf{x} - \mathbf{s}||_2^2 && \text{(C.21)} \\ &= \arg\min_{\mathbf{a},\omega_0,k} ||\mathbf{x} - \mathbf{Z}\mathbf{a}||_2^2. && \text{(C.22)} \end{aligned}$$

Here, we are interested in the joint estimation of the fundamental frequency and chirp rate, and, therefore, the amplitudes are substituted with their least squares estimate [21],

$$\widehat{\mathbf{a}} = (\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{x}, \tag{C.23}$$

to give the estimator:

$$\begin{aligned}\{\widehat{\omega}_0, \widehat{k}\} &= \arg\min_{\omega_0,k} ||\mathbf{x} - \mathbf{Z}(\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{x}||_2^2 && \text{(C.24)}\\ &= \arg\min_{\omega_0,k} \left(\mathbf{x}^H(\mathbf{I}_N - \mathbf{Z}(\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H)\mathbf{x}\right) && \text{(C.25)}\\ &= \arg\min_{\omega_0,k} \left(\mathbf{x}^H\Pi^{\perp}\left(\omega_0, k\right)\mathbf{x}\right), && \text{(C.26)}\end{aligned}$$

where Π is an orthogonal projection matrix

$$\Pi\left(\omega_0, k\right) = \mathbf{Z}(\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H \tag{C.27}$$

and $\Pi^{\perp}$ its orthogonal complement

$$\Pi^{\perp}\left(\omega_0, k\right) = \mathbf{I}_N - \Pi\left(\omega_0, k\right). \tag{C.28}$$

This optimisation includes a two dimensional optimization over ω_0 and k. To solve the problem in a computational efficient manner, we propose to do it by iterating between two one dimensional searches [14]. First, the chirp rate in step i, k^i, is estimated using the fundamental frequency estimate from the previous iteration, $\omega_0^{(i-1)}$, $i = 1, 2, ...$

$$k^{(i)} = \arg\min_{k} \left(\mathbf{x}^H\Pi^{\perp}(\omega_0^{(i-1)}, k)\mathbf{x}\right). \tag{C.29}$$

This estimate of the chirp rate is used to find a new estimate of the fundamental frequency

$$\omega_0^{(i)} = \arg\min_{\omega_0} \left(\mathbf{x}^H\Pi^{\perp}(\omega_0, k^{(i)})\mathbf{x}\right). \tag{C.30}$$

The estimates of ω_0 and k are found by iterating between (C.29) and (C.30) until convergence of the cost function in (E.11), but could alternatively be ended by the convergence of the estimated parameters. The fundamental frequency and chirp rate minimising the cost function in (E.11) are found by searching among candidates in a grid centred at the value of the parameter from the previous iteration, $i-1$. The grid search is followed by a Dichotomous search [22] to get a refined estimate of the minimum. It is expected that the fundamental frequency estimate is close to the estimate found under the assumption of stationarity within the analysis frame. Therefore, a fundamental frequency estimate found under the traditional harmonic assumption, e.g., by using one of

Table C.1: Estimation of fundamental frequency and chirp rate.

```
for each sample
initialisation
$\omega_0^{(0)} = \omega_{0,h}$
$k^{(0)} = 0$
$\Delta k = 2\alpha_k/(K-1)$
$\Delta\omega = 2\alpha_\omega/(K-1)$
repeat
    $K = \{k^{(i-1)} - \alpha_k, \Delta k, ...., k^{i-1} + \alpha_k\}$
    $\Omega = \{\omega_0^{(i-1)} - \alpha_\omega, \Delta\omega, ...., \omega_0^{i-1} + \alpha_\omega\}$
    $k^{(i)} = \arg\min_{k \in K} \left(\mathbf{x}^H \Pi^{\perp}(\omega_0^{(i-1)}, k)\mathbf{x}\right)$
    $\omega_0^{(i)} = \arg\min_{\omega_0 \in \Omega} \left(\mathbf{x}^H \Pi^{\perp}(\omega_0, k^{(i)})\mathbf{x}\right)$
until (convergence)
```

the methods in [19], will be a good choice as initialisation of the iterations, i.e., $\omega_0^{(0)} = \omega_{0,h}$. The chirp rate is expected to be small and the first grid search is, therefore, centred around zero, i.e., $k^{(0)} = 0$. The estimation process is summarised in Table C.1.

The best obtainable performance of an unbiased estimator is given by the Cramer-Rao bound (CRB). The CRB sets a lower limit to the variance of the parameter estimate

$$\mathrm{var}(\widehat{\theta}_g) \geq [\mathcal{I}(\boldsymbol{\theta})^{-1}]_{gg}, \tag{C.31}$$

where θ_g is the g'th parameter of the parameter vector $\boldsymbol{\theta}$ of length G, $[\cdot]_{gg}$ denotes the matrix element of row g and column g, and $\mathcal{I}(\boldsymbol{\theta})$ is the Fisher information matrix (FIM) [23] of size $G \times G$:

$$[\mathcal{I}(\boldsymbol{\theta})]_{gh} = -\mathbb{E}\left\{\frac{\partial^2 \ln(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial\theta_g \partial\theta_h}\right\}. \tag{C.32}$$

Under the assumptions of white Gaussian noise and a noise covariance matrix independent of the parameters, the FIM reduces to:

$$\mathcal{I}(\boldsymbol{\theta}) = \frac{2}{\sigma_v^2}\mathrm{Re}\left\{\frac{\partial \mathbf{s}^H}{\partial \boldsymbol{\theta}} \frac{\partial \mathbf{s}}{\partial \boldsymbol{\theta}^T}\right\} \tag{C.33}$$

$$= \frac{2}{\sigma_v^2}\mathrm{Re}\left\{\mathbf{D}^H(\boldsymbol{\theta})\mathbf{D}(\boldsymbol{\theta})\right\} \tag{C.34}$$

with

$$\mathbf{D}(\boldsymbol{\theta}) = [\mathbf{d}(\omega_0)\,\mathbf{d}(k)\,\mathbf{d}(A_1)\,\mathbf{d}(\phi_1)\,\ldots\,\mathbf{d}(A_L)\,\mathbf{d}(\phi_L)], \tag{C.35}$$

$$\mathbf{d}(y) = \frac{\partial \mathbf{s}}{\partial y}. \tag{C.36}$$

For the signal model at hand, the elements of the $\mathbf{d}$ vectors are:

$$[\mathbf{d}(\omega_0)]_n = \sum_{l=1}^{L} jlnA_l e^{jl(\omega_0 n + k/2n^2)+j\phi_l}, \tag{C.37}$$

$$[\mathbf{d}(k)]_n = \sum_{l=1}^{L} \frac{1}{2} jln^2 A_l e^{jl(\omega_0 n + k/2n^2)+j\phi_l}, \tag{C.38}$$

$$[\mathbf{d}(A_l)]_n = e^{jl(\omega_0 n + k/2n^2)+j\phi_l}, \tag{C.39}$$

$$[\mathbf{d}(\phi_l)]_n = jA_l e^{jl(\omega_0 n + k/2n^2)+j\phi_l}. \tag{C.40}$$

The CRB depends on the choice of n_0. The best estimates can be obtained if the segment is centred around $n = 0$ [24], and, thereby, n_0 should be chosen as $n_0 = -(N-1)/2$ for N odd and $n_0 = -N/2$ for N even.

4 MAP criteria and model selection

Model selection and segmentation can be done with a maximum a posteriori (MAP) model selection criterion. The principle behind the MAP criterion is to choose the model, $\mathcal{M}$, that maximises the posterior probability given the observed data, $\mathbf{x}$:

$$\widehat{\mathcal{M}} = \arg\max_{\mathcal{M}} p(\mathcal{M}|\mathbf{x}). \tag{C.41}$$

Using Bayes' theorem [25] this can be rewritten as:

$$\widehat{\mathcal{M}} = \arg\max_{\mathcal{M}} \frac{p(\mathbf{x}|\mathcal{M})p(\mathcal{M})}{p(\mathbf{x})}. \tag{C.42}$$

Choosing the same prior probability, $p(\mathcal{M})$, for every model to avoid favouring any model beforehand, and noting that the probability of a given data vector, $p(\mathbf{x})$, is constant once it has been observed, the MAP estimate can be reduced to:

$$\widehat{\mathcal{M}} = \arg\max_{\mathcal{M}} p(\mathbf{x}|\mathcal{M}), \tag{C.43}$$

which is the likelihood of the observed data given the model. The likelihood is also dependent on other parameters like the fundamental frequency and the model order. As opposed to the maximum likelihood approach, in the Bayesian framework these have to be integrated out to give the marginal density of the data given the model [19]:

$$p(\mathbf{x}|\mathcal{M}) = \int_{\mathbf{\Theta}} p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})d\boldsymbol{\theta}, \tag{C.44}$$

An approximation to this integral can be found assuming high amounts of data and a likelihood that is highly peaked around the maximum likelihood estimates of $\boldsymbol{\theta}$ [7, 19, 26]

$$p(\mathbf{x}|\mathcal{M}) = \pi^{G/2}\det(\widehat{\mathbf{H}})^{-1/2} p(\mathbf{x}|\widehat{\boldsymbol{\theta}},\mathcal{M})p(\widehat{\boldsymbol{\theta}}|\mathcal{M}), \tag{C.45}$$

where $\widehat{\mathbf{H}}$ is the Hessian of the log-likelihood function evaluated at $\widehat{\boldsymbol{\theta}}$:

$$\widehat{\mathbf{H}} = -\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta},\mathcal{M})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}. \tag{C.46}$$

Now an expression for the MAP estimator can be found by taking the negative logarithm of (C.45). The term $\pi^{G/2}$ can be assumed constant for large N and is neglected, and a weak prior on $p(\boldsymbol{\theta}|\mathcal{M})$ has been used [7] to obtain the expression [19]:

$$\widehat{\mathcal{M}} = \arg\min_{\mathcal{M}} -\ln\mathcal{L}(\widehat{\boldsymbol{\theta}}|\mathbf{x}) + \frac{1}{2}\ln\det(\widehat{\mathbf{H}}). \tag{C.47}$$

This corresponds to minimising a cost function, where the first part is the likelihood from (C.16), and the second part is a model dependent penalty term.

The penalty term is found by noting that the Hessian is related to the Fisher information matrix in (C.32). Evaluating the Fisher information matrix at $\theta = \widehat{\theta}$ gives the expected value of the Hessian, and, therefore, the elements in the Hessian can be found by using (C.35)-(C.40). To ease complexity, an asymptotic expression for the Hessian can be found by looking at the elements of the matrix. The diagonal elements of the Hessian for the harmonic chirp model is given by:

$$\widehat{\mathbf{H}}_{\omega_0\omega_0} = \sum_{l=1}^{L}\frac{1}{12}(N^3 - N)l^2\widehat{A}_l^2, \tag{C.48}$$

$$\widehat{\mathbf{H}}_{kk} = \sum_{l=1}^{L}\frac{1}{960}(3N^5 - 10N^3 + 7N)l^2\widehat{A}_l^2, \tag{C.49}$$

$$\widehat{\mathbf{H}}_{A_lA_l} = N, \tag{C.50}$$

$$\widehat{\mathbf{H}}_{\phi_l\phi_l} = N\widehat{A}_l^2, \tag{C.51}$$

for N odd and $n_0 = -(N-1)/2$. From this, it is seen that when the Hessian is evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, the model order and amplitudes can be considered constant, and the Hessian is then only dependent on N. To make this dependency negligible, a diagonal normalisation matrix, $\mathbf{K}$, is introduced [19, 27]

$$\mathbf{K} = \begin{bmatrix} N^{-3/2} & & \mathbf{0} \\ & N^{-5/2} & \\ \mathbf{0} & & N^{-1/2}\,\mathbf{I}_{2L} \end{bmatrix}, \tag{C.52}$$

and, thereby,

$$\widehat{\mathbf{H}} = \mathbf{K}^{-1}\mathbf{K}\widehat{\mathbf{H}}\mathbf{K}\mathbf{K}^{-1}. \tag{C.53}$$

The definition of the elements in $\mathbf{K}$ as $N^{-x/2}$ instead of N^{-x}, $x = 1, 3, 5$, and multiplication with $\mathbf{K}$ from both sides is done to ensure that also the off-diagonal elements of $\widehat{\mathbf{H}}$ are compensated for in the right way. The determinant of the Hessian is then given by:

$$\det(\widehat{\mathbf{H}}) = \det(\mathbf{K}^{-2})\det(\mathbf{K}\widehat{\mathbf{H}}\mathbf{K}), \tag{C.54}$$

where the main dependency on N is now moved to the term $\mathbf{K}^{-2}$ whereas $\mathbf{K}\widehat{\mathbf{H}}\mathbf{K}$ is assumed small and constant for large N. Taking the natural logarithm of the determinant gives:

$$\begin{aligned} \ln\det(\widehat{\mathbf{H}}) &= \ln\det(\mathbf{K}^{-2}) + \ln\det(\mathbf{K}\widehat{\mathbf{H}}\mathbf{K}) && \text{(C.55)} \\ &= 3\ln N + 5\ln N + 2L\ln N + \mathcal{O}(1). && \text{(C.56)} \end{aligned}$$

An expression for the cost associated with the harmonic chirp model can now be found by combining the log likelihood for the harmonic chirp model in (C.20) with the penalty term in (C.56) where the term $\mathcal{O}(1)$ is ignored:

$$\begin{aligned} J_c = {} & N\ln\pi + N\ln\frac{1}{N}||\mathbf{x} - \mathbf{Z}\mathbf{a}||_2^2 + N \\ & + \frac{3}{2}\ln N + \frac{5}{2}\ln N + L\ln N. \end{aligned} \tag{C.57}$$

For the traditional harmonic model, the Hessian will not contain a term related to the chirp rate, k, and the penalty for the MAP estimator will, therefore, also be short of this term:

$$\begin{aligned} J_h = {} & N\ln\pi + N\ln\frac{1}{N}||\mathbf{x} - \mathbf{Z_0}\mathbf{a}||_2^2 + N \\ & + \frac{3}{2}\ln N + L\ln N, \end{aligned} \tag{C.58}$$

where $\mathbf{Z}_0$ equals $\mathbf{Z}$ for $k = 0$. The MAP expressions for the harmonic chirp model and the traditional harmonic model can be used to choose between the two models by choosing the one with the smallest cost. Due to Occam's razor [28], the simplest model is always preferred if the models describe the signal equally well. This is assured by the extra penalty that naturally appears in the MAP expression for the chirp model. The error between the chirp model and the observed signal has to decrease enough relative to the traditional harmonic model to outweigh this penalty term before the chirp model is favoured over the traditional harmonic model. Besides from choosing between the two different harmonic models, the MAP estimator can also be used for voiced/unvoiced

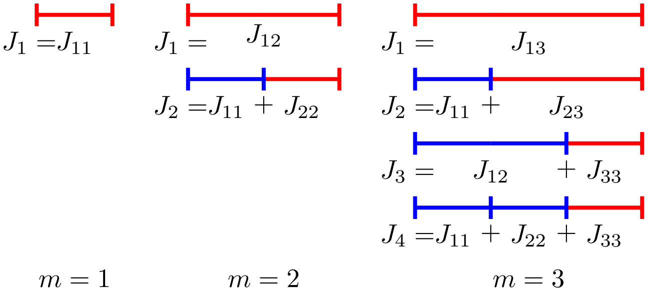


Fig. C.1: Principle of segmentation. $M = 3$. Modified from [11].

detection by determining whether a harmonic signal is present or not by comparing the two models with a zero order model,

$$J_0 = N \ln \pi + N \ln \sigma_x^2 + N, \tag{C.59}$$

where σ_x^2 is the variance of the observed signal. The voiced/unvoiced detection can also be done by use of the generalised likelihood ratio test (GLRT) [29, 30]. In this method, the ratio of the likelihood of the presence of voiced speech found based on the harmonic model to the likelihood of a noise-only signal is calculated and compared to a threshold. The method has a constant false alarm ratio (CFAR) and the threshold is, therefore, set to assure a given false alarm ratio independent of the signal-to-noise ratio (SNR). Other methods as, e.g, described in [31, 32] can also be used.

5 Segmentation

The characteristics of the observed signal is varying over time, sometimes faster than others, which means that a fixed segment length is not optimal. Using the MAP criteria, the cost associated with different segment lengths can be compared and the most optimal chosen as the one minimising (C.57). The segmentation is based on the principle in [11, 12] which is sketched in Fig. C.1. In the figure, J_{xy} is the cost of a segment starting at block x and ending at block y, with both block x and y included in the segment.

A minimal segment length, $N_{\min}$, is chosen, generating a block of $N_{\min}$ samples and dividing the signal into M blocks. Since this will give 2^{M-1} ways of segmenting the signal, a maximum number of blocks in one segment, $K_{\max}$, is also set since very long segments are highly unlikely, and setting a maximum will bound the computational complexity. The maximum number of samples in one segment is, therefore, $K_{\max} N_{\min}$. Using a dynamic programming algorithm the optimal number of blocks in a segment, k_{opt}, is found for all blocks, $m = 1, ..., M$, starting at $m = 1$ moving continuously to $m = M$. For each block,

Table C.2: Segmentation.

```
while $m \times N_{\min} \leq$ length(signal)
    $K = \min([m, K_{\max}])$
    for $k = 1 : K$
        blocks of signal to use is $m - k + 1, ...., m$
        find analytic signal and downsample
        estimate $\omega_0$ and $k$ using Table C.1
        estimate $\mathbf{a}$ and $\mathbf{Z}$ from (C.23), (C.8) and (C.9)
        calculate $J_{(m-k+1)m}$ from (C.57)
        $J(k) = \begin{cases} J_{(m-k+1)m} + J_{1(m-k)} & \text{if } m - k > 0, \\ J_{(m-k+1)m} & \text{otherwise.} \end{cases}$
    end for
    $k_{\text{opt}}(m) = \arg\min J(k)$
    $m = m + 1$
end while

backtrack
$m = M$
while $m > 0$
    number of blocks in segment is $k_{\text{opt}}(m)$
    $m = m - k_{\text{opt}}(m)$
end while
```

the cost of all new block combinations is calculated whereas old combinations are reused from earlier blocks. Relating to Fig. C.1, the red segments are calculated whereas the blue segments are reused from earlier. To decrease the number of calculations further, only a block combination minimising the cost is used in a later step, which in Fig. C.1 means that only one of J_3 and J_4 is considered for $m = 3$, corresponding to the block combination that minimised the cost at $m = 2$. When the end of the signal is reached, backtracking is used to find the optimal segmentation of the signal, starting at the last block, and jumping through the signal to the beginning. This is done by starting at $m = M$ and setting the number of blocks in the last segment of the signal to $k_{\text{opt}}(M)$. Thereby, the next segment ends at block $m = M - k_{\text{opt}}(M)$ and includes $k_{\text{opt}}(M - k_{\text{opt}}(M))$ blocks. This is continued until $m = 0$. The segmentation is summarised in Table C.2.

6 Prewhitening

The maximum likelihood estimates of the fundamental frequency and chirp rate and the MAP model selection and segmentation criteria were found under

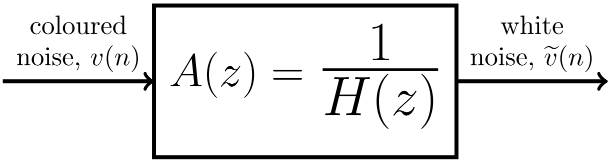


Fig. C.2: Prewhitening of noise by passing it through the filter $A(z)$.

the assumption of white Gaussian noise. However, in real life scenarios the noise is not always white. A prewhitening step is, therefore, required. The observed signal can be prewhitened by passing it through a filter that changes the noise from coloured to white. This is illustrated in Fig. C.2. In the figure $H(z)$ is a filter having a frequency response similar to the spectrum of the noise. The coloured noise can be seen as white noise filtered by a filter with coefficients given by $H(z)$. Therefore, to obtain a flat frequency spectrum of the noise, the action is reversed by dividing by $H(z)$, here denoted by $A(z)$. Of course the desired signal will also be altered by the passage through the filter. This can have an influence on the results dependent on how much the signal is changed, and what the prewhitened signal is used for. At the very best, the linear transformation of the signal will not affect the CRB of the parameter estimation.

To obtain $H(z)$, information about the noise spectrum is needed. Different methods exist to estimate the power spectral density (PSD) of the noise given a mixture of desired signal and noise [15–18]. The PSD can be used directly to generate a simple finite impulse response (FIR) filter based on the frequency coefficients of the PSD. Alternatively, also based on the PSD, linear prediction (LP) can be used to find the characteristic parts of the noise spectrum and filter the observed signal based on this. In linear prediction the present sample is estimated based on P prior samples:

$$\widehat{v}(n) = -\sum_{p=1}^{P} a_p v(n-p), \tag{C.60}$$

which leads to a filter of the form:

$$A(z) = 1 + \sum_{p=1}^{P} a_p z^{-p}. \tag{C.61}$$

After filtering, the signal is normalised to have the same standard deviation before and after the filtering. To ensure that the desired signal has a smooth evolution over time after filtering, i.e., no drastic changes in amplitude or phase, it is important that the PSD is smooth. This is ensured by most recent PSD methods where the value in one time frame is a weighted combination of the preceding time frame and an estimate from the current time frame.

7 Simulations

In the following, the different proposed methods are tested through simulations on synthetic signals and speech. The synthetic signals are made according to (C.7). Unless otherwise stated in the specific subsections, the signals were generated with $L = 10$, $A_l = 1 \, \forall \, l$, random phase, fundamental frequency, and fundamental chirp rate, in the intervals $\phi_l \in [0, 2\pi]$, $f_0 \in [100, 300]$ Hz, $k \in [-500, 500]$ Hz^2 and the sampling frequency, f_s, was set to 8000 Hz.

The speech signal used for most of the simulations was a recording with a female uttering the sentence "Why were you away a year, Roy?" sampled with a frequency of 8000 Hz. This signal is chosen because it primarily contains voiced speech. Besides from this, the fundamental frequency estimation is also tested on speech from the NOIZEUS database [33].

In most experiments, the signals were added noise with a variance calculated to fit the desired input SNR defined as

$$\text{iSNR} = \frac{\sigma_s^2}{\sigma_v^2}, \tag{C.62}$$

where σ_s^2 is the variance of the desired signal. The noise signals used are white Gaussian noise, and different types of noise from the AURORA database [34].

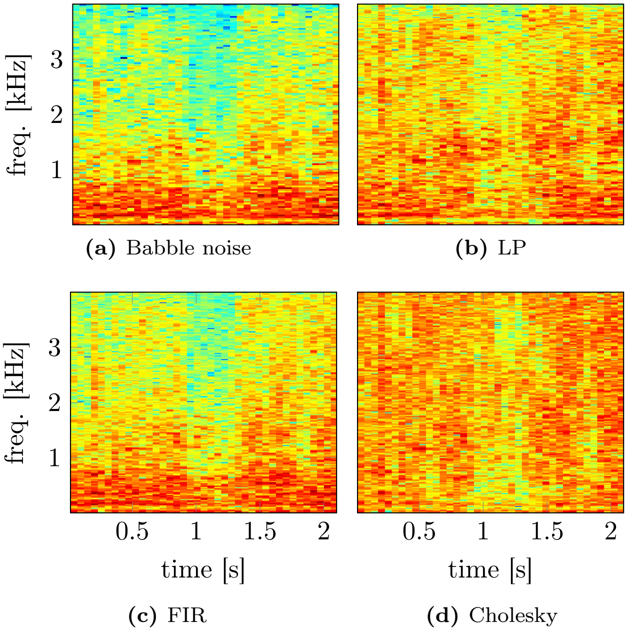
For each segment of noisy speech, the discrete-time analytic signal [20] and the parameter estimation is performed on this complex, downsampled version of the signal.

7.1 Prewhitening

The prewhitening using the FIR filter and LP is tested on "Why were you away a year, Roy?" and compared to prewhitening using Cholesky factorisation [35]. The signal is added noise at input SNRs of 0 and 10 dB, and the prewhitening filters are generated based on the noisy signal. The PSD is found using an implementation of [16] given in [15]. The spectrum of babble noise at an input SNR of 10 dB before and after prewhitening is shown in Fig. C.3. Here, it seems that the whitest noise signal is obtained with the Cholesky factorisation, followed by LP, and the FIR filter seems to make a minor change to the original noise.

The prewhitening methods are compared by means of the spectral flatness, $\mathcal{F}$, which is the ratio of the geometric mean to the arithmetic mean of the power spectrum, $S(k)$, [36]:

$$\mathcal{F} = \frac{\left(\prod_{k=0}^{K-1} S(k)\right)^{1/K}}{\frac{1}{K}\sum_{k=0}^{K-1} S(k)}. \tag{C.63}$$



(a) Babble noise **(b)** LP

(c) FIR **(d)** Cholesky

Fig. C.3: Spectrograms of babble noise before (a) and after prewhitening with (b) LPC filter, (c) FIR filter and (d) Cholesky factorisation. The four spectrograms are plotted with the same limits in dB.

The spectral flatness gives a number between zero and one, where perfect white noise has a value of one. The spectral flatness for four different noise types at 0 and 10 dB is shown in Fig. C.4 where also the spectral flatness of the original noise and a white noise signal generated with MATLABs `randn` are shown for comparison. The spectral flatness is very similar at 0 and 10 dB for all noise types using a given prewhitening method. The results confirm the tendencies seen in Fig. C.3. The Cholesky factorisation leads to the highest spectral flatness for all noise types followed by linear prediction in the case of babble, car and street noise whereas the FIR filter is better than linear prediction for exhibition noise. There is, however, big differences between the different noise types in how big the advantage is of using one prewhitening method over another. The Cholesky factorisation is clearly best in terms of whitening the noise, but as is seen in Fig. C.5, it is also the method that has the largest influence on the desired signal. Here, it seems that the LP filtering preserves the desired signal best, the FIR filter is almost just as good in that respect whereas the Cholesky factorisation clearly changes the appearance of the desired signal. Using the Cholesky factorisation for prewhitening, the signal model has to be redefined including the Cholesky matrix in the model as was

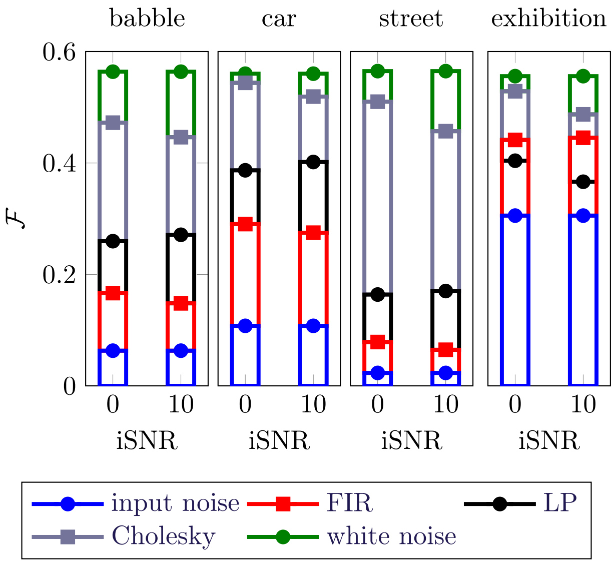
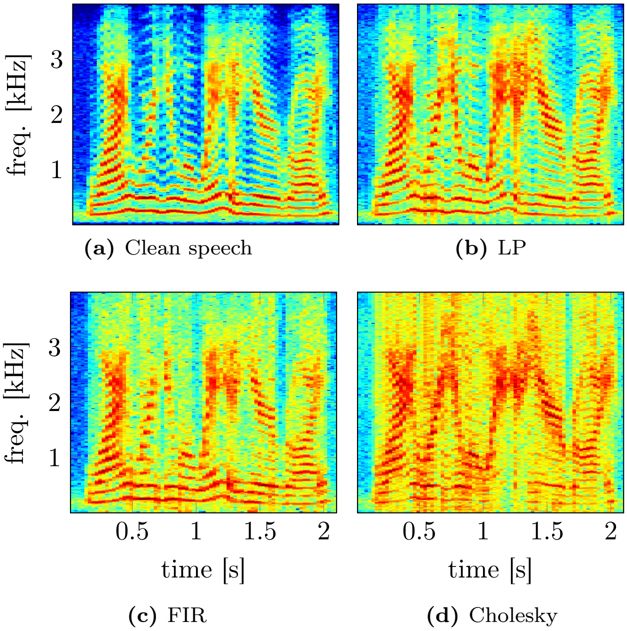


Fig. C.4: Spectral flatness, $\mathcal{F}$, at 0 and 10 dB input SNR for original noise, prewhitened noise using FIR, LPC and Cholesky factorisation. The spectral flatness for white noise is added for comparison.

done in [37]. Thus it cannot be applied directly with the proposed model, and has been excluded from the following simulations. The FIR and LP filters only change the amplitude and phase, and, thereby, they only change the complex amplitude vector **a**.

7.2 Fundamental frequency and chirp rate

The proposed estimator of fundamental frequency and chirp rate is first evaluated on synthetic chirp signals. Two experiments were made. In the first, the segment length, N, was varied from 49 to 199 samples with a fixed input SNR of 10 dB, in the second, the input SNR was varied from -10 to 10 dB with a fixed segment length of 199 samples. For each generated signal, noise was added, and an initial fundamental frequency estimate was found using a harmonic NLS estimator [19] with lower and upper limits of the search interval of 80 and 320 Hz. Hereafter, the fundamental frequency and chirp rate were estimated, and the squared error was found. This was repeated 2000 times and the mean was taken to give the mean squared error (MSE). In Figs. C.6 and C.7 the MSE as a function of N and the input SNR is shown and compared to the CRB and estimates obtained using a harmonic NLS estimator [19]. The chirp estimates reach the CRB around a segment length of 110 and at an input SNR of around -5 dB under the given settings. The harmonic estimates are



(a) Clean speech

(b) LP

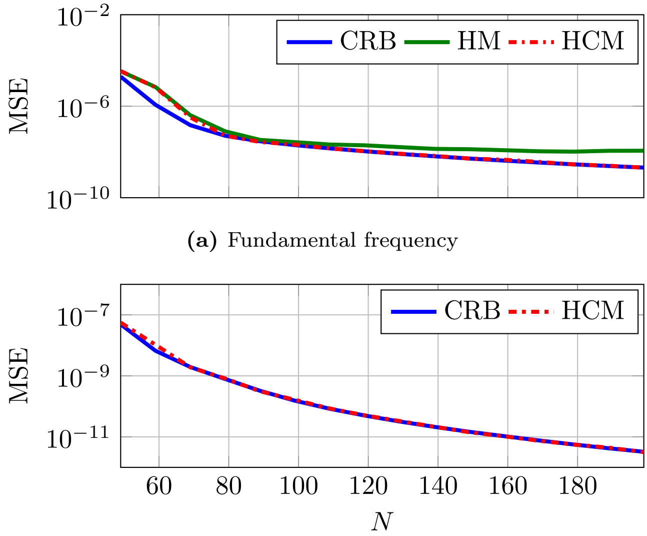
(c) FIR

(d) Cholesky

Fig. C.5: Spectrograms of speech signal before (a) and after prewhitening with (b) LPC filter, (c) FIR filter and (d) Cholesky factorisation. The four spectrograms are plotted with the same limits in dB.

close to reaching the bound too but as the CRB decreases for higher segment lengths and input SNRs, the error on the harmonic estimates do not decrease with the same rate resulting in a gap between the CRB and the estimates.

The estimator was used to estimate the fundamental frequency and chirp rate of "Why were you away a year, Roy?" with the spectrum shown in Fig. C.5a. Here, the parameters are estimated directly from the clean signal in segments with a length of 198 samples (24.8 ms). To confirm that a good initialisation is made, an example of a two dimensional cost function for a segment of the speech signal is shown in Fig. C.8. The initialisation as a combination of the harmonic fundamental frequency estimate and a chirp rate of zero is marked by a yellow cross whereas the final estimate of fundamental frequency and chirp rate is marked by a red cross. As seen, the function is locally convex around the initial and true fundamental frequency and chirp rate. Now, the parameters are estimated in steps of 5 samples. The resulting estimates are shown in Fig. C.9. The chirp rate can be interpreted as the tangent to the fundamental frequency curve in a given point. This means that the chirp rate should be negative when the fundamental frequency is decreasing, positive when it is increasing, and zero at a local maximum or minimum. To



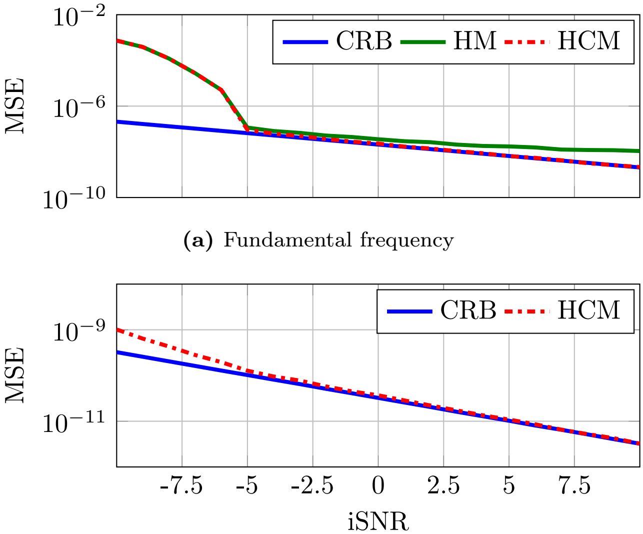
(a) Fundamental frequency

(b) Chirp rate

Fig. C.6: Mean squared error (MSE) of the fundamental frequency and chirp rate as a function of N.

illustrate this, some maxima and minima of the fundamental frequency are marked by red stars in the figure and the chirp rates at the same points in time are marked as well. The difference in fundamental frequency estimate between the traditional harmonic model and harmonic chirp model is calculated for 30 sentences from the NOIZEUS database [33] in segments with a length of 240 samples (30 ms). Only voiced speech segments are used for this, located by use of the GLRT [29, 30]. The distribution of occurrences as a function of the difference in fundamental frequency estimate is shown in Fig. C.10(a). In most cases, the fundamental frequency estimate is changed due to the use of the chirp model. The signal is reconstructed using (C.7), and the difference in SNR between using the chirp model and the harmonic model is depicted in Fig. C.10(b). This histogram is clearly skewed towards positive differences, indicating that the signal generated based on the chirp model in general bears a stronger resemblance to the desired signal than the signal generated using the traditional harmonic model. However, in some cases, it is better to use the harmonic model which could, e.g., be due to erroneous estimates of fundamental frequency and chirp rate.

The estimation is repeated after addition of noise to give an input SNR of 0 and 10 dB, but this time the parameters are only estimated once per segment of 198 samples. The estimation is done both for white Gaussian noise, babble noise and after prewhitening of the signal with babble noise using the FIR and LP



(a) Fundamental frequency

(b) Chirp rate

Fig. C.7: Mean squared error for the fundamental frequency and chirp rate as a function of the input SNR.

Table C.3: Sum of absolute error between noisy estimate and clean estimate of fundamental frequency in Hz at input SNRs of 0 and 10 dB.

	white noise	babble	FIR	LP
0 dB	585	2653	2483	1201
10 dB	167	408	714	787

filter. The sum of the absolute error between noisy and clean estimates is given in Table C.3 at 0 and 10 dB. Here, only the time interval shown in Figs. C.9 is considered since the beginning and end of the signal contain no speech, and it, therefore, does not make much sense to talk about a fundamental frequency. The white noise gives the best estimate at both 0 and 10 dB. At 0 dB, the LP prewhitened signal gives a lower error than the FIR filtered and clean babble noise whereas at 10 dB, the babble noise gives the lowest error followed by the FIR and LP filtered noise. This suggests that for the proposed ML estimator, the dominance of the desired signal at 10 dB decreases the importance of the noise shape relative to the effects of prewhitening on the signal. However, at 0 dB the noise is more dominant, and the importance of prewhitening increases.

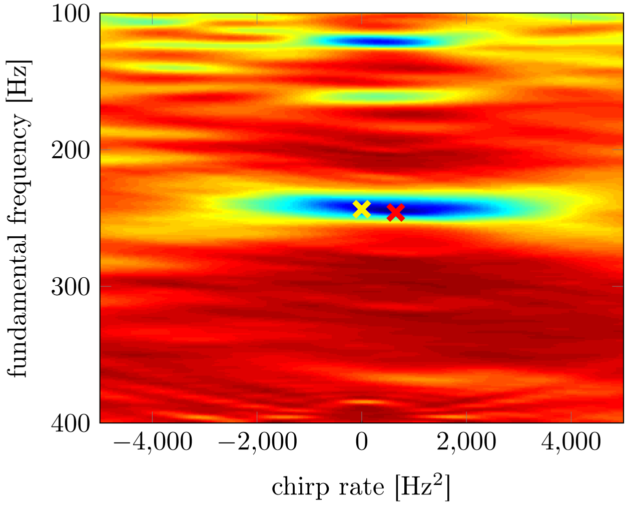


Fig. C.8: Example of a cost function for a speech signal as a function of fundamental frequency and chirp rate

7.3 Model selection

The model selection was first tested on synthetic signals in an input SNR of 10 dB white Gaussian noise. In this part, the possible models included in the test is the traditional harmonic model and the harmonic chirp model. The model selection was tested for different chirp rates and different segment lengths. For each combination of chirp rate and segment length, 2000 signals were generated, the selected model was noted for each signal and the percent of the chirp model chosen is shown in Fig. C.11. Even though all generated signals, except for the ones with a chirp rate of zero, are chirp signals, the chirp model is not chosen in all cases. As mentioned in Section 4, this is due to the extra penalty term introduced for the chirp model compared to the harmonic model. The longer the signal is, the more prone it is to be denoted as a chirp signal since the error term $||\mathbf{x} - \mathbf{Za}||_2^2$ will increase with signal length when the model does not fit, making the cost of the harmonic model greater than that for the chirp model despite the extra penalty for the chirp model.

Model selection was also performed on the speech signal "Why were you away a year, Roy?" in white Gaussian noise at different segment lengths. Here, the noise model is also included. The percentage of each model chosen is found as the number of segments in the signal labelled with a given model out of the total number of segments in the signal. The result is shown in Fig. C.12. The percentage of noise model chosen is fairly independent on the segment length since the noise model is primarily chosen in the beginning and end of the signal where there is no speech present. For short segment lengths, the

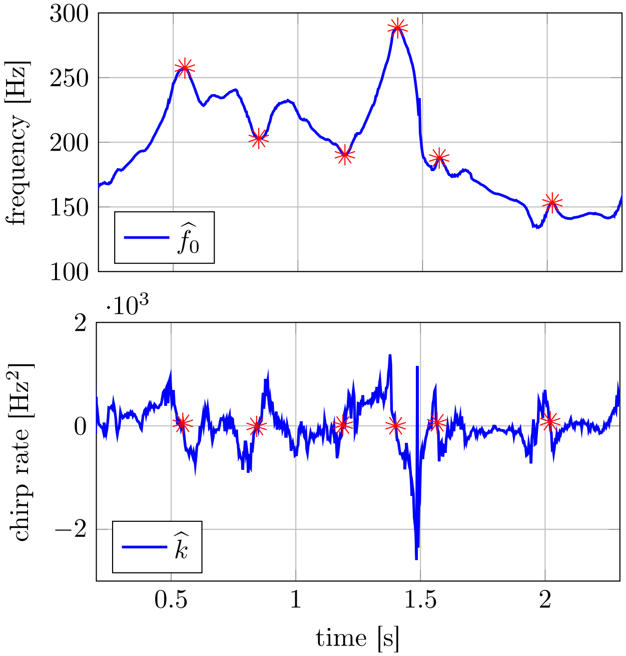
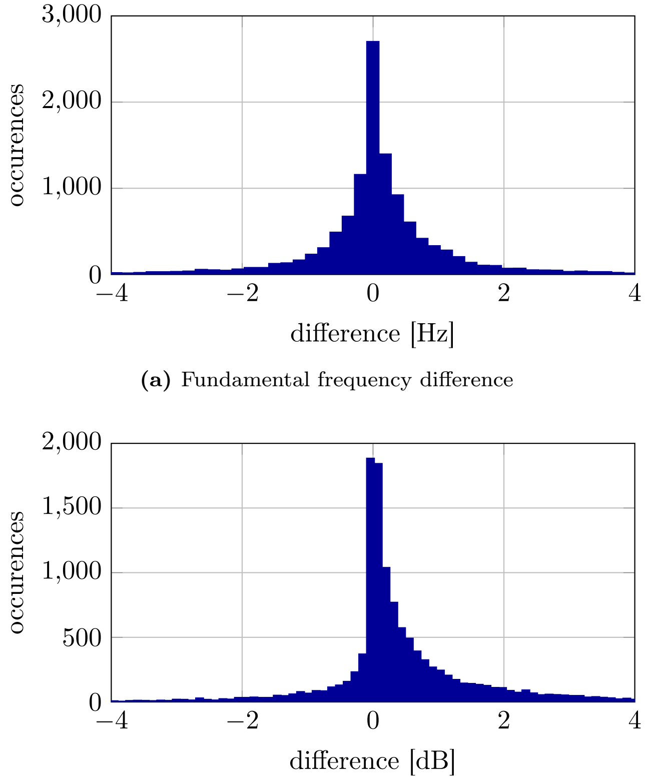


Fig. C.9: Fundamental frequency and chirp rate estimation.

harmonic model is preferred over the chirp model, but as the segment length is increased, the preferred model is the chirp model.

7.4 Segmentation

The segmentation is tested on the signal "Why were you away a year, Roy?". The signal is added white Gaussian noise to give an input SNR of 10 dB. The signal is segmented according to the harmonic chirp model and the traditional harmonic model where in both cases $N_{\text{min}} = 40$ and $K_{\text{max}} = 10$ which means that the minimum length of a segment is 40 samples (5 ms) and the maximum length of a segment is 400 samples (50 ms). A representative example of the chosen segment length as a function of time is shown in Fig. C.13. For comparison, the fundamental frequency estimate is plotted as well. In general, the chirp model gives rise to longer segment lengths than the traditional harmonic model. For this example the average segment length is 195 samples (24.4 ms) using the chirp model whereas it is 137 samples (17.1 ms) using the traditional harmonic model. A typical choice of fixed segment length is 20–30 ms [5] which would on average be a good choice when using the harmonic chirp model whereas shorter segments would be better if the traditional harmonic model is used. The longer segments for the chirp model of course also means



(a) Fundamental frequency difference

(b) Difference in reconstruction SNR

Fig. C.10: Difference in fundamental frequency estimate and reconstruction SNR between the traditional harmonic model and the harmonic chirp model.

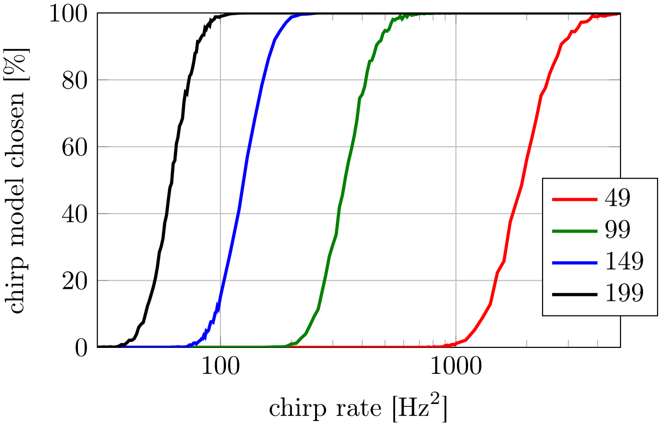


Fig. C.11: Model selection for synthetic signals as a function of the chirp rate for different segment lengths from 49 to 199.

that the total number of segments for the chirp model is lower than for the harmonic model. The chirp model is dividing the signal into 105 segments and with the harmonic model the number of segments is 150. Three areas in Fig. C.13 are marked with circles as examples of the longer segments obtained with the chirp model. In the light blue circle, the fundamental frequency is decreasing quite fast, but the change is constant over time, and, therefore, a long segment is obtained using the chirp model whereas shorter segments represent this piece when the harmonic model is used. In the purple circle, the piece of speech is divided into four segments with the chirp model. Two segments of maximum length where the fundamental frequency is almost constant and two shorter but still fairly long segments where the fundamental frequency is increasing and decreasing, respectively. For the harmonic model, there are two long segments where the fundamental frequency is close to constant, but the rest of the piece is divided into shorter segments. In the brown circle, the piece is divided into two segments using the chirp model. One piece when the fundamental frequency is decreasing and one when it is increasing. The harmonic model covers the area in the middle where the fundamental frequency is fairly constant with two somewhat long segments, but to cover the full area, shorter segments are added on both sides of the segments in the middle. The longer segments chosen for the chirp model compared to the harmonic model also suggests that the chirp model describes the signal in a better way than the traditional harmonic model since it to some extent takes the non-stationarity of the speech into account whereas the traditional harmonic model assumes the signal stationary within the segments.

The segmentation is also tested for the signal in babble noise and prewhitened babble noise also at an input SNR of 10 dB. The average segment lengths in

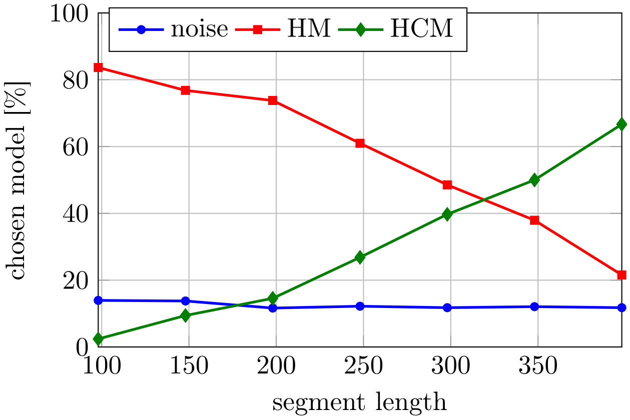


Fig. C.12: Model selection as a function of the segment length

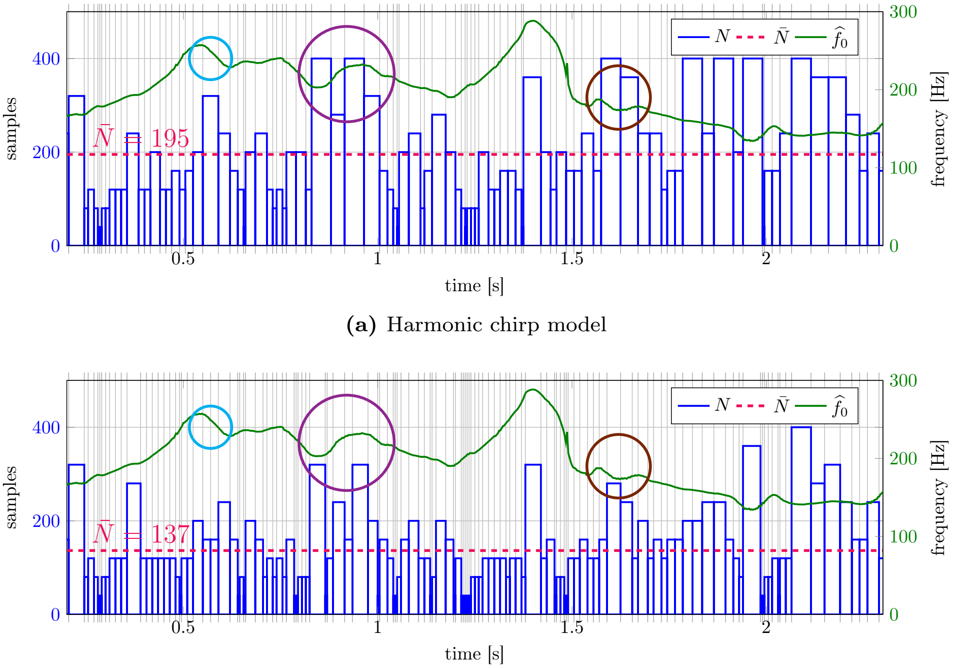
Table C.4: Average segment length, $\bar{N}$, for chirp and harmonic signal for different noise types at 10 dB.

	chirp	harmonic
babble	69	62
FIR	73	65
LP	119	91

the different cases are shown for the two models in Table C.4. In all cases, the signal is divided into longest segments when the chirp model is used. With respect to the different noise scenarios, the tendency is the same for the two models. The segments are shortest when the signal in babble noise is considered, hereafter comes the prewhitened signal using FIR filtering and the longest segments are obtained with the LP filtered signal.

8 Conclusion

Traditionally, non-stationarity, fixed segment lengths and noise assumptions have limited the performance of pitch estimators. In this paper, we take these factors into account. We described the voiced part of a speech signal by a harmonic chirp model that allows the fundamental frequency to vary linearly within each segment. We proposed an iterative maximum likelihood estimator of the fundamental frequency and chirp rate based on this model. The estimator reaches the Cramer-Rao bound and shows expected correspondence between the estimate of the fundamental frequency and fundamental chirp rate



(a) Harmonic chirp model

(b) Traditional harmonic model

Fig. C.13: Segment length as a function of time for (a) the harmonic chirp model and (b) the traditional harmonic model. The average segment length, $\bar{N}$, is marked by the red line. The total number of segments is 105 for the chirp model and 150 for the harmonic model.

of speech. Based on the maximum a posteriori (MAP) model selection criterion, the chirp model was shown to be preferred over the traditional harmonic model for long segments, suggesting that the chirp model is better at describing the non-stationary behaviour of voiced speech. Since the extent of the non-stationarity of speech changes over time, a fixed segment length is not optimal. Therefore, we also proposed to vary the segment length based on the MAP criterion. Longer segments were obtained when the chirp model was used compared to the traditional harmonic model, again suggesting a better fit of the model to the speech. The maximum likelihood and MAP estimators are based on an assumption of white Gaussian noise. However, in real life the noise is rarely white. Therefore, we also suggested two filters to prewhiten the noise, a simple FIR filter and one based on linear prediction (LP). They both have a minor influence on the speech signal, but the LP filter is a more effective prewhitener than the FIR filter.

References

[1] D. Ealey, H. Kelleher, and D. Pearce, “Harmonic tunnelling: tracking non-stationary noises during speech,” in *Proc. Eurospeech*, Sep. 2001, pp. 437–440.

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

[3] T. Nilsson, S. I. Adalbjornsson, N. R. Butt, and A. Jakobsson, “Multi-pitch estimation of inharmonic signals,” in *Proc. European Signal Processing Conf.*, 2013, pp. 1–5.

[4] M. Krawczyk and T. Gerkmann, “STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1931–1940, 2014.

[5] L. Deng and D. O’Shaughnessy, *Speech processing: a dynamic and optimization-oriented approach.* CRC Press, 2003.

[6] F. R. Drepper, “A two-level drive-response model of non-stationary speech signals,” *Nonlinear Analyses and Algorithms for Speech Processing*, vol. 1, pp. 125–138, Apr. 2005.

[7] P. M. Djuric, “A model selection rule for sinusoids in white gaussian noise,” *IEEE Trans. Signal Process.*, vol. 44, no. 7, pp. 1744–1751, 1996.

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

[9] Y. Doweck, A. Amar, and I. Cohen, “Joint model order selection and parameter estimation of chirps with harmonic components,” *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1765–1778, Apr. 2015.

[10] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, “Enhancement of non-stationary speech using harmonic chirp filters,” in *Proc. Interspeech*, Sep. 2015, accepted for publication.

[11] P. Prandoni, M. M. Goodwin, and M. Vetterli, “Optimal time segmentation for signal modeling and compression,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1997, pp. 2029–2032.

[12] P. Prandoni and M. Vetterli, “R/D optimal linear prediction,” *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 646–655, 2000.

[13] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP J. on Advances in Signal Processing*, vol. 2007, no. 1, p. 24, Jun. 2007.

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

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

[16] K. V. Sørensen and S. V. Andersen, "Speech enhancement with natural sounding residual noise based on connected time-frequency speech presence regions," *EURASIP J. on Advances in Signal Processing*, vol. 2005, no. 18, pp. 2954–2964, 2005.

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

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

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

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

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

[22] A. Antoniou and W. S. Lu, *Practical Optimization - Algorithms and Engineering Applications.* Springer Science+Business Media, 2007.

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

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

[25] C. M. Bishop, *Pattern recognition and machine learning.* Springer, 2006, vol. 1.

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

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

[28] D. J. MacKay, *Information theory, inference and learning algorithms.* Cambridge university press, 2003.

[29] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory.* Prentice Hall, Inc., 1998.

[30] E. Fisher, J. Tabrikian, and S. Dubnov, “Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 2, pp. 502–510, 2006.

[31] K. I. Molla, K. Hirose, N. Minematsu, and K. Hasan, “Voiced/unvoiced detection of speech signals using empirical mode decomposition model,” in *Int. Conf. Information and Communication Technology*, Mar. 2007, pp. 311–314.

[32] Y. Qi and B. R. Hunt, “Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier,” *IEEE Trans. Speech Audio Process.*, vol. 1, no. 2, pp. 250–255, 1993.

[33] Y. Hu and P. C. Loizou, “Subjective comparison and evaluation of speech enhancement algorithms,” *Speech Communication*, vol. 49, no. 7–8, pp. 588 – 601, 2007.

[34] D. Pearce and H. G. Hirsch, “The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *Proc. Int. Conf. Spoken Language Process.*, Oct 2000.

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

[36] N. S. Jayant and P. Noll, *Digital coding of wafeforms.* Prentice-Hall, 1984.

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

Paper D

On the Influence of Inharmonicities in Model-Based Speech Enhancement

Sidsel Marie Nørholm, Jesper Rindom Jensen and Mads Græsbøll Christensen

The paper has been published in the
Proc. European Signal Processing Conference, 2013.

Abstract

In relation to speech enhancement, we study the influence of modifying the harmonic signal model for voiced speech to include small perturbations in the frequencies of the harmonics. A perturbed signal model is incorporated in the nonlinear least squares method, the Capon filter and the amplitude and phase estimation filter. Results show that it is possible to increase the performance, in terms of the signal reduction factor and the output signal-to-noise ratio, at the cost of increased complexity in the estimation of the model parameters. It is found that the perturbed signal model performs better than the harmonic signal model at input signal-to-noise ratios above approximately -10 *dB, and that they are equally good below.*

Index Terms: Single-channel speech enhancement, perturbed signal models, inharmonicity, parameter estimation.

1 Introduction

In systems such as mobile phones, teleconferencing systems and hearing aids, noise interferes with the speech signal which has a detrimental effect on the quality of the resulting signal. Speech enhancement is therefore an important component in such systems. Speech enhancement can be performed using different approaches. A common one is filtering based on the noise statistics, e.g., using the Wiener filter. This method is very vulnerable to nonstationary noise because the problem of estimating noise statistics in the presence of speech is non-trivial [1, 2]. Another approach is to optimise filtering by assuming a model of the speech signal, as for example the harmonic signal model used in [2–6]. However, some problems arise when the harmonic signal model is used. The first is that only the voiced part of the speech signal can be modelled by a harmonic signal model. A second is due to the voiced speech being quasistationary, which means that the fundamental frequency changes over time. To minimise the effect of this, the processing is done on small segments, where the signal can be assumed periodic. A third problem is that voiced speech is not perfectly harmonic [7]. There are small perturbations in the frequencies of the harmonics and therefore they do not coincide completely with the harmonics of the assumed model. This causes unwanted distortion in the resulting speech signal when using a signal driven approach. The phenomenon of inharmonicity is well known from musical instruments, where the perturbations of the harmonics are very well defined and have to be taken into account, for example in the tuning of pianos [8]. Inharmonic models are also used in [6, 9] for fundamental frequency estimation in musical signals, but the research of the influence of inharmonicities in speech is very sparse. The inharmonicity in voiced speech is not as predictable as in musical instruments and a less restrictive model is therefore used in speech, (see e.g. [5, 7]). Inharmonicities are taken into account in the estimation of the amplitudes of the harmonics in [10], but the influence of using a perturbed signal model on the filter performance in speech enhancement has not been studied.

The purpose of this paper is, therefore, to investigate whether using a perturbed signal model will have an effect on filter performance, in terms of the signal reduction factor and the output signal-to-noise ratio (oSNR). The perturbations in synthetic signals and a set of voiced speech signals are estimated by incorporating the perturbed signal model in a nonlinear least squares (NLS) method [11] and the Capon and amplitude and phase estimation (APES) filters [12]. The estimated perturbations are then used in filtering of the signals with the APES filter [13] in order to find the gain in signal reduction factor and oSNR when compared to filtering based on the harmonic signal model.

In Section 2, the used signal model is presented along with the applied methods for estimation of the perturbations and filtering. In Section 3, the choices for the setup of experiments are explained followed by the results in Section 5, and Section 6 concludes the work.

2 Methods

2.1 Signal model

A commonly used model of N samples of voiced speech or musical instrument recordings is given by a sum of complex sinusoids, $s(n)$, corrupted by noise, $e(n)$, as

$$x(n) = \sum_{l=1}^{L} a_l e^{j\psi_l n} + e(n) = s(n) + e(n), \tag{D.1}$$

where L is the model order. The l'th complex sinusoid has frequency ψ_l and complex amplitude $a_l = A_l e^{j\phi}$ with $A_l > 0$ and ϕ_l being the real amplitude and phase, respectively. The noise term, $e(n)$, is assumed to be zero mean and complex. Measurements of speech are real valued but can be converted to the complex representation by use of the Hilbert transform and be downsampled by a factor of two if N is sufficiently large [5].

Defining a subvector of samples $\mathbf{x}(n) = [\, x(n) \; x(n-1) \; \ldots \; x(n-M+1)\,]^T$, where $M \leq N$ and $(\cdot)^T$ denotes the transpose, the signal model can be written as

$$\mathbf{x}(n) = \mathbf{Z} \begin{bmatrix} e^{-j\psi_1 n} & & 0 \\ & \ddots & \\ 0 & & e^{-j\psi_L n} \end{bmatrix} \mathbf{a} + \mathbf{e}(n), \tag{D.2}$$

where $L < M$ and $\mathbf{Z} \in \mathbb{C}^{M \times L}$ is a matrix with Vandermonde structure given by

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}(\psi_1) & \mathbf{z}(\psi_2) & \dots & \mathbf{z}(\psi_L) \end{bmatrix}, \tag{D.3}$$

$$\mathbf{z}(\psi_l) = \begin{bmatrix} 1 & e^{-j\psi_l} & \dots & e^{-j\psi_l(M-1)} \end{bmatrix}^T, \tag{D.4}$$

$\mathbf{a} = [a_1 \dots a_L]^T$ is a vector containing the complex amplitudes of the signal and $\mathbf{e}(n)$ is defined like $\mathbf{x}(n)$, but containing the noise terms $e(n)$.

Often, voiced speech is characterised using a harmonic signal model obtained by setting $\psi_l = \omega_0 l$. The harmonics are then exact multiples of the fundamental frequency, ω_0. In many musical instruments, the frequencies of the harmonics deviate slightly in a very predictable manner, leading to $\psi_l = \omega_0 l \sqrt{1 + Bl^2}$, where $B \ll 1$ is an instrument dependent stiffness parameter [5]. In speech, perturbations of the harmonics are also present, however, they are not as predictable as in music, leading to a less restrictive model for speech with [5].

$$\psi_l = \omega_0 l + \Delta_l. \tag{D.5}$$

Here, the perturbations, Δ_l, are assumed to be small and evenly distributed in the interval $P_l = [-\delta_l, +\delta_l]$, where δ_l is a small and positive number. Further, it is assumed that $\psi_l < \psi_k \,\forall\, l < k$.

The considered problem can either be solved by estimating ψ_l and from this find estimates of ω_0 and Δ_l [14], or the fundamental frequency can be estimated first and thereafter Δ_l. The second approach is taken in this paper and the fundamental frequency is therefore assumed known. Further, the model order is assumed to be known as well. Both the fundamental frequency and the model order can be found, e.g., using one of the methods in [13].

2.2 Nonlinear least squares method

The maximum a posteriori estimatior, which is asymptotically optimal, will, under the assumption of white Gaussian noise and a uniform distribution of Δ_l in P_l, reduce to the NLS method [5]. NLS minimises the error between the recorded data and the signal model from (D.2) with $M = N$ [5]

$$\{\hat{\Delta}_l\} = \arg \min_{\mathbf{a}, \{\Delta_l \in P_l\}} \|\mathbf{x}(n) - \mathbf{Za}\|_2^2, \tag{D.6}$$

with $\|\cdot\|_2$ denoting the ℓ_2-norm. Minimisation of (D.6) with respect to $\mathbf{a}$ followed by insertion of the result in (D.6) will lead to the concentratred NLS estimator of the perturbations given by [5]

$$\{\hat{\Delta}_l\} = \arg \max_{\{\Delta_l \in P_l\}} \mathbf{x}^H \mathbf{Z} (\mathbf{Z}^H \mathbf{Z})^{-1} \mathbf{Z}^H \mathbf{x}, \tag{D.7}$$

where $(\cdot)^H$ denotes the Hermitian transpose.

When the noise is colored or when several speakers are present, the NLS estimator might not be the optimal choice and therefore it is instructive to look at other estimation methods as well.

2.3 Capon filter

The Capon filter is designed to minimise the output of the filter while having unit gain at the harmonic frequencies. This minimisation problem can be expressed as [5]

$$\min_{\mathbf{h}} \mathbf{h}^H \mathbf{R}_x \mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^H \mathbf{Z} = \mathbf{1}, \tag{D.8}$$

where $\mathbf{h} = [h(0)\, h(1) \ldots h(M-1)]^H$ is the filter response, $\mathbf{1} = [1 \ldots 1]^T$ and $\mathbf{R}_x$ is the covariance matrix of $\mathbf{x}$ defined as

$$\mathbf{R}_x = \mathrm{E}\{\mathbf{x}(n)\mathbf{x}^H(n)\}, \tag{D.9}$$

with $\mathrm{E}\{\cdot\}$ denoting statistical expectation. When $s(n)$ and $e(n)$ are uncorrelated, the covariance matrix of $\mathbf{x}$ is given by the sum of the covariance matrices of the signal, $\mathbf{R}_s$, and the noise, $\mathbf{R}_e$, i.e., $\mathbf{R}_x = \mathbf{R}_s + \mathbf{R}_e$. However, none of these are known and $\mathbf{R}_x$ has to be estimated as, e.g.,

$$\widehat{\mathbf{R}}_x = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{x}(n)\mathbf{x}^H(n). \tag{D.10}$$

The filter that minimises (D.8) is given by [5]

$$\mathbf{h} = \mathbf{R}_x^{-1}\mathbf{Z}(\mathbf{Z}^H\mathbf{R}_x^{-1}\mathbf{Z})^{-1}\mathbf{1}. \tag{D.11}$$

By maximising the output power of this filter, the perturbations can be estimated as

$$\{\hat{\Delta}_l\} = \arg \max_{\{\Delta_l \in P_l\}} \mathbf{1}^H(\mathbf{Z}^H\mathbf{R}_x^{-1}\mathbf{Z})^{-1}\mathbf{1}, \tag{D.12}$$

2.4 Amplitude and phase estimation filter

The APES filter uses the same principle as the Capon filter. The only difference is that another covariance matrix is used in (D.8) which is estimated by subtracting from $\mathbf{R}_x$ the covariance corresponding to the part of $\mathbf{x}$ that resembles the signal model [13]

$$\widehat{\mathbf{R}}_e = \widehat{\mathbf{R}}_x - \mathbf{G}^H\mathbf{W}^{-1}\mathbf{G}, \tag{D.13}$$

with

$$\mathbf{G} = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{w}(n)\mathbf{x}^H(n), \tag{D.14}$$

$$\mathbf{W} = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{w}(n)\mathbf{w}^H(n), \tag{D.15}$$

where $\mathbf{w}(n) = \left[e^{j\psi_1 n} \; \ldots \; e^{j\psi_L n} \right]^T$.

The optimisation problem for the APES filter is then given by (D.8) with $\mathbf{R}_x$ replaced by $\widehat{\mathbf{R}}_e$ and the solutions for the optimal filter and the perturbations are given by (D.11) and (D.12) also with $\mathbf{R}_x$ replaced by $\widehat{\mathbf{R}}_e$.

2.5 Numerical optimisation

The estimation of the perturbations by means of (D.7) or (D.12) is a multidimensional, nonlinear and nontrivial problem. Direct estimation is therefore not feasible [11] and approximate solutions have been found as explained in what follows.

The perturbations are found one at a time by a grid search in the intervals P_l. An approximate position of the maximum is found at first, followed by a Fibonacci search [15] to give an increased resolution. If the cost functions in (D.7) and (D.12) for a given harmonic have no peak inside P_l, the perturbation is set to zero.

The NLS algorithm needs information about the perturbations of all harmonics in order to find the minimum distance between $\mathbf{x}(n)$ and the signal model $\mathbf{Za}$ in (D.6). In the first approach, denoted NLS-I, the perturbations are initialised with zeros and continuously updated with the estimated values of the perturbations. In the second approach, denoted NLS-II, the perturbations are initialised with the correct values of the perturbations and only the value of the perturbation under investigation is changed. With this second approach, the estimation of the perturbations is not influenced by errors in the frequencies of the other harmonics. Estimates based on NLS-II are therefore expected to reach the Cramér-Rao bound (CRB) and can in that case be used to bound the performance of other methods. It will of course only be possible to use NLS-II on synthetic signals where the perturbations are known. Using the Capon and APES filters for estimation, it is found that the best results are obtained using a single order filter fitted to the harmonic under investigation, compared to using a filter of order L. Therefore, first order filters have been used.

3 Experimental setup

The different ways to estimate Δ_l were evaluated through Monte Carlo simulations (MCS). A signal of the form (D.1) with $\{\psi_l\}$ given by (D.5) was generated and the performance of the different methods was evaluated by means of the mean squared error (MSE), $\frac{1}{LK}\sum_{l=1}^{L}\sum_{k=1}^{K}(\Delta_{l,k}-\hat{\Delta}_{l,k})^2$, where K is the number of MCS. The MSE was evaluated as a function of the input signal-to-noise ratio (iSNR) and the number of samples, N, and compared to the CRB for unconstrained frequency estimation [11].

The signal was generated with $L = 5$, $A_l = 1 \;\forall\; l$, random phase, fundamental frequency and perturbations in the intervals $\phi_l \in [0, 2\pi]$, $f_0 \in [150, 250]$ Hz, $\Delta_l \in [-15, 15]$ Hz, and δ_l was chosen to be 30 Hz. The Fibonacci search was performed with 14 iterations. The noise was white Gaussian with a standard deviation calculated from the desired iSNR. When N was varied, the iSNR was set to 10 dB, whereas when the iSNR was varied, N was fixed at 200. In the Capon and APES filters, the filter length was set to $\lfloor N/4 \rfloor$, with $\lfloor\cdot\rfloor$ denoting the floor operator. According to [4], this should be a good choice of filter length for both filter types. The number of MCS was $K = 500$. The importance of including perturbations in the filter design was tested by making APES filters with the estimated perturbations included and comparing them to a filter based on the harmonic assumption, $\Delta_l = 0 \,\forall\, l$. The APES filter was chosen since it was found to perform better than the Capon filter, when filtering based on already estimated frequency components is considered, which is consistent with frequency and amplitude estimation results in [12]. The performance of the filters with a perturbed and a harmonic signal model was evaluated by calculation of the signal reduction factor, $\xi_{\text{sr}}(\mathbf{h})$, and the oSNR($\mathbf{h}$) given by [2]

$$\xi_{\text{sr}}(\mathbf{h}) = \frac{\sigma_s^2}{\sigma_{s,\text{nr}}^2} = \frac{\sigma_s^2}{\mathbf{h}^H\mathbf{R}_s\mathbf{h}}, \tag{D.16}$$

$$\text{oSNR}(\mathbf{h}) = \frac{\sigma_{s,\text{nr}}^2}{\sigma_{e,\text{nr}}^2} = \frac{\mathbf{h}^H\mathbf{R}_s\mathbf{h}}{\mathbf{h}^H\mathbf{R}_e\mathbf{h}}, \tag{D.17}$$

where σ_s and $\sigma_{s,\text{nr}}$ are the variances of the signal before and after filtering and $\sigma_{e,\text{nr}}$ is the variance of the noise after filtering. Without signal distortion, the variance of the desired signal before and after filtering is the same, and, therefore, $\xi_{\text{sr}}(\mathbf{h})$ should preferably be one. However, even though $\xi_{\text{sr}}(\mathbf{h}) = 1$, the signal can still be distorted in subbands [2]. Further, better performance after filtering requires oSNR($\mathbf{h}$) > iSNR.

In order to test the perturbed signal model on voiced speech, recordings from the Keele database [16] were used. Four different speakers were used, two men and two women. The speech signal was downsampled to have a sample frequency of 8 kHz and divided into four non-overlapping segments, one for each speaker. Voiced sections and uncertain voiced sections with periodicity in the laryngograph were treated as voiced speech and extracted from the speech signal. Hereafter, voiced speech segments with a length shorter than $3N$ were discarded. In total, the performance measures were calculated for 49013 samples of voiced speech and averaged. Random white noise was added to give the desired iSNR and the performance was evaluated for the harmonic signal model and for perturbations estimated with NLS-I and Capon. Since the lowest fundamental frequency in the speech signal was 57 Hz, δ_l was set to 25 Hz.

4 Experimental results

The MSEs of the estimated perturbations were averaged over all harmonics and are shown in Fig. D.1 as a function of N and the iSNR. NLS-II reaches the CRB for all N, whereas NLS-I and Capon follow the same course from 100 samples and up with a small but constant gap to the CRB. The APES filter does not perform well for estimation of the perturbations, as was also found in [12] in the case of fundamental frequency estimation. No method reaches the CRB at low iSNRs, but above 0 dB the tendency is the same as when N was varied. It should be kept in mind, that when no peak was found in the search interval, the perturbation was set to zero, which is seen to have an influence on the result at low iSNRs as well as for the APES filter at $N = 50$.

The performance measures according to the perturbations found in Fig. D.1 are shown in Fig. D.2 along with the performance of a filter based on the harmonic signal model, i.e., $\Delta_l = 0 \, \forall \, l$. NLS-I, NLS-II and Capon perform equally well and better than both APES and the harmonic signal model when the sample length is larger than 50 and the iSNR is larger than -10 dB. The similarity between the performance using NLS-I, NLS-II and Capon means that it is not crucial to use an estimation method for the perturbations that reaches the CRB. The signal distortion is clearly decreased when taking perturbations into account. When the perturbations are estimated with NLS-I, NLS-II and Capon, $\xi_{\text{sr}}(\mathbf{h})$ is very close to 0 dB independently of N and iSNR, whereas it is increasing as a function of both N and iSNR when a harmonic signal model is used. The $\text{oSNR}(\mathbf{h})$ is also increased using the perturbed signal model. When using NLS-I instead of the harmonic signal model, the gains in $\text{oSNR}(\mathbf{h})$ are 3.1 dB and 10.5 dB at iSNRs of 0 dB and 10 dB, respectively. The performance on real speech is shown in Fig. D.3 as a function of the iSNR. The tendency here is the same as in the case of synthetic signals, and the perturbed signal model leads to improvements in both $\xi_{\text{sr}}(\mathbf{h})$ and $\text{oSNR}(\mathbf{h})$. The speech signal is more distorted than the synthetic signal in Fig. D.2, but, nevertheless, when using NLS-I, $\xi_{\text{sr}}(\mathbf{h})$ is lowered by 2.1 dB and 3.4 dB compared to the harmonic signal model at 0 dB and 10 dB, respectively. The gain in $\text{oSNR}(\mathbf{h})$ is 2.2 dB and 3.8 dB at the same iSNRs.

5 Conclusion

The influence of using the perturbed signal model as a basis for filtering of voiced speech signals was investigated and evaluated by means of the signal reduction factor and output signal-to-noise ratio. It was found that the performance was increased for input signal-to-noise ratios above approximately -10 dB when compared to the harmonic signal model. The perturbed and the harmonic signal models perform equally well for input signal-to-noise ratios below -10 dB. The perturbed signal model definitely has a potential of increasing the quality of the filtered speech signal, but with the perturbations found by grid searches, it comes with the cost of increased complexity in the estimation process.

References

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

[2] J. R. Jensen, “Enhancement of periodic signals: with applications to speech signals,” Ph.D. dissertation, Aalborg University, Jul. 2012.

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

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

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

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

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

[8] R. A. Rasch and V. Heetvelt, “String inharmonicity and piano tuning,” *Music Perception: An Interdisciplinary Journal*, vol. 3, no. 2, pp. 171–189, Winter 1985.

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

[10] Y. Pantazis, O. Rosec, and Y. Stylianou, “Iterative estimation of sinusoidal signal parameters,” *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 461–464, Feb. 2010.

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

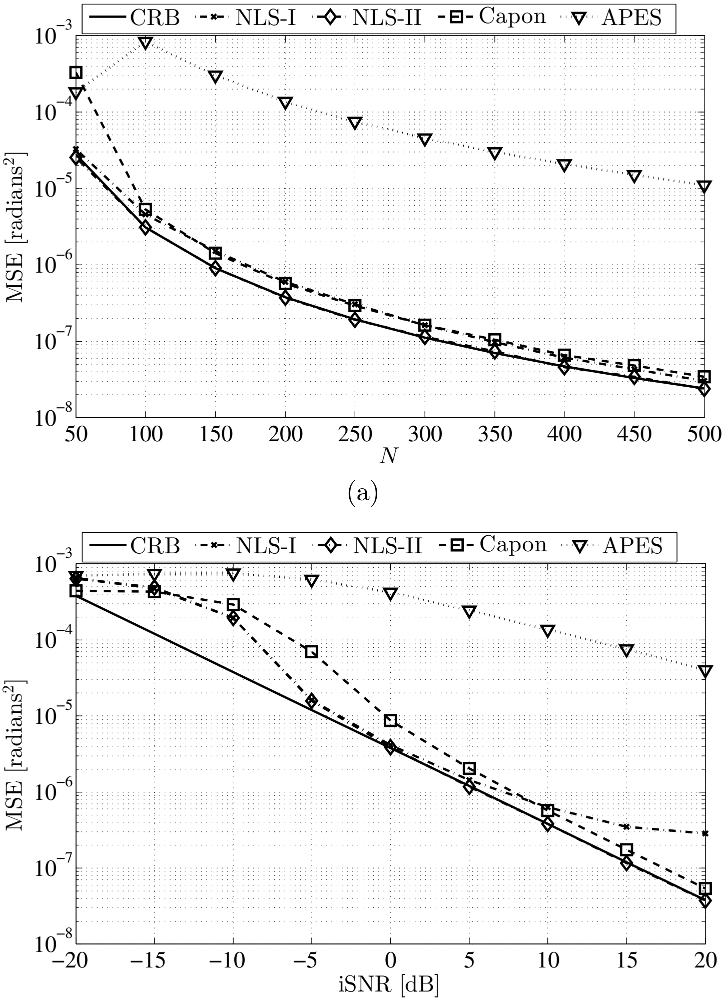
[12] A. Jakobsson and P. Stoica, “Combining Capon and APES for estimation of spectral lines,” *Circuits, Systems and Signal Processing*, vol. 19, pp. 159–169, Mar. 2000.

[13] M. G. Christensen and A. Jakobsson, “Optimal filter designs for separating and enhancing periodic signals,” *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

[14] H. Li, P. Stoica, and J. Li, “Computationally efficient parameter estimation for harmonic sinusoidal signals,” *Signal Process.*, vol. 80, no. 9, pp. 1937 – 1944, Sep. 2000.

[15] A. Antoniou and W. S. Lu, *Practical Optimization - Algorithms and Engineering Applications.* Springer Science+Business Media, 2007.

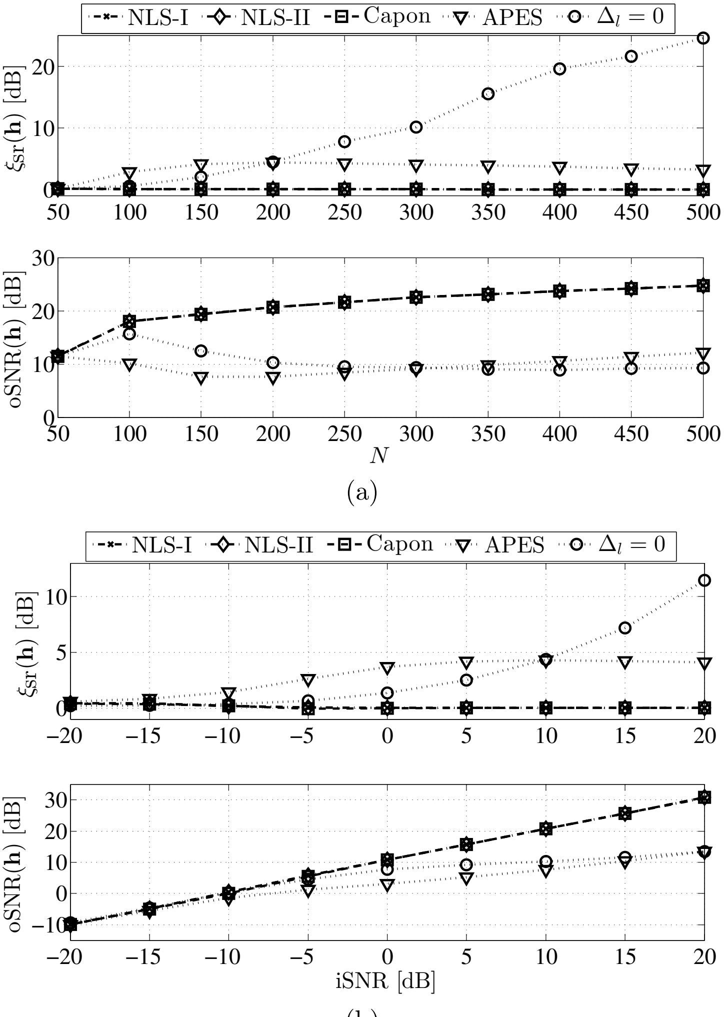
[16] F. Plante, G. F. Meyer, and W. A. Ainsworth, “A pitch extraction reference database,” in *Proc. Eurospeech*, Sep. 1995, pp. 837–840.



(a)

(b)

Fig. D.1: Mean squared error (MSE) of the estimated perturbations as a function of (a) N and (b) iSNR.



(a)

(b)

Fig. D.2: Performance measures of synthetic signal as a function of (a) N and (b) iSNR.

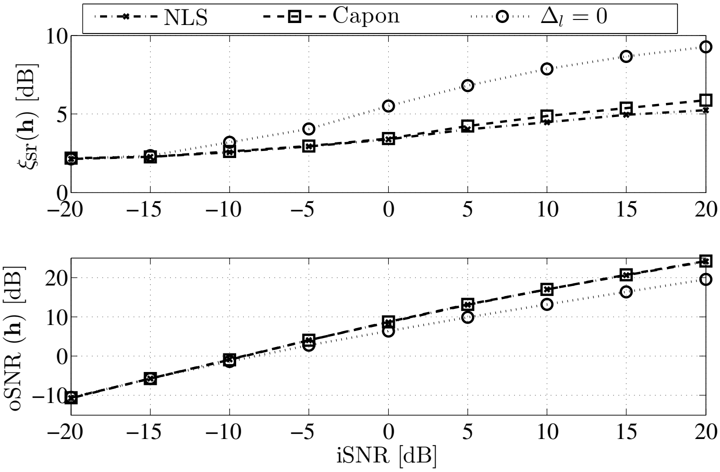


Fig. D.3: Performance measures of speech signal as a function of iSNR.

Paper E

Spatio-Temporal Audio Enhancement Based on IAA Noise Covariance Matrix Estimates

Sidsel Marie Nørholm, Jesper Rindom Jensen and Mads Græsbøll Christensen

The paper has been published in the
Proc. European Signal Processing Conference, 2014.

Abstract

A method for estimating the noise covariance matrix in a multichannel setup is proposed. The method is based on the iterative adaptive approach (IAA), which only needs short segments of data to estimate the covariance matrix. Therefore, the method can be used for fast varying signals. The method is based on an assumption of the desired signal being harmonic, which is used for estimating the noise covariance matrix from the covariance matrix of the observed signal. The noise covariance estimate is used in the linearly constrained minimum variance (LCMV) filter and compared to an amplitude and phase estimation (APES) based filter. For a fixed number of samples, the performance in terms of signal-to-noise ratio can be increased by using the IAA method, whereas if the filter size is fixed and the number of samples in the APES based filter is increased, the APES based filter performs better.

Index Terms: Speech enhancement, iterative adaptive approach, multichannel, covariance estimates, harmonic signal model.

1 Introduction

In many applications such as teleconferencing, surveillance systems and hearing aids, it is desirable to extract one signal from an observation of the desired signal buried in noise. This can be done in several ways, in general separated in three groups: the spectral-subtractive methods, the statistical-model-based methods and the subspace methods [1]. In this work, we focus on the filtering methods, which are in the group of statistical-model-based methods. A filter will, preferably, pass the desired signal undistorted, whereas the noise is reduced. In the design of the filter, an estimate of the noise statistics is often needed. Therefore, this is a widely studied problem in the single-channel case, and several methods for estimating the noise statistics exist [2–6]. In the multichannel case the problem is more difficult due to the cross-correlation between microphones. Some methods are proposed in [7–11]: in [7–10], the cross correlation elements are only updated in periods of unvoiced speech, which can be problematic in the case of non-stationary noise, whereas, in [11], the elements are updated continuously under the assumption that the position of the source is known. However, this is done by steering a null in the direction of the source which means that the filtering has to be done in two steps; a spatial filtering followed by a temporal. Another approach, used in the present work, is to take advantage of the nature of the desired signal. This signal is often voiced speech or musical instruments which is quasi-periodic, and, therefore, the focus in this paper is signals that can be modelled using the harmonic signal model. For speech signals, voiced/unvoiced detectors [12] make it possible to use the approach only on the voiced segments, which are the primary components of a speech signal. Knowing the parameters of the harmonic model, the noise statistics can be estimated by subtracting the desired signal contribution from the statistics of the observed signal. This approach is also taken in the amplitude and phase estimation (APES) filter [13–15]. However, since the APES filter is based on the sample covariance matrix, the number of samples has to be large, a problem which is even more pronounced in the multichannel setup. This can cause problems if the signal is fast varying and, therefore, not stationary over the interval used for estimating the sample covariance matrix.

In the present paper, the multichannel noise covariance matrix is estimated by the iterative adaptive approach (IAA) [16, 17], and the need for a high number of samples is, therefore, not present.The IAA covariance matrix estimate is modified according to the harmonic signal model to get an estimate of the noise covariance matrix and compared to an APES based filter for a harmonic signal.

The rest of the paper is organised as follows: in Section 2, the signal model is set up in the multichannel case. In Section 3, the used filtering method and the sample covariance matrix are introduced, elaborating the motivation for the IAA method. In Section 4, the IAA method for noise covariance matrix estimation is explained. Section 5 shows results, and Section 6 ends the work with a discussion.

2 Signal model

Considering an array of N_s microphones, the observed signal measured by the n_s'th microphone, for time index $n_t = 0, ..., N_t - 1$ and microphone $n_s = 0, ..., N_s - 1$ is: $x_{n_s}(n_t) = s_{n_s}(n_t) + v_{n_s}(n_t)$, where $s_{n_s}(n_t)$ is the desired signal and $v_{n_s}(n_t)$ is the noise. If the desired signal is harmonic, it can be written as a sum of complex sinusoids:

$$s_{n_s}(n_t) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_t n_t} e^{-jl\omega_s n_s}, \tag{E.1}$$

where L is the number of harmonics in the signal, α_l is the complex amplitude of the l'th harmonic, ω_t is the temporal and ω_s is the spatial frequency. If the signal is real it can easily be transformed to its complex counterpart by use of the Hilbert transform [18]. In this paper, we assume anechoic far field conditions and sampling by a uniform linear array (ULA) with an equal spacing, d, between the microphones. Thereby, the relation between the temporal and spatial frequency is $\omega_s = \omega_t f_s c^{-1} d \sin\theta$, for the temporal sampling frequency f_s, the speed of sound in air c, and the direction of arrival (DOA) $\theta \in [-90°; 90°]$.

The processing of the observed signal is done on a subset of M_t observations in time and M_s observations in space defined by the matrix:

$$\mathbf{X}_{n_s}(n_t) = \begin{bmatrix} x_{n_s}(n_t) & \dots & x_{n_s}(n_t - M'_t) \\ \vdots & \ddots & \vdots \\ x_{n_s+M'_s}(n_t) & \dots & x_{n_s+M'_s}(n_t - M'_t) \end{bmatrix}, \tag{E.2}$$

with $M'_t = M_t - 1$ and $M'_s = M_s - 1$. The matrix is then put into vector format using the column-wise stacking operator vec$\{\cdot\}$, i.e., $\mathbf{x}_{n_s}(n_t) = \text{vec}\{\mathbf{X}_{n_s}(n_t)\}$.

3 Filtering

To obtain an estimate of the desired signal, $\tilde{s}(n_t)$, from measurements of the noisy observation, $\mathbf{x}_{n_s}(n_t)$ is filtered by the filter $\mathbf{h}_{\omega_{t,s}}$, optimised for a harmonic signal with temporal fundamental frequency ω_t and spatial frequency ω_s. The spatio-temporal linearly constrained minimum variance (LCMV) filter is a good choice for filtering of periodic signals since the filter gain can be chosen to be one at the harmonic frequencies at the DOA of the observed signal whereas the overall output power of the filter is minimised. The filter is the solution to the minimisation problem [19]

$$\min_{\mathbf{h}} \mathbf{h}^H_{\omega_{t,s}} \mathbf{R} \mathbf{h}_{\omega_{t,s}} \quad \text{s.t.} \quad \mathbf{h}^H_{\omega_{t,s}} \mathbf{a}_{l\omega_{t,s}} = 1 \qquad \text{for} \quad l = 1, ..., L. \tag{E.3}$$

Here, $\{\cdot\}^H$ denotes complex conjugate transpose, $\mathbf{R}$ is the covariance matrix of $\mathbf{x}_{n_s}(n_t)$, i.e., $\mathbf{R} = \text{E}\{\mathbf{x}_{n_s}(n_t)\mathbf{x}^H_{n_s}(n_t)\}$, and

$$\mathbf{a}_{l\omega_{t,s}} = \mathbf{a}_{l\omega_t} \otimes \mathbf{a}_{l\omega_s}, \tag{E.4}$$

$$\mathbf{a}_{\omega} = \begin{bmatrix} 1 & e^{-j\omega} & \dots & e^{-j\omega M'} \end{bmatrix}^T, \tag{E.5}$$

with $\otimes$ denoting the Kronecker product and $\{\cdot\}^T$ the transpose. The solution is given by:

$$\mathbf{h}_{\omega_{t,s}} = \mathbf{R}^{-1}\mathbf{A}_{\omega_{t,s}}(\mathbf{A}^H_{\omega_{t,s}}\mathbf{R}^{-1}\mathbf{A}_{\omega_{t,s}})^{-1}\mathbf{1}, \tag{E.6}$$

where $\mathbf{1}$ is an $L \times 1$ vector containing ones and $\mathbf{A}_{\omega_{t,s}}$ is the spatio-temporal steering matrix

$$\mathbf{A}_{\omega_{t,s}} = \begin{bmatrix} \mathbf{a}_{\omega_{t,s}} & \dots & \mathbf{a}_{L\omega_{t,s}} \end{bmatrix}. \tag{E.7}$$

The covariance matrix is an unknown quantity and is most often replaced by the sample covariance matrix

$$\widehat{\mathbf{R}} = \sum_{p=0}^{N_t - M_t} \sum_{q=0}^{N_s - M_s} \frac{\mathbf{x}_q(n_t - p)\mathbf{x}^H_q(n_t - p)}{(N_t - M'_t)(N_s - M'_s)}. \tag{E.8}$$

If the covariance matrix in (E.3) is replaced by the noise covariance matrix, only the noise power output, and not the overall output power, will be minimised. This will, most often, give better filtering results since perturbations in DOA and fundamental frequency estimates cause a mismatch between the DOA and fundamental frequency of the signal and those used for constraining the LCMV filter, leading to badly regularised filters and signal cancellation. The noise covariance matrix can, for example, be estimated by an amplitude and phase estimation (APES) based approach, as in [20], where a spatio-temporal form of the APES filter [14] is derived. A harmonic signal model is assumed for the desired signal and the part of the sample covariance matrix resembling this signal is then subtracted to give an estimate of the noise covariance matrix. One drawback of both the sample covariance estimate and the APES based covariance estimate is that, in order to make the covariance matrix full rank, the following relation between N_t, N_s, M_t and M_s has to be fulfilled: $(N_t - M_t + 1)(N_s - M_s + 1) \geq M_t M_s$. Normally, there will be a restriction on the number of microphones available, and N_s will, therefore, be fairly small. In order to get a good spatial resolution it is then desirable to choose M_s close or equal to N_s, thereby forcing N_t to be very large compared to M_t. This can be problematic if the signal is not stationary for longer periods of time. Therefore, an alternative method for estimation of the covariance matrix is proposed, where, preferably, $M_t = N_t$ and $M_s = N_s$.

4 IAA covariance matrix estimates

The iterative adaptive approach (IAA) is a method for estimating the spectral amplitudes, $\alpha_{\Omega_{g,k}}$, in the observed signal for temporal and spatial frequency bins:

$$\boldsymbol{\Omega}_G = \big[0 \quad 2\pi\frac{1}{G} \quad \dots \quad 2\pi\frac{G-1}{G}\big], \tag{E.9}$$

$$\boldsymbol{\Omega}_K = \big[0 \quad 2\pi\frac{1}{K} \quad \dots \quad 2\pi\frac{K-1}{K}\big], \tag{E.10}$$

where G and K are the temporal and spatial frequency grid sizes. Element g and k in (E.9) and (E.10) are denoted as Ω_g and Ω_k, respectively, and a combination of frequencies Ω_g and Ω_k is denoted by $\Omega_{g,k}$. The amplitudes are estimated by minimisation of a weighted least squares (WLS) cost function [17, 20]

$$\begin{aligned} J_{\text{WLS}} =& \big[\mathbf{x}_{n_s}(n_t) - \alpha_{\Omega_{g,k}}\mathbf{a}_{\Omega_{g,k}}\big]^H \\ & \mathbf{Q}^{-1}_{\Omega_{g,k}}\big[\mathbf{x}_{n_s}(n_t) - \alpha_{\Omega_{g,k}}\mathbf{a}_{\Omega_{g,k}}\big], \end{aligned} \tag{E.11}$$

Table E.1: IAA for spatio-temporal covariance matrix estimation.

```
initialisation
    $\widetilde{\alpha}_{\Omega_{g,k}} = \frac{\mathbf{a}_{\Omega_{g,k}}^H \mathbf{x}_{n_s}(n_t)}{\mathbf{a}_{\Omega_{g,k}}^H \mathbf{a}_{\Omega_{g,k}}},$
        $g = 0, ...., G-1, \quad k = 0, ...., K-1.$
repeat
    $\widetilde{\mathbf{R}} = \sum_{g=0}^{G-1} \sum_{k=0}^{K-1} |\widetilde{\alpha}_{\Omega_{g,k}}|^2 \mathbf{a}_{\Omega_{g,k}} \mathbf{a}_{\Omega_{g,k}}^H,$
    $\widetilde{\alpha}_{\Omega_{g,k}} = \frac{\mathbf{a}_{\Omega_{g,k}}^H \widetilde{\mathbf{R}}^{-1} \mathbf{x}_{n_s}(n_t)}{\mathbf{a}_{\Omega_{g,k}}^H \widetilde{\mathbf{R}}^{-1} \mathbf{a}_{\Omega_{g,k}}},$
        $g = 0, ...., G-1, \quad k = 0, ...., K-1.$
until (convergence)
```

where $\mathbf{a}_{\Omega_{g,k}}$ is given by (E.4) and (E.5) for $l = 1$, and $\mathbf{Q}_{\Omega_{g,k}}$ is the noise covariance matrix

$$\mathbf{Q}_{\Omega_{g,k}} = \mathbf{R} - |\alpha_{\Omega_{g,k}}|^2 \mathbf{a}_{\Omega_{g,k}} \mathbf{a}_{\Omega_{g,k}}^H. \tag{E.12}$$

The covariance matrix, $\mathbf{R}$, is not known, but is estimated as

$$\widetilde{\mathbf{R}} = \sum_{g=0}^{G-1} \sum_{k=0}^{K-1} |\alpha_{\Omega_{g,k}}|^2 \mathbf{a}_{\Omega_{g,k}} \mathbf{a}_{\Omega_{g,k}}^H. \tag{E.13}$$

The solution to the minimisation of (E.11) is [17, 20]

$$\widetilde{\alpha}_{\Omega_{g,k}} = \frac{\mathbf{a}_{\Omega_{g,k}}^H \mathbf{R}^{-1} \mathbf{x}_{n_s}(n_t)}{\mathbf{a}_{\Omega_{g,k}}^H \mathbf{R}^{-1} \mathbf{a}_{\Omega_{g,k}}}. \tag{E.14}$$

Since the estimate of the spectral amplitudes depends on the estimate of the covariance matrix and vice versa, they are estimated by iterating between (E.13) and (E.14). Typically, 10 to 15 iterations are sufficient for convergence [21]. The process is summarised in Table E.1. With the IAA covariance matrix as a starting point, we find the noise covariance matrix as

$$\mathbf{Q}_{\omega_{t,s}} = \mathbf{R} - \sum_{l=1}^{L} |\alpha_{l\omega_{t,s}}|^2 \mathbf{a}_{l\omega_{t,s}} \mathbf{a}_{l\omega_{t,s}}^H. \tag{E.15}$$

Since the covariance matrix is estimated with a limited number of samples, the desired signal will leak into neighbouring frequency components. Therefore, we estimate the noise covariance matrix by also subtracting the neighbouring grid points to those corresponding to the harmonic frequencies:

$$\widetilde{\mathbf{Q}}_{\omega_{t,s}} = \widetilde{\mathbf{R}} - \sum_{l=1}^{L} \sum_{y=g_l-\delta}^{g_l+\delta} \sum_{z=k_l-\delta}^{k_l+\delta} |\widetilde{\alpha}_{\Omega_{y,z}}|^2 \mathbf{a}_{\Omega_{y,z}} \mathbf{a}_{\Omega_{y,z}}^H,$$

where g_l and k_l are the grid indices corresponding to the l'th harmonic and 2δ is the number of subtracted neighbouring frequency grid points.

5 Results

The IAA noise covariance estimates are tested by use of a synthetic harmonic signal with $\omega_t = 0.5027$ (corresponding to 200 Hz), $f_s = 2500$ Hz, $L = 5$, $\theta = 10^\circ$ and $\alpha_l = 1\,\forall\, l$. The speed of sound is set to $c = 343.2$ m/s and $d = c/f_s$. The individual microphone signals are artificially delayed according to d and θ. Noise is added to give a desired average input signal-to-noise ratio (SNR). The noise is white Gaussian noise passed through a 10'th order auto-regressive filter made using a harmonic signal with seven harmonics and a fundamental frequency of 137 Hz. For the IAA estimate $N_t = M_t = 20$, $N_s = M_s = 10$. To decrease computational complexity, the grid is modified to make a uniform grid containing the harmonic frequencies, and, thereby, the number of grid points can be decreased, here, $G = 400$ and $K = 71$, and the number of iterations is 10. Alternatively, if the harmonics are not placed on the grid, the relaxation in [22] can be utilised. When the covariance matrices of consecutive samples are estimated, the first estimate is initialised as in Table E.1, the rest are initialised with the former estimate of the covariance matrix, and only one iteration is made [21]. The number of subtracted neighbouring frequency grid points is set to eight since this was observed to give the highest SNR.

The performance after filtering is measured by means of the output SNR, $\text{oSNR}(\mathbf{h}) = \frac{\sigma^2_{s,\text{nr}}}{\sigma^2_{v,\text{nr}}}$, with $\sigma^2_{s,\text{nr}}$ and $\sigma^2_{v,\text{nr}}$ being the variances of signal and noise after noise reduction. The variances are computed over 50 consecutive samples and the resulting output SNR is averaged over 100 runs.

The IAA noise covariance estimate, $\widetilde{\mathbf{Q}}_{\omega_{t,s}}$ ($\text{IAA}_{\widetilde{\mathbf{Q}}_{\omega_t,\omega_s}}$) is compared to the IAA covariance estimate $\widetilde{\mathbf{R}}$ ($\text{IAA}_{\widetilde{\mathbf{R}}}$), the IAA noise covariance estimate based on the clean noise signal ($\text{IAA}_{\mathbf{Q}}$) and to the APES based estimate with two different configurations. In the first (APES_1), the number of samples is the same as for the IAA filter whereas the filter length is shorter, $N_t = 20$, $M_t = 10$, $N_s = 10$ and $M_s = 5$. In the second (APES_2), the filter length is the same as in the IAA, but longer data segments are used, $N_t = 224$, $M_t = 20$, $N_s = 10$ and $M_s = 10$. The methods are compared by using the covariance matrix estimates in the LCMV filter. Examples of filter responses are shown in Fig. G.2 for an average input SNR of 10 dB. Comparing (a) to (b), it is seen that taking account for the desired signal in the generation of the filter gives a much more well conditioned filter. Comparing to (c), (a) has more attenuation at other DOAs and frequencies than the ones of the desired signal, whereas it is difficult to say whether the filter in (a) or (d) will have the best performance.

The output SNR are shown as a function of the input SNR in Fig. E.2.

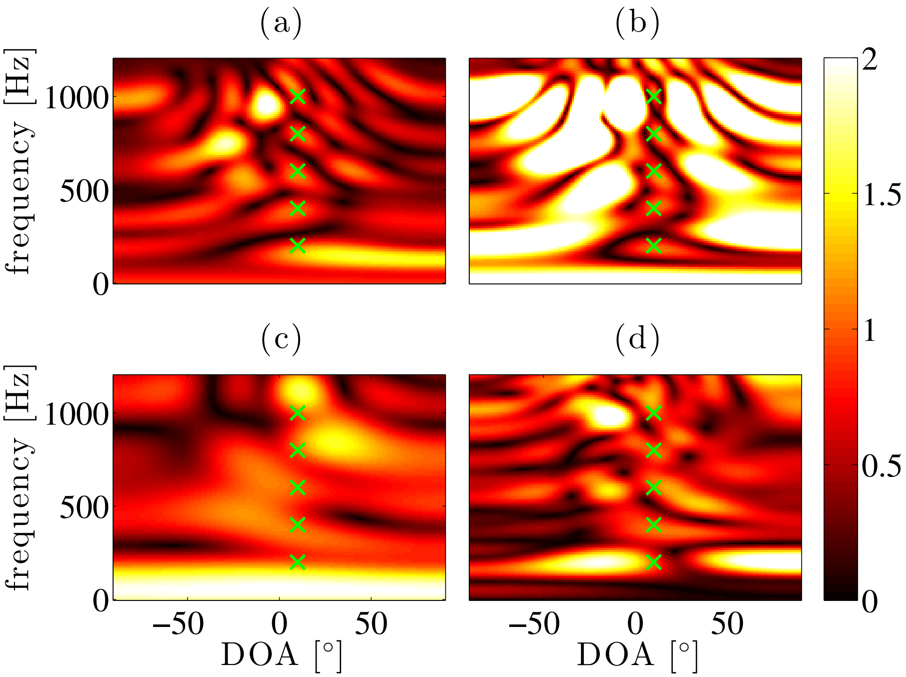


Fig. E.1: Filter responses for (a) IAA based on noise covariance matrix estimate $\widetilde{\mathbf{Q}}_{\omega_{t,s}}$ (b) IAA based on covariance matrix estimate, $\widetilde{\mathbf{R}}$ (c) APES based estimate with $N_t = 20$, $M_t = 10$, $N_s = 10$, and $M_s = 5$ (d) APES based estimate with $N_t = 224$, $M_t = 20$, $N_s = 10$, and $M_s = 10$. Harmonics of desired signal are marked by green crosses. The average input SNR is 10 dB.

For input SNRs from 0 to 10 dB, a gain in SNR of approximately 8 dB can be obtained compared to APES$_1$. At higher input SNRs, the gain decreases. If more samples are available, APES$_2$ outperforms IAA, but then the noise covariance matrix has been estimated on the basis of 4480 samples of the signal compared to only 200 with the IAA method.

The IAA method is tested on a piece of a speech signal sampled at 8 kHz. The fundamental frequency is estimated from the desired signal with an approximate nonlinear least squares estimator [23], and the model order is set to 18. Due to the high model order, here $N_t = M_t = 50$. The DOA, N_s, M_s, c and d are the same as before. Based on the fundamental frequency estimate, we design the grid at each time instance such that the harmonics lie on the grid, which means that the grid size varies slightly over time, with approximate values of $G = 400$ and $K = 100$. The ten microphone recordings are made using the room impulse response generator [24] under anechoic conditions with a distance of 5 m between source and microphone array. Babble noise from the AURORA database [25] is added to the microphone signals to give an average

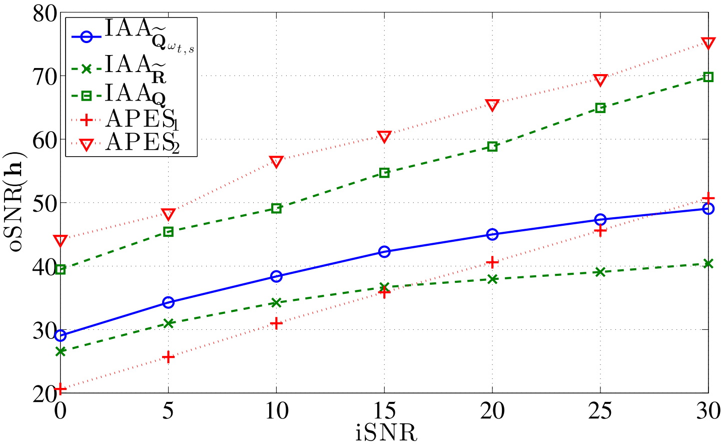


Fig. E.2: Output SNR as a function of the input SNR.

input SNR of 10 dB.

A short segment of the noisy, desired and estimated signal using, respectively, the proposed IAA noise covariance matrix estimate, $\widetilde{\mathbf{Q}}_{\omega_{t,s}}$, and the IAA covariance matrix estimate, $\widetilde{\mathbf{R}}$, are plotted in Fig. E.3. It is seen in the figure that $\text{IAA}_{\widetilde{\mathbf{Q}}_{\omega_{t,s}}}$ gives a good estimate of the desired signal and follows the desired signal more closely than the $\text{IAA}_{\widetilde{\mathbf{R}}}$ estimate.

6 Discussion

In the present paper, we suggest a method for estimation of the noise covariance matrix based on the iterative adaptive approach (IAA). The method only needs a single snapshot of data to estimate the covariance matrix. This makes it advantageous when fast varying signals are considered. In speech enhancement, IAA has formerly been used for fundamental frequency estimation [20] and joint direction of arrival (DOA) and fundamental frequency estimation [22], both assumed known in the present paper. Here, the covariance matrix estimate from the IAA is modified, under the assumption of a harmonic desired signal, to give an estimate of the noise covariance matrix. This estimate is then used in the linearly constrained minimum variance (LCMV) filter and compared to a spatio-temporal APES based filter proposed in [15]. The proposed method shows better performance in terms of signal-to-noise ratio (SNR) when the number of samples is limited, whereas the APES based filter has a better performance when the number of samples is not an issue. Compared to [11], where the filtering has to be done in two steps, the work presented here does the spatial and temporal filtering jointly.

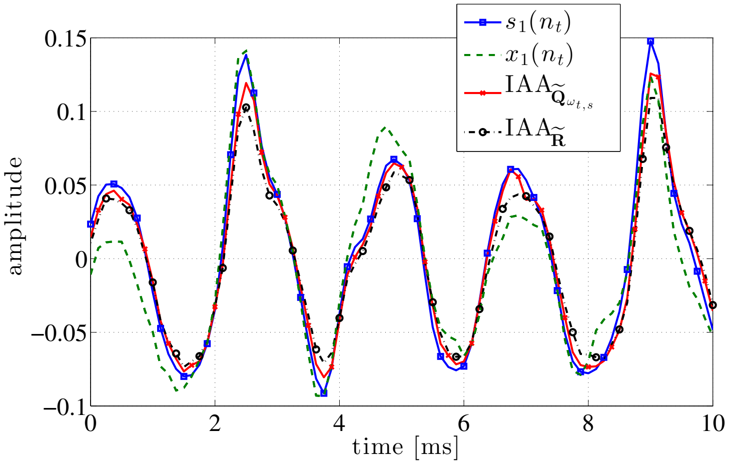


Fig. E.3: Reconstructed signal using $\text{IAA}_{\widetilde{\mathbf{Q}}_{\omega_{t,s}}}$ compared to reconstruction using $\text{IAA}_{\widetilde{\mathbf{R}}}$ and the desired and noisy signal from the first microphone.

References

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

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

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

[4] L. Lin, W. H. Holmes, and E. Ambikairajah, “Subband noise estimation for speech enhancement using a perceptual Wiener filter,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 2003, pp. 80–83.

[5] R. C. Hendriks, R. Heusdens, and J. Jensen, “MMSE based noise psd tracking with low complexity,” in *IEEE Trans. Acoust., Speech, Signal Process.*, 2010, pp. 4266–4269.

[6] D. Ealey, H. Kelleher, and D. Pearce, “Harmonic tunnelling: tracking non-stationary noises during speech,” in *Proc. Eurospeech*, Sep. 2001, pp. 437–440.

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

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

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

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

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

[12] K. I. Molla, K. Hirose, N. Minematsu, and K. Hasan, “Voiced/unvoiced detection of speech signals using empirical mode decomposition model,” in *Int. Conf. Information and Communication Technology*, Mar. 2007, pp. 311–314.

[13] J. Li and P. Stoica, “An adaptive filtering approach to spectral estimation and SAR imaging,” *IEEE Trans. Signal Process.*, vol. 44, no. 6, pp. 1469–1484, Jun. 1996.

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

[15] J. R. Jensen, M. G. Christensen, and S. H. Jensen, “An optimal spatio-temporal filter for extraction and enhancement of multi-channel periodic signals,” in *Rec. Asilomar Conf. Signals, Systems, and Computers*, Nov. 2010, pp. 1846–1850.

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

[17] W. Roberts, P. Stoica, J. Li, T. Yardibi, and F. A. Sadjadi, “Iterative adaptive approaches to mimo radar imaging,” *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 1, pp. 5–20, Feb. 2010.

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

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

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

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

[22] Z. Zhou, M. G. Christensen, J. R. Jensen, and H. C. So, “Joint DOA and fundamental frequency estimation based on relaxed iterative adaptive approach and optimal filtering,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013.

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

[24] E. A. P. Habets, “Room impulse response generator,” Technische Universiteit Eindhoven, Tech. Rep., 2010, ver. 2.0.20100920.

[25] D. Pearce and H. G. Hirsch, “The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *Proc. Int. Conf. Spoken Language Process.*, Oct 2000.

Paper F

Least Squares Estimate of the Initial Phases in STFT based Speech Enhancements

Sidsel Marie Nørholm, Martin Krawczyk-Becker, Timo Gerkmann, Steven van de Par, Jesper Rindom Jensen and Mads Græsbøll Christensen

The paper has been accepted for publication in the
Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2015.

Abstract

In this paper, we consider single-channel speech enhancement in the short time Fourier transform (STFT) domain. We suggest to improve an STFT phase estimate by estimating the initial phases. The method is based on the harmonic model and a model for the phase evolution over time. The initial phases are estimated by setting up a least squares problem between the noisy phase and the model for phase evolution. Simulations on synthetic and speech signals show a decreased error on the phase when an estimate of the initial phase is included compared to using the noisy phase as an initialisation. The error on the phase is decreased at input SNRs from -10 to 10 dB. Reconstructing the signal using the clean amplitude, the mean squared error is decreased and the PESQ score is increased.

Index Terms: speech enhancement, single-channel, STFT domain, phase estimation, signal reconstruction.

1 Introduction

Single-channel speech enhancement is important in many systems such as mobile phones and hearing aids where it is desirable to estimate a speech signal from a mixture of the signal buried in noise. Some enhancement methods work directly in the time domain [1, 2] whereas other methods work by transforming the signal into another domain. This could for example be the subspace methods where, e.g., the eigenvalue decomposition of a signal matrix is computed [3]. Another domain, that we will focus on in this paper because it is computational effective [4], is the short time Fourier transform (STFT) domain. Here, some well-known methods are spectral subtraction [5] and the Short-Time Spectral Amplitude Estimator [6]. Common for these methods, and most other methods in this domain, is that they enhance the STFT amplitude, whereas the phase is left unaltered. This is motivated by [7, 8] who conclude that modifying the noisy STFT phase only gives a minor gain compared to modifying the noisy STFT amplitude. However, later work by [9] shows that the importance of the phase depends on the settings and that it can be beneficial to estimate the STFT phase. Recently, in [10, 11], improved STFT amplitude estimates are obtained by using STFT phase estimates in the process.

Different approaches have been taken to modify the noisy STFT phase. In [12, 13], the change of STFT phase is based on the fact that not all STFT representations are consistent. Given a spectrum of a speech signal, an inverse STFT followed by an STFT leads back to the same spectrum, but if changes are made to the amplitude or phase of the spectrum, this is not necessarily the case for the altered spectrum, and it is, therefore, not consistent [14]. The quality of the resulting signal can be improved by minimising this inconsistency. In [12, 13] this is done by modifying the STFT phase to make a better match to the STFT amplitude estimate. The error on the phase is, therefore, not guaranteed to decrease because the phase is only modified to match the enhanced STFT amplitude. In [15], the STFT phase change in voiced speech periods is estimated based on the harmonic model and knowledge about the fundamental frequency. The phase in unvoiced periods is left unaltered, but since the major constituent of speech is voiced, changing the phase in these periods can still make a difference in terms of speech enhancement. Since only the phase change is estimated in [15], an initial phase estimate is needed as an anchor. In [15], the noisy phase is used as the initial STFT phase at the harmonic frequencies which gives a constant offset at each harmonic between the clean speech phase and the estimated phase and changes the relation between harmonics. This results in a significant error on the enhanced STFT phase, and the waveform of the resulting signal will be changed. In terms of perception, this is not a major problem if only a single harmonic is present, but in the case of more harmonics, as is the case in speech signals, it can have an influence on how the sound is perceived [16, 17].

To minimise the error on the phase, we propose a method to estimate the initial STFT phases in voiced speech periods. The method is based on the harmonic model and the model for phase evolution over time presented in [15]. The initial phases are estimated by setting up a least squares (LS) problem between the noisy phase and the signal model.

The paper is organised as follows: in Section 2 the harmonic signal model and the STFT are shortly introduced, in Section 3 the method from [15] is introduced, in Section 4 the proposed method is explained, results are presented in Section 5, and Section 6 concludes the work.

2 Signal Model

We here use the harmonic signal model which is a good approximation to voiced speech. With this model the signal is composed of a set of harmonics with sinusoids having frequencies given by multiples of a fundamental frequency. For discrete time indices, $m = 0, ..., M-1$, the signal can be represented as:

$$s(m) = \sum_{h=1}^{H} 2A_h \cos(\omega_0 hm + \varphi_h), \tag{F.1}$$

where H is the number of harmonics, A_h the amplitude of the h'th harmonic, $\omega_0 = 2\pi f_0/f_s$ the normalised fundamental angular frequency, with f_0 being the fundamental frequency and f_s the sampling frequency, and φ_h is the initial phase of the h'th harmonic. The desired signal is estimated from a mixture,

$x(m)$, of the desired signal, $s(m)$, and additive noise, $v(m)$,

$$x(m) = s(m) + v(m). \tag{F.2}$$

The processing is done in the short-time Fourier transform (STFT) domain. The transformation to this domain is done by splitting the noisy signal into segments of length N, overlapping by $N - L$ samples, applying a window function $w(n)$ and computing the Discrete Fourier Transform (DFT), i.e.,

$$\begin{aligned} X(k,l) &= \sum_{n=0}^{N-1} x(lL+n)w(n)e^{-j\omega_k n} && \text{(F.3)} \\ &= |X(k,l)|e^{j\phi_X(k,l)}, && \text{(F.4)} \\ &= S(k,l) + V(k,l), && \text{(F.5)} \\ &= |S(k,l)|e^{j\phi_S(k,l)} + |V(k,l)|e^{j\phi_V(k,l)}, && \text{(F.6)} \end{aligned}$$

with k being the frequency index, l the segment index and $\omega_k = 2\pi k/N$ the normalised angular frequency of frequency band k. It can be seen in (F.4) that the signal in the STFT domain can be split into an amplitude part $|X(k,l)|$ and a phase part $e^{j\phi_X(k,l)}$. In many existing approaches only the amplitude is modified whereas the phase is not estimated, and the noisy phase is used directly, i.e., $\widehat{S(k,l)} = |\widehat{S(k,l)}|e^{j\phi_X(k,l)}$, where $\widehat{\{\cdot\}}$ denotes an estimated quantity. In this paper we will focus on estimating the clean phase $\phi_S(k,l)$ from the noisy phase $\phi_X(k,l)$.

3 Phase Reconstruction

In [15], the change in instantaneous phase in frequency bins containing the harmonic frequencies is estimated as a piecewise linear function when the harmonic frequency $\omega_h^{k,l} = h\omega_0^{k,l}$ is known, i.e.,

$$\begin{aligned} \Delta\phi_S(k,l) &= \phi_S(k,l) - \phi_S(k,l-1) \\ &= \omega_h^{k,l} L. && \text{(F.7)} \end{aligned}$$

The last equality holds under the assumption that the fundamental frequency in segments $l - 1$ and l are the same. Reformulation of (F.7) gives the instantaneous phase in segment l from the phase in segment $l - 1$

$$\widehat{\phi}_S(k,l) = \widehat{\phi}_S(k,l-1) + \omega_h^{k,l} L. \tag{F.8}$$

To get the instantaneous phase in segment l, it is therefore necessary to have information about the instantaneous phase in segment $l - 1$. In the very beginning of a piece of voiced speech, the algorithm has to be initialised with a

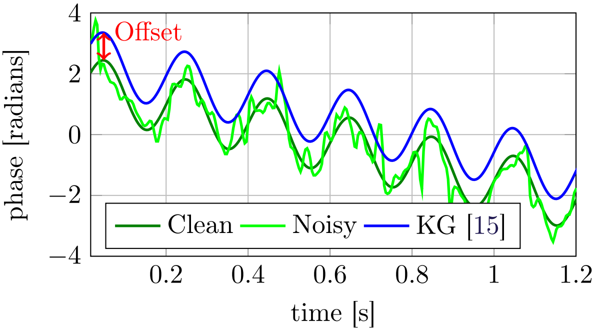


Fig. F.1: Reconstruction of the STFT phase based on KG [15] where the noisy phase is used as initialisation leading to a constant offset between the clean phase and the reconstructed phase.

phase for the first segment, i.e., information about the initial phases, φ_h, is needed. In [15], the noisy phase is used as an initialisation. This is illustrated in Fig. F.1 where the baseband transformed phase (see [15]) in a frequency band containing a single harmonic of a frequency modulated signal is shown. It is seen that even though the phase evolution over time is correctly estimated with the method in [15] (KG [15]), using the noisy phase as an initialisation will give a constant offset between the clean phase and the estimated phase due to a wrong initial phase, ϕ_h. If only a single sinusoid is present, the initial phase is not that important in terms of perception, but if several harmonics are present, the relationship between the initial phases of the different harmonics has an influence on the shape of the waveform of the resulting signal and can also have an influence on how the sound is perceived [16, 17]. Therefore, we estimate the initial phases in the next section.

4 Estimation of Initial Phases

The estimation of the initial phases is set up as a least squares (LS) problem between the instantaneous phases estimated using (F.8) with an initialisation of $\varphi_h = 0$ and the noisy phase for each harmonic separately

$$\widehat{\varphi}_h = \arg\min_{\varphi_h} \sum_{l=l_0}^{l_0+P-1} (\phi_X(k,l) - \widehat{\phi}_S(k,l) - \varphi_h)^2, \tag{F.9}$$

where P is the number of segments used for the estimation. The solution is found by differentiating the expression and equating with zero, i.e.,

$$\widehat{\varphi}_h = \frac{1}{P} \sum_{l=l_0}^{l_0+P-1} \phi_X(k,l) - \widehat{\phi}_S(k,l). \tag{F.10}$$

Due to the properties of the phase seen in (F.4), every $b2\pi, b \in \mathbb{Z}$, multiple of the phase gives rise to the same phase contribution to the resulting signal. This has to be taken into account in the estimation of the initial phase and, therefore, every phase difference in (F.10) is mapped to the interval $[-\pi, \pi]$, and the final estimate of the initial phase of harmonic h is given by:

$$\widehat{\varphi}_h = \frac{1}{P} \sum_{l=l_0}^{l_0+P-1} \angle(e^{j\phi_X(k,l) - j\widehat{\phi}_S(k,l)}), \tag{F.11}$$

where $\angle(\cdot)$ denotes the angle of the argument. To keep the right relation between frequency bins, all bins dominated by the given harmonic (see [15]) are also shifted according to the given estimate.

The method is implemented in two different ways. One where an entire piece of voiced speech is used for the estimation of the initial phase (denoted LS1 in the results section) and one where the initial phase of a given harmonic is reestimated each time the harmonic jumps to a new frequency bin (denoted LS2 in the results section). The first method has the advantage of more data used in the estimation and, therefore, if the model is perfectly correct, it should give a better estimate. However, it is vulnerable to errors in the model, e.g., a slightly wrong fundamental frequency estimation would lead to a model that over time deviates more and more from the clean signal and, thereby, gives larger errors in the estimation of the initial phase with more time segments used. The second method should do a better job in the case of a erroneous fundamental frequency estimate. However, in the transformation to the STFT domain the signal is overlapped which means that the noise in neighbouring time frames is not uncorrelated and, therefore, an estimation based on only a few frames would give an unreliable estimate.

The estimate of the initial phases introduces a latency in the system according to P. LS1 introduces a delay of one voiced speech period. The latency introduced by LS2 will depend on when the harmonics jump from one frequency bin to another and will, therefore, be smaller or equal to the latency introduced by LS1.

5 Results

The least squares estimates of the initial phases are first tested by means of a synthetic signal. After testing the concept on synthetic data, we turn to real speech signals. The synthetic signal used is a frequency modulated harmonic signal, i.e.,

$$s(m) = \sum_{h=1}^{H} A_h \cos(\omega_0 hm + \frac{\omega_\Delta}{\omega_m} h \cos(\omega_m m) + \varphi_h).$$

Here, $\omega_\Delta = 2\pi f_\Delta / f_s$ is the maximum deviation of the first harmonic away from ω_0 in one direction and $\omega_m = 2\pi f_m / f_s$ is the normalised angular modulation frequency. The signal is chosen because of its harmonic structure which is the basis of the proposed method and, further, it is a more interesting case than a pure harmonic signal since the fundamental frequency is modulated and, therefore, the harmonics will jump between different frequency bins when it is transformed to the STFT domain. Due to the multiplication by h in the modulation, the maximum deviation away from the harmonic frequency is increasing for higher harmonics, and they will, therefore, also have a higher tendency to jump between frequency bins. This will also be the case for speech signals. In the simulations $H = 10$, $f_s = 8000$, $M = 20000$, and f_0, f_Δ, f_m and φ_h are chosen randomly in intervals as $f_0 \in [100, 200]$ Hz, $f_\Delta \in [0, 10]$ Hz, $f_m \in [0, 10]$ Hz and $\varphi_h \in [-\pi, \pi]$. The frequency modulated signal is degraded by white Gaussian noise at signal-to-noise ratios (SNRs) from -10 dB to 10 dB in steps of 2.5 dB. The signal is transformed to the STFT domain in segments of 256 samples (corresponding to 32 ms) with an overlap of 87.5% and the window applied is a square root Hann window. In the evolution of the phase in the frequency domain, the true fundamental frequency is assumed to be known. The results are averaged over 1000 Monte Carlo simulations (MCS) [18]. The methods are evaluated both in the frequency and in the time domain. In the frequency domain, the phase ambiguities are again taken into account by using the circular phase error [6]:

$$\varepsilon(k, l) = 1 - \cos(\phi^S_{k,l} - \widehat{\phi}^S_{k,l}), \tag{F.12}$$

which is in the range [0,2]. In the time domain they are evaluated by means of the mean squared error (MSE) between the clean signal, $s(m)$, and the reconstructed signal, $\widehat{s}(m)$, $\text{MSE} = (s(m) - \widehat{s}(m))^2$. The two methods are compared to the method in [15] where the noisy phase is used as an initialisation, here denoted by KG [15], and the noisy phase denoted by Noisy. In Fig. F.2, the phase error averaged over all frequency bins and time is shown. It is seen that at all input SNRs considered here there is an advantage in estimating the instantaneous phase compared to using the noisy phase. Also, a smaller error can be obtained by estimating the initial phase. Both LS estimates give smaller errors than KG [15] up to approximately 0 dB input SNR, above 0 dB, LS2 gives a smaller error than KG [15] whereas LS1 gives the same error as KG [15]. The signal is thereafter reconstructed using an inverse STFT. Before doing that, the STFT phase term has to be multiplied with the STFT amplitude. For calculation of the mean squared error, we have used the clean speech amplitude, and the result is shown in Fig. F.3. Now, KG [15] gives the highest error at all input SNRs, LS2 gives the lowest error whereas LS1 and the noisy phase give errors in between. The lower error of LS2 compared to LS1 shows that it is reasonable to take the jumps between frequency bins into account in the estimation process.

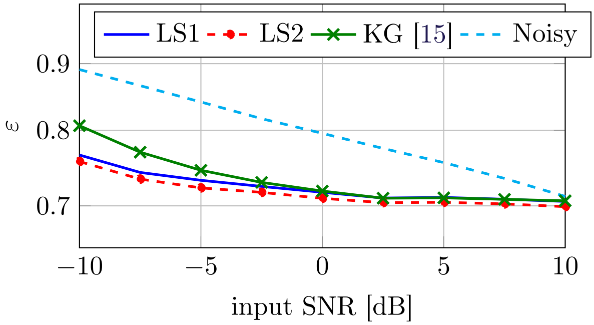


Fig. F.2: Phase error, ε, as a function of the input SNR averaged over all frequency bins and time for a synthetic signal. Averaged over 1000 MCS.

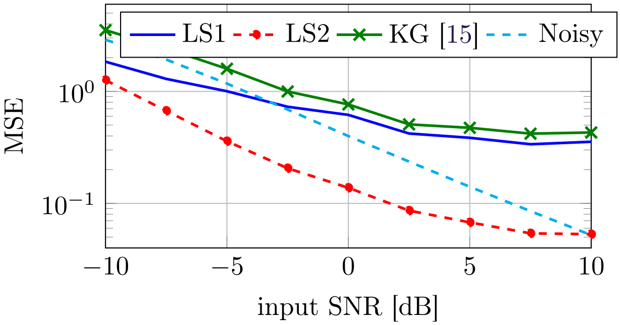


Fig. F.3: Mean squared error of reconstructed signal as a function of the input SNR for a synthetic signal. Combination of phase and clean amplitude. Average over 1000 MCS.

The methods are also evaluated using five male and five female speech signals from the TIMIT database degraded by white Gaussian noise. The signals are downsampled to 8 kHz and the fundamental frequency is estimated from the clean speech signal using a nonlinear least squares estimator [19]. In the estimation, a search interval around (±10 Hz) the pitch obtained from the corresponding laryngograph signal [17] is used. The voiced periods are also chosen using the laryngograph track as the periods where the fundamental frequency is larger than zero. It is found that best results are obtained if only the lowest harmonics are modified so here the initial phases for the three first harmonics are estimated and changed. As in [15], the noisy phase is used directly in periods of unvoiced speech. The phase error is shown in Fig. F.4, this time averaged over 50 MCS for each speaker, voiced speech periods and all frequency bins. The error on the phase using LS1 or LS2 is considerably decreased compared to KG [15], and LS2 again performs slightly better than LS1. Here, however, the error on the noisy phase is very similar to the error of LS1 and LS2, being slightly higher below 5 dB input SNR and slightly lower

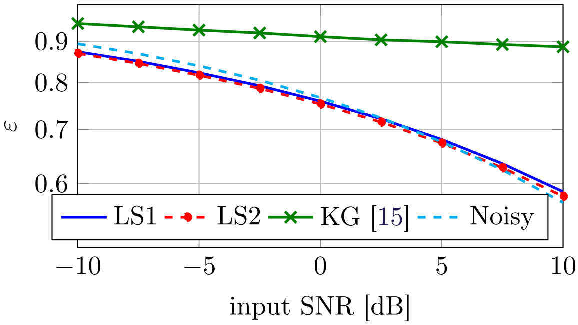


Fig. F.4: Phase error, ε, as a function of the input SNR averaged over all frequency bins and voiced speech periods for 5 male and 5 female speakers from the TIMIT database. Average over 50 MCS for each speaker.

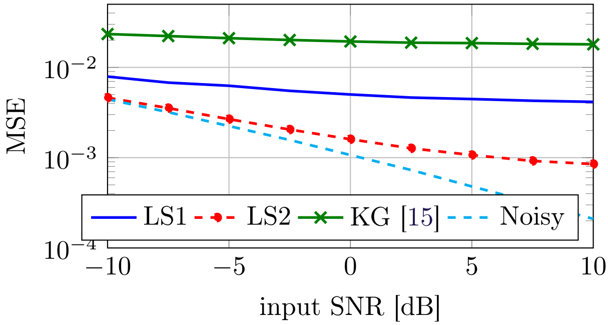
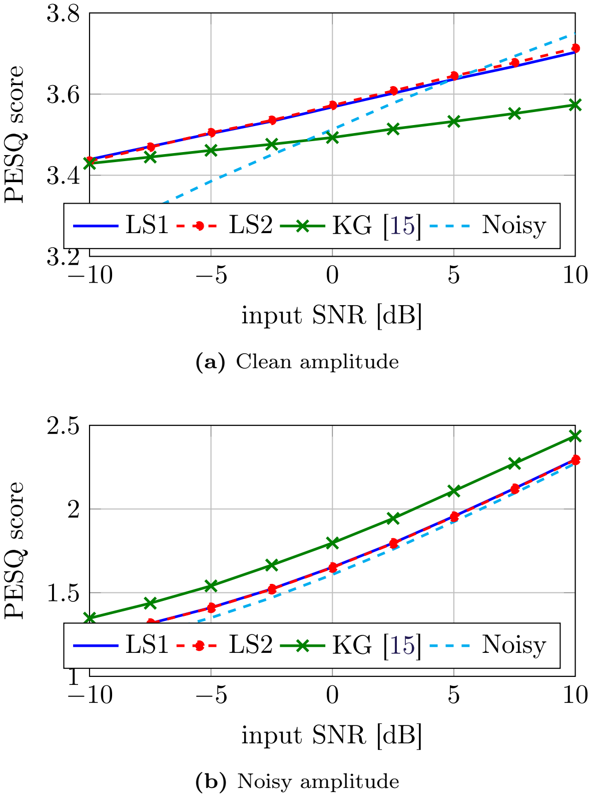


Fig. F.5: Mean squared error of reconstructed voiced speech parts as a function of the input SNR for 5 male and 5 female speakers from the TIMIT database. Combination of phase and clean amplitude. Average over 50 MCS for each speaker.

above 5 dB input SNR. Looking at the mean squared error of the reconstructed voiced speech parts in Fig. F.5, it is seen that the error is again decreased when estimating the initial phase with LS1 or LS2 compared to using the noisy initial phase in KG [15], and again it is also more beneficial to use LS2 than LS1. Here, on the other hand, using the noisy phase at all times gives a slightly lower error on the reconstructed signal than LS2. The Perceptual Evaluation of Speech Quality (PESQ) score [20] of the reconstructed speech signals is also found. We have used two different choices of amplitudes in the reconstruction. These are the clean amplitude and the noisy amplitude. Using the clean amplitude, LS1 and LS2 performs best over the most of the range of input SNRs as seen in Fig. F.6a whereas Fig. F.6b shows that using the noisy amplitude, KG [15] gives the best PESQ score over the entire range. It would be more intuitive if a smaller error on the phase always would lead to a better reconstructed signal. The reason for this might be due to the



(a) Clean amplitude

(b) Noisy amplitude

Fig. F.6: PESQ score of reconstructed signal as a function of the input SNR for 5 male and 5 female speakers from the TIMIT database. Combination of phase and (a) clean amplitude and (b) noisy amplitude. Average over 50 MCS for each speaker.

inconsistency discussed in [14] and suggest that more work should be put into making consistent STFT representations based on both an amplitude and a phase estimate. However, better phase estimates on its own can still be used in, e.g., [10, 11] to give better reconstructed signals.

6 Conclusion

In this paper, we considered speech enhancement in the STFT domain. Most prior work has been done on enhancing the noisy STFT amplitude, but the focus of this paper was the STFT phase. We suggest a least squares method to estimate the initial STFT phases in voiced speech periods. The initial phases are found by minimising the squared error between the noisy phase and the model-based phase estimates suggested in [15]. Simulations show that the error on the phase can be decreased considerably when estimating the initial phase as compared to using the noisy phase as the initial phase as proposed in [15]. The error on the phase is also reduced compared to the noisy phase in the ideal case with a synthetic signal and also slightly up to an input SNR of 5 dB when speech signals are considered. Reconstruction in combination with the clean amplitude gives an increase in PESQ score relative to KG [15] and an increase relative to the noisy phase up to an input SNR of 5 dB.

References

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

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

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

[4] J. Benesty, J. Chen, and E. A. Habets, *Speech enhancement in the STFT domain.* Springer, 2012.

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

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[7] D. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 4, pp. 679–681, 1982.

[8] P. Vary, "Noise suppression by spectral magnitude estimation - mechanism and theoretical limits," *Signal Processing*, vol. 8, no. 4, pp. 387 – 400, 1985.

[9] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011.

[10] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Process. Lett.*, vol. 20, no. 2, pp. 129–132, 2013.

[11] T. Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4199–4208, Aug. 2014.

[12] J. Le Roux, E. Vincent, Y. Mizuno, H. Kameoka, N. Ono, and S. Sagayama, "Consistent wiener filtering: Generalized time-frequency masking respecting spectrogram consistency," in *Latent Variable Analysis and Signal Separation.* Springer Berlin Heidelberg, 2010, pp. 89–96.

[13] D. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr 1984.

[14] N. Sturmel and L. Daudet, "Signal reconstruction from STFT magnitude: a state of the art," *International Conference on Digital Audio Effects (DAFx)*, pp. 375–386, 2011.

[15] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1931–1940, 2014.

[16] B. C. J. Moore, *An introduction to the psychology of hearing.* Brill, 2012.

[17] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models.* Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[18] N. Metropolis, "The beginning of the monte carlo method," *Los Alamos Science*, no. 15, pp. 125–130, 1987.

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

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

Paper G

Single-Channel Noise Reduction using Unified Joint Diagonalization and Optimal Filtering

Sidsel Marie Nørholm, Jacob Benesty, Jesper Rindom Jensen and
Mads Græsbøll Christensen

The paper has been published in the
EURASIP Journal on Advances in Signal Processing, 2014.

Abstract

In this paper, the important problem of single-channel noise reduction is treated from a new perspective. The problem is posed as a filtering problem based on joint diagonalization of the covariance matrices of the desired and noise signals. More specifically, the eigenvectors from the joint diagonalization corresponding to the least significant eigenvalues are used to form a filter, that effectively estimates the noise when applied to the observed signal. This estimate is then subtracted from the observed signal to form an estimate of the desired signal, i.e., the speech signal. In doing this, we consider two cases, where, respectively, no distortion and distortion is incurred on the desired signal. The former can be achieved when the covariance matrix of the desired signal is rank deficient, which is the case, for example, for voiced speech. In the latter case, the covariance matrix of the desired signal is full rank, as is the case, for example, in unvoiced speech. Here, the amount of distortion incurred is controlled via a simple, integer parameter, and the more distortion allowed, the higher the output signal-to-noise ratio (SNR). Simulations demonstrate the properties of the two solutions. In the distortionless case, the proposed filter achieves only a slightly worse output SNR, compared to the Wiener filter, along with no signal distortion. Moreover, when distortion is allowed, it is possible to achieve higher output SNRs compared to the Wiener filter. Alternatively, when a lower output SNR is accepted, a filter with less signal distortion than the Wiener filter can be constructed.

Keywords: noise reduction, speech enhancement, single-channel, time domain filtering, joint diagonalization.

1 Introduction

Speech signals corrupted by additive noise suffer from a lower perceived quality and lower intelligibility than their clean counterparts and cause listeners to suffer from fatigue after extended exposure. Moreover, speech processing systems are frequently designed under the assumption that only a single, clean speech signal is present at the time. For these reasons, noise reduction plays an important role in many communication and speech processing systems and continues to be an active research topic today. Over the years, many different methods for noise reduction have been introduced, including optimal filtering methods [1], spectral subtractive methods [2], statistical methods [3–5], and subspace methods [6, 7]. For an overview of methods for noise reduction, we refer the interested reader to [1, 8, 9] and to [10] for a recent and complete overview of applications of subspace methods to noise reduction.

In the past decade or so, most efforts in relation to noise reduction seem to have been devoted to tracking of noise power spectral densities [11–14] to allow for better noise reduction during speech activity, extensions of noise reduction methods to multiple channels [15–18], and improved optimal filtering techniques for noise reduction [1, 8, 19–21]. However, little progress has been made on subspace methods.

In this paper, we explore the noise reduction problem from a different perspective in the context of single-channel noise reduction in the time domain. This perspective is different from traditional approaches in several respects. Firstly, it combines the ideas behind subspace methods and optimal filtering via joint diagonalization of the desired and noise signal covariance matrices. Since joint diagonalization is used, the method will work for all kinds of noise, as opposed to, e.g., when an eigenvalue decomposition is used where preprocessing has to be performed when the noise is not white. Secondly, the perspective is based on obtaining estimates of the noise signal by filtering of the observed signal and, thereafter, subtracting the estimate of the noise from the observed signal. This is opposite to a normal filtering approach where the observed signal is filtered to get the estimated signal straight away. The idea of first estimating the noise is known from the generalized sidelobe canceller technique in a multichannel scenario [22]. Thirdly, when the covariance matrix of the desired signal has a rank that is lower than that of the observed signal, the perspective leads to filters that can be formed such that no distortion is incurred on the desired signal, and distortion can be introduced so that more noise reduction is achieved. The amount of distortion introduced can be controlled via a simple, integer parameter.

The rest of the paper is organized as follows. In Section 2, the basic signal model and the joint diagonalization perspective is introduced, and the problem of interest is stated. We then proceed, in Section 3, to introduce the noise reduction approach for the case where no distortion is incurred on the desired signal. This applies in cases where the rank of the observed signal covariance matrix exceeds that of the desired signal covariance matrix. In Section 4, we then relax the requirement of no distortion on the desired signal to obtain filters that can be applied more generally, i.e., when the ranks of the observed and desired signals are the same. Simulation results demonstrating the properties of the obtained noise reduction filters are presented in Section 5, whereafter we conclude on the work in Section 6.

2 Signal Model and Problem Formulation

The speech enhancement (or noise reduction) problem considered in this work is the one of recovering the desired (speech) signal $x(k)$, k being the discrete-time index, from the noisy observation (sensor signal) [1, 8, 9]:

$$y(k) = x(k) + v(k), \tag{G.1}$$

where $v(k)$ is the unwanted additive noise which is assumed to be uncorrelated with $x(k)$. All signals are considered to be real, zero mean, broadband, and stationary.

The signal model given in (G.1) can be put into a vector form by considering the L most recent successive time samples of the noisy signal, i.e.,

$$\mathbf{y}(k) = \mathbf{x}(k) + \mathbf{v}(k), \tag{G.2}$$

where

$$\mathbf{y}(k) = \left[\begin{array}{cccc} y(k) & y(k-1) & \cdots & y(k-L+1) \end{array} \right]^T \tag{G.3}$$

is a vector of length L, the superscript ^{T} denotes transpose of a vector or a matrix, and $\mathbf{x}(k)$ and $\mathbf{v}(k)$ are defined in a similar way to $\mathbf{y}(k)$ from (G.3). Since $x(k)$ and $v(k)$ are uncorrelated by assumption, the covariance matrix (of size $L \times L$) of the noisy signal can be written as

$$\mathbf{R}_{\mathbf{y}} = E\left[\mathbf{y}(k)\mathbf{y}^T(k)\right] = \mathbf{R}_{\mathbf{x}} + \mathbf{R}_{\mathbf{v}}, \tag{G.4}$$

where $E[\cdot]$ denotes mathematical expectation, and $\mathbf{R}_{\mathbf{x}} = E\left[\mathbf{x}(k)\mathbf{x}^T(k)\right]$ and $\mathbf{R}_{\mathbf{v}} = E\left[\mathbf{v}(k)\mathbf{v}^T(k)\right]$ are the covariance matrices of $\mathbf{x}(k)$ and $\mathbf{v}(k)$, respectively. The noise covariance matrix, $\mathbf{R}_{\mathbf{v}}$, is assumed to be full rank, i.e., equal to L. In the rest, we assume that the rank of the speech covariance matrix, $\mathbf{R}_{\mathbf{x}}$, is equal to $P \leq L$. Then, the objective of speech enhancement (or noise reduction) is to estimate the desired signal sample, $x(k)$, from the observation vector, $\mathbf{y}(k)$. This should be done in such a way that the noise is reduced as much as possible with little or no distortion of the desired signal.

Using the joint diagonalization technique [23], the two symmetric matrices $\mathbf{R}_{\mathbf{x}}$ and $\mathbf{R}_{\mathbf{v}}$ can be jointly diagonalized as follows:

$$\mathbf{B}^T\mathbf{R}_{\mathbf{x}}\mathbf{B} = \mathbf{\Lambda}, \tag{G.5}$$

$$\mathbf{B}^T\mathbf{R}_{\mathbf{v}}\mathbf{B} = \mathbf{I}_L, \tag{G.6}$$

where $\mathbf{B}$ is a full-rank square matrix (of size $L \times L$), $\mathbf{\Lambda}$ is a diagonal matrix whose main elements are real and nonnegative, and $\mathbf{I}_L$ is the $L \times L$ identity matrix. Furthermore, $\mathbf{\Lambda}$ and $\mathbf{B}$ are the eigenvalue and eigenvector matrices, respectively, of $\mathbf{R}_{\mathbf{v}}^{-1}\mathbf{R}_{\mathbf{x}}$, i.e.,

$$\mathbf{R}_{\mathbf{v}}^{-1}\mathbf{R}_{\mathbf{x}}\mathbf{B} = \mathbf{B}\mathbf{\Lambda}. \tag{G.7}$$

Since $\mathbf{R}_{\mathbf{x}}$ is semidefinite and its rank is equal to P, the eigenvalues of $\mathbf{R}_{\mathbf{v}}^{-1}\mathbf{R}_{\mathbf{x}}$ can be ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_P > \lambda_{P+1} = \cdots = \lambda_L = 0$. In other words, the last $L-P$ eigenvalues of the matrix product $\mathbf{R}_{\mathbf{v}}^{-1}\mathbf{R}_{\mathbf{x}}$ are exactly zero while its first P eigenvalues are positive, with λ_1 being the maximum eigenvalue. We denote by $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_L$, the corresponding eigenvectors. The noisy signal covariance matrix can also be diagonalized as

$$\mathbf{B}^T\mathbf{R}_{\mathbf{y}}\mathbf{B} = \mathbf{\Lambda} + \mathbf{I}_L. \tag{G.8}$$

We end this section by defining the input and output signal-to-noise ratios (SNRs):

$$\mathrm{iSNR} = \frac{\mathrm{tr}\left(\mathbf{R_x}\right)}{\mathrm{tr}\left(\mathbf{R_v}\right)} = \frac{\sigma_x^2}{\sigma_v^2}, \tag{G.9}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a square matrix, and $\sigma_x^2 = E\left[x^2(k)\right]$ and $\sigma_v^2 = E\left[v^2(k)\right]$ are the variances of $x(k)$ and $v(k)$, respectively, and

$$\mathrm{oSNR}_{\mathrm{nr}}(\mathbf{h}) = \frac{\sigma_{x,\mathrm{nr}}^2}{\sigma_{v,\mathrm{nr}}^2}, \tag{G.10}$$

where $\mathbf{h}$ is a filter applied to the observation signal (see Section 3), and $\sigma_{x,\mathrm{nr}}^2$ and $\sigma_{v,\mathrm{nr}}^2$ are the variances of $x(k)$ and $v(k)$ after noise reduction.

3 Noise Reduction Filtering without Distortion

In this section, we assume that $P < L$; as a result, the speech covariance matrix is rank deficient.

The approach proposed here is based on two successive stages. Firstly, we apply the filter of length L:

$$\mathbf{h} = \left[\begin{array}{cccc} h_0 & h_1 & \cdots & h_{L-1} \end{array}\right]^T \tag{G.11}$$

to the observation signal vector, $\mathbf{y}(k)$, to get the filter output:

$$z(k) = \mathbf{h}^T\mathbf{y}(k) = \mathbf{h}^T\mathbf{x}(k) + \mathbf{h}^T\mathbf{v}(k). \tag{G.12}$$

From (G.4) and (G.12), we deduce that the output SNR from the filter is

$$\mathrm{oSNR}_{\mathrm{f}}\left(\mathbf{h}\right) = \frac{\sigma_{x,\mathrm{f}}^2}{\sigma_{v,\mathrm{f}}^2} = \frac{\mathbf{h}^T\mathbf{R_x}\mathbf{h}}{\mathbf{h}^T\mathbf{R_v}\mathbf{h}}, \tag{G.13}$$

which, in this case, is not the same as the output SNR after noise reduction stated in (G.10). Since the objective is to estimate the noise, we find $\mathbf{h}$ that minimizes $\mathrm{oSNR}_{\mathrm{f}}\left(\mathbf{h}\right)$. Due to the relation $\mathbf{b}_i^T\mathbf{R_x}\mathbf{b}_i = \lambda_i$, it is easy to see that the solution is

$$\mathbf{h}_P = \sum_{i=P+1}^{L} \beta_i \mathbf{b}_i, \tag{G.14}$$

where $\beta_i,\ i = P+1, \ldots, L$, are arbitrary real numbers with at least one of them different from 0. With the filter having the form of (G.14), $\mathrm{oSNR}_{\mathrm{f}}\left(\mathbf{h}_P\right) = 0$ and $z(k)$ can be seen as an estimate of the noise, $\widehat{v}(k) = z(k) = \mathbf{h}_P^T\mathbf{y}(k)$.

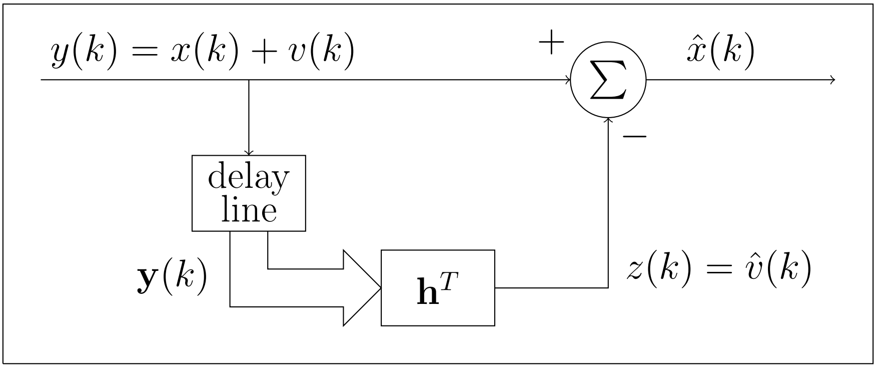


Fig. G.1: Block diagram of the estimation process.

Secondly, we estimate the desired signal, $x(k)$, as

$$\widehat{x}(k) = y(k) - \widehat{v}(k) = x(k) + v(k) - \sum_{i=P+1}^{L} \beta_i \mathbf{b}_i^T \mathbf{v}(k). \tag{G.15}$$

An overview of the estimation process is shown in the block diagram in Figure G.1. Now, we find the β_i's that minimize the power of the residual noise, i.e.,

$$J_{\text{rn}} = E\left\{\left[v(k) - \sum_{i=P+1}^{L} \beta_i \mathbf{b}_i^T \mathbf{v}(k)\right]^2\right\} = \sigma_v^2 - 2\sum_{i=P+1}^{L} \beta_i \mathbf{i}_L^T \mathbf{R}_\mathbf{v} \mathbf{b}_i + \sum_{i=P+1}^{L} \beta_i^2, \tag{G.16}$$

where $\mathbf{i}_L$ is the first column of the $L \times L$ identity matrix. We get

$$\beta_i = \mathbf{i}_L^T \mathbf{R}_\mathbf{v} \mathbf{b}_i. \tag{G.17}$$

Substituting (G.17) into (G.15), the estimator becomes

$$\begin{aligned}\widehat{x}(k) &= x(k) + v(k) - \sum_{i=P+1}^{L} \mathbf{i}_L^T \mathbf{R}_\mathbf{v} \mathbf{b}_i \mathbf{b}_i^T \mathbf{v}(k) \\ &= x(k) + v(k) - \mathbf{i}_L^T \mathbf{R}_\mathbf{v} \left(\mathbf{R}_\mathbf{v}^{-1} - \sum_{p=1}^{P} \mathbf{b}_p \mathbf{b}_p^T\right) \mathbf{v}(k) \\ &= x(k) + \sum_{p=1}^{P} \mathbf{i}_L^T \mathbf{R}_\mathbf{v} \mathbf{b}_p \mathbf{b}_p^T \mathbf{v}(k). \end{aligned} \tag{G.18}$$

The variance of $\widehat{x}(k)$ is

$$\sigma_{\widehat{x}}^2 = \sigma_x^2 + \sigma_v^2 - \sum_{i=P+1}^{L} \left(\mathbf{i}_L^T \mathbf{R}_\mathbf{v} \mathbf{b}_i\right)^2 = \sigma_x^2 + \sum_{p=1}^{P} \left(\mathbf{i}_L^T \mathbf{R}_\mathbf{v} \mathbf{b}_p\right)^2. \tag{G.19}$$

We deduce that the output SNR after noise reduction is

$$\mathrm{oSNR}_{\mathrm{nr}}(\mathbf{h}_P) = \frac{\sigma_x^2}{\sigma_v^2 - \sum_{i=P+1}^{L}\left(\mathbf{i}_L^T\mathbf{R}_{\mathbf{v}}\mathbf{b}_i\right)^2} = \frac{\sigma_x^2}{\sum_{p=1}^{P}\left(\mathbf{i}_L^T\mathbf{R}_{\mathbf{v}}\mathbf{b}_p\right)^2} \geq \mathrm{iSNR}. \tag{G.20}$$

It is clear that the larger $L - P$ is, the larger is the value of the output SNR. Also, from (G.18), we observe that the desired signal is not distorted so that the speech distortion index [1] is

$$\upsilon_{\mathrm{sd}}(\mathbf{h}_P) = \frac{E\{[x_{\mathrm{nr}}(k) - x(k)]^2\}}{E[x^2(k)]} = \frac{E\{[\mathbf{h}_P^T\mathbf{x}(k)]^2\}}{E[x^2(k)]} = 0. \tag{G.21}$$

The noise reduction factor [1] is

$$\xi_{\mathrm{nr}}(\mathbf{h}_P) = \frac{\sigma_v^2}{\sigma_{v,\mathrm{nr}}^2} = \frac{\sigma_v^2}{\sigma_v^2 - \sum_{i=P+1}^{L}\left(\mathbf{i}_L^T\mathbf{R}_{\mathbf{v}}\mathbf{b}_i\right)^2}, \tag{G.22}$$

and since there is no signal distortion, we also have the relation:

$$\frac{\mathrm{oSNR}_{\mathrm{nr}}(\mathbf{h}_P)}{\mathrm{iSNR}} = \xi_{\mathrm{nr}}(\mathbf{h}_P). \tag{G.23}$$

From (G.18), we find a class of distortionless estimators:

$$\widehat{x}_Q(k) = x(k) + \sum_{q=1}^{Q}\mathbf{i}_L^T\mathbf{R}_{\mathbf{v}}\mathbf{b}_q\mathbf{b}_q^T\mathbf{v}(k), \tag{G.24}$$

where $P \leq Q \leq L$. We have $\widehat{x}_P(k) = \widehat{x}(k)$ and $\widehat{x}_L(k) = y(k)$. The latter is the observation signal itself. It is obvious that the output SNR corresponding to $\widehat{x}_Q(k)$ is

$$\mathrm{oSNR}_{\mathrm{nr}}(\mathbf{h}_Q) = \frac{\sigma_x^2}{\sum_{q=1}^{Q}\left(\mathbf{i}_L^T\mathbf{R}_{\mathbf{v}}\mathbf{b}_q\right)^2} \geq \mathrm{iSNR} \tag{G.25}$$

and

$$\mathrm{oSNR}_{\mathrm{nr}}(\mathbf{h}_P) \geq \mathrm{oSNR}_{\mathrm{nr}}(\mathbf{h}_{P+1}) \geq \mathrm{oSNR}_{\mathrm{nr}}(\mathbf{h}_L) = \mathrm{iSNR}. \tag{G.26}$$

4 Noise Reduction Filtering with Distortion

In this section, we assume that the speech covariance matrix is full rank, i.e., equal to L. We can still use the method presented in the previous section, but this time we should expect distortion of the desired signal.

Again, we apply the filter:

$$\mathbf{h}' = \begin{bmatrix} h'_0 & h'_1 & \cdots & h'_{L-1} \end{bmatrix}^T \tag{G.27}$$

of length L to the observation signal vector. Then, the filter output and output SNR are, respectively,

$$z'(k) = \mathbf{h}'^T \mathbf{x}(k) + \mathbf{h}'^T \mathbf{v}(k) \tag{G.28}$$

and

$$\text{oSNR}_{\text{f}}\left(\mathbf{h}'\right) = \frac{\mathbf{h}'^T \mathbf{R}_{\mathbf{x}} \mathbf{h}'}{\mathbf{h}'^T \mathbf{R}_{\mathbf{v}} \mathbf{h}'}. \tag{G.29}$$

Now, we choose

$$\mathbf{h}'_{P'} = \sum_{i=P'+1}^{L} \beta'_i \mathbf{b}_i, \tag{G.30}$$

where $\beta'_i, \ i = P' + 1, \ldots, L$, are arbitrary real numbers. With this choice of $\mathbf{h}'$, the output SNR becomes

$$\text{oSNR}_{\text{f}}\left(\mathbf{h}'_{P'}\right) = \frac{\sum_{i=P'+1}^{L} \beta'^2_i \lambda_i}{\sum_{i=P'+1}^{L} \beta'^2_i}. \tag{G.31}$$

This time, however, the output SNR cannot be equal to 0 but we can make it as small as we desire. The larger is the value of $\text{oSNR}_{\text{f}}\left(\mathbf{h}'_{P'}\right)$, the more the speech signal is distorted. If we can tolerate a small amount of distortion, then we can still consider $z'(k)$ as an estimate of the noise, $\widehat{v}'(k) = z'(k) = \mathbf{h}'^T_{P'} \mathbf{y}(k)$.

In the second stage, we estimate the desired signal as

$$\widehat{x}'(k) = y(k) - \widehat{v}'(k) = x(k) - \sum_{i=P'+1}^{L} \beta'_i \mathbf{b}_i^T \mathbf{x}(k) + v(k) - \sum_{i=P'+1}^{L} \beta'_i \mathbf{b}_i^T \mathbf{v}(k). \tag{G.32}$$

By minimizing the power of the residual noise:

$$J'_{\text{rn}} = E\left\{\left[v(k) - \sum_{i=P'+1}^{L} \beta'_i \mathbf{b}_i^T \mathbf{v}(k)\right]^2\right\} = \sigma_v^2 - 2\sum_{i=P'+1}^{L} \beta'_i \mathbf{i}_L^T \mathbf{R}_{\mathbf{v}} \mathbf{b}_i + \sum_{i=P'+1}^{L} \beta'^2_i, \tag{G.33}$$

we find that

$$\beta'_i = \mathbf{i}_L^T \mathbf{R}_{\mathbf{v}} \mathbf{b}_i = \frac{1}{\lambda_i} \mathbf{i}_L^T \mathbf{R}_{\mathbf{x}} \mathbf{b}_i. \tag{G.34}$$

Substituting (G.34) into (G.32), we obtain

$$\widehat{x}'(k) = x(k) - \sum_{i=P'+1}^{L} \frac{1}{\lambda_i} \mathbf{i}_L^T \mathbf{R}_{\mathbf{x}} \mathbf{b}_i \mathbf{b}_i^T \mathbf{x}(k) + v(k) - \sum_{i=P'+1}^{L} \mathbf{i}_L^T \mathbf{R}_{\mathbf{v}} \mathbf{b}_i \mathbf{b}_i^T \mathbf{v}(k). \tag{G.35}$$

The variance of $\widehat{x}'(k)$ is

$$\sigma_{\widehat{x}'}^2 = \sigma_x^2 - \sum_{i=P'+1}^{L} \frac{1}{\lambda_i} \left(\mathbf{i}_L^T \mathbf{R}_{\mathbf{x}} \mathbf{b}_i\right)^2 + \sigma_v^2 - \sum_{i=P'+1}^{L} \left(\mathbf{i}_L^T \mathbf{R}_{\mathbf{v}} \mathbf{b}_i\right)^2. \tag{G.36}$$

We deduce that the output SNR and speech distortion index are, respectively,

$$\mathrm{oSNR}_{\mathrm{nr}}(\mathbf{h}'_{P'}) = \frac{\sigma_x^2 - \sum_{i=P'+1}^{L} \frac{1}{\lambda_i} \left(\mathbf{i}_L^T \mathbf{R}_{\mathbf{x}} \mathbf{b}_i\right)^2}{\sigma_v^2 - \sum_{i=P'+1}^{L} \left(\mathbf{i}_L^T \mathbf{R}_{\mathbf{v}} \mathbf{b}_i\right)^2} \tag{G.37}$$

and

$$\upsilon_{\mathrm{sd}}(\mathbf{h}'_{P'}) = \frac{1}{\sigma_x^2} \sum_{i=P'+1}^{L} \frac{1}{\lambda_i} \left(\mathbf{i}_L^T \mathbf{R}_{\mathbf{x}} \mathbf{b}_i\right)^2. \tag{G.38}$$

The smaller P' is compared to L, the larger is the distortion. Further, the speech distortion index is independent of the input SNR, as is the gain in SNR. This can be observed by multiplying either $\mathbf{R}_{\mathbf{x}}$ in (G.5) or $\mathbf{R}_{\mathbf{v}}$ in (G.6) by a constant c, which leads to a corresponding change in the input SNR. Insertion of the resulting λ_i's and $\mathbf{b}_i$'s in (G.37) and (G.38) will show that the output SNR is changed by the factor c and that the speech distortion index is independent of c.

The output SNR and the speech distortion index are related as follows:

$$\frac{\mathrm{oSNR}_{\mathrm{nr}}(\mathbf{h}'_{P'})}{\mathrm{iSNR}} = \left[1 - \upsilon_{\mathrm{sd}}(\mathbf{h}'_{P})\right] \xi_{\mathrm{nr}}(\mathbf{h}'_{P}), \tag{G.39}$$

where

$$\xi_{\mathrm{nr}}(\mathbf{h}'_{P'}) = \frac{\sigma_v^2}{\sigma_v^2 - \sum_{i=P'+1}^{L} \left(\mathbf{i}_L^T \mathbf{R}_{\mathbf{v}} \mathbf{b}_i\right)^2} \tag{G.40}$$

is the noise reduction factor.

Interestingly, the exact same estimator is obtained by minimizing the power of the residual desired signal:

$$J'_{\mathrm{rd}} = E\left\{\left[x(k) - \sum_{i=P'+1}^{L} \beta'_i \mathbf{b}_i^T \mathbf{x}(k)\right]^2\right\} = \sigma_x^2 - 2 \sum_{i=P'+1}^{L} \beta'_i \mathbf{i}_L^T \mathbf{R}_{\mathbf{x}} \mathbf{b}_i + \sum_{i=P'+1}^{L} \lambda_i \beta_i'^2. \tag{G.41}$$

Again, minimizing J'_{rn} or J'_{rd} leads to the estimator $\widehat{x}'(k)$.

Alternatively, another set of estimators can be obtained by minimizing the mean squared error between $x(k)$ and $\widehat{x}'(k)$:

$$\begin{aligned} J'_{\text{mse}} &= E\left\{\left[v(k) - \sum_{i=P'+1}^{L} \beta'_i \mathbf{b}_i^T \mathbf{v}(k) - \sum_{i=P'+1}^{L} \beta'_i \mathbf{b}_i^T \mathbf{x}(k)\right]^2\right\} \\ &= \sigma_v^2 - 2\sum_{i=P'+1}^{L} \beta'_i \mathbf{i}_L^T \mathbf{R}_{\mathbf{v}} \mathbf{b}_i + \sum_{i=P'+1}^{L} (1+\lambda_i)\beta_i'^2, \end{aligned} \tag{G.42}$$

which leads to

$$\beta'_i = \frac{\mathbf{i}_L^T \mathbf{R}_{\mathbf{v}} \mathbf{b}_i}{1+\lambda_i}. \tag{G.43}$$

In the special case where $P' = 0$ the estimator is the well known Wiener filter.

5 Simulations

In this section, the filter design with and without distortion are evaluated through simulations. Firstly, the distortionless case is considered in order to verify that the basics of the filter design hold and the filter works as expected. Secondly, we turn to the filter design with distortion to investigate the influence of the input SNR and the choice of P' on the output SNR and the speech distortion index.

The distortionless filter design was tested by use of a synthetic harmonic signal. The use of such a signal makes it possible to control the rank of the signal covariance matrix, which is a very important feature in the present study. Further, the harmonic signal model is used to model voiced speech, e.g., in [24]. The harmonic signal model has the form:

$$x(k) = \sum_{m=1}^{M} A_m \cos(m 2\pi f_0/f_{\text{s}} k + \phi_m) \tag{G.44}$$

where M is the model order, $A_m > 0$ and $\phi_m \in [0, 2\pi]$ are the amplitude and phase of the m'th harmonic, $f_0 \in [0, \pi/m]$ is the fundamental frequency and f_{s} is the sampling frequency. The rank of the signal covariance matrix, $\mathbf{R}_{\mathbf{x}}$, is then $P = 2M$. In the simulations $M = 5$, the amplitudes are decreasing with the frequency, f, as $1/f$, normalised to give $A_1 = 1$, the fundamental frequency is chosen randomly such that $f_0 \in [150, 250]$ Hz, the sampling frequency is 8 kHz and the phases are random. The covariance matrices of $\mathbf{R}_{\mathbf{x}}$ and $\mathbf{R}_{\mathbf{v}}$ are estimated from segments of 230 samples and are updated along with the filter for each sample. The number of samples is 1000.

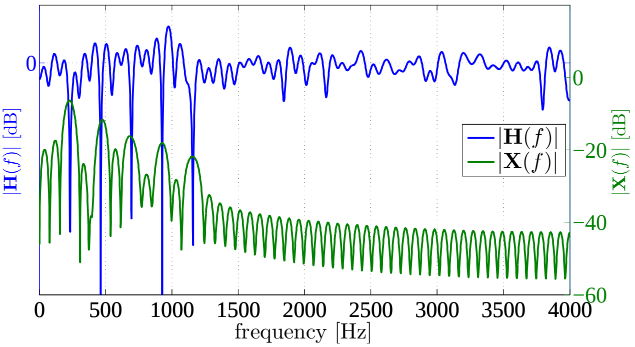


Fig. G.2: Spectrum of the signal vector, $\mathbf{x}(k)$, and the corresponding filter, $\mathbf{h}_P$.

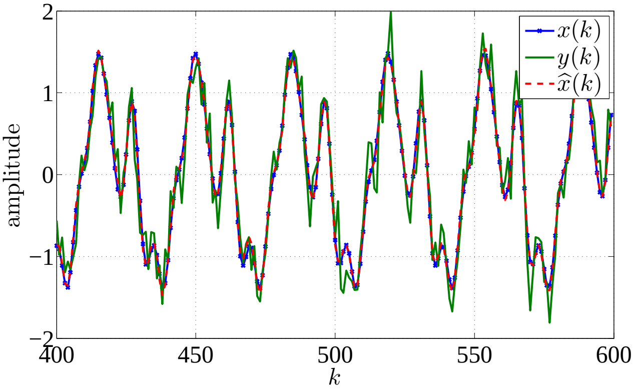


Fig. G.3: Desired signal, $x(k)$, noisy observation, $y(k)$, and estimated signal, $\widehat{x}(k)$.

As an example, the spectrum of a synthetic signal is shown in Fig. G.2 along with the frequency response of the corresponding filter. The fundamental frequency is in this case $f_0 = 200$ Hz and the filter has a length of $L = 110$. After subtraction of the filter output from the noisy observation the estimate of the desired signal, shown in Fig. G.3, results. The desired signal and the noisy observation are shown as well. Comparing the signals it is easily seen that the filtering has improved the output SNR in the estimated signal relative to the noisy observation.

In order to support this, 100 Monte Carlo simulations have been performed for different lengths of the filter, and the performance are evaluated by the output SNR and speech distortion index. The output SNR is calculated according to (G.10) as the ratio of the variances of the desired signal after noise reduction, $[x(k)-\mathbf{h}_P^T\mathbf{x}(k)]$, and the noise after noise reduction, $[v(k)-\mathbf{h}_P^T\mathbf{v}(k)]$, whereas the speech distortion index is calculated according to (G.21) as the ratio of the variance of the filtered desired signal to the variance of the original

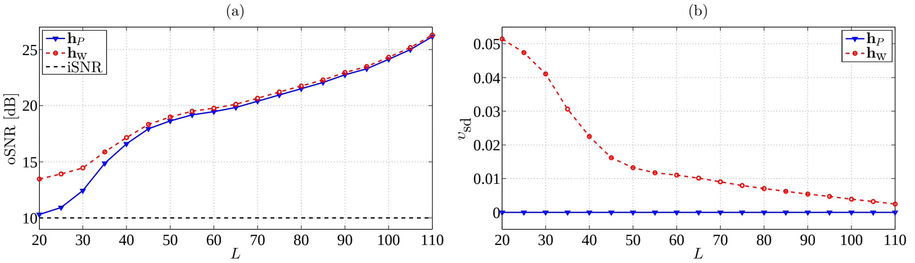


Fig. G.4: Performance as a function of L for a signal with rank deficient covariance matrix. (a) Output SNR and (b) speech distortion index as a function of the filterlength, L, for a real synthetic harmonic signal simulating voiced speech.

desired signal. As is seen in Fig. G.4(a), it is definitely possible to increase the SNR, but the extent is highly dependent on the length of the filter. For short filter lengths, the filter has almost no effect and oSNR $\approx$ iSNR, but as the filter length is increased, the output SNR is increased as well. Even though the estimates of the covariance matrices worsen when the filter length is increased, the longest filter gives rise to the best output SNR. By increasing the filter length from 20 to 110, a gain in SNR of more than 15 dB can be obtained. The corresponding speech distortion index, shown in Fig. G.4(b), is zero for all filter lengths, as was the basis for the filter design. As a reference, results for the Wiener filter ($\mathbf{h}_\mathrm{W}$) are shown as well. The Wiener filter is constructed based on [15] where it is derived based on joint diagonalization. The proposed method has a slightly lower output SNR, especially at short filter lengths. On the other hand, the Wiener filter introduces distortion of the desired signal at all filter lengths, whereas the proposed filter is distortionless.

When the covariance matrix of the desired signal is full rank, speech distortion is introduced in the reconstructed speech signal. This situation was evaluated by use of auto regressive (AR) models, since these can be used to describe unvoiced speech [25]. The models used were of second order and the coefficients were found based on ten segments of unvoiced speech from the Keele database [26], resampled to give a sampling frequency of 8 kHz and a length of 400 samples after resampling. Again, P' was set to 10, the signal was added white Gaussian noise to give an average input SNR of 10 dB and 100 Monte Carlo simulations were run on each of the ten generated signals in order to see the influence of the filter length when the signal covariance matrix is full rank. The results are shown in Fig. G.5. As was the case for voiced speech, it is possible to gain approximately 15 dB in SNR by increasing the filter length from 20 to 110. However, this time the speech distortion is also dependent on the filterlength, and the longer the filter the more signal distortion. In this case, comparison to the Wiener filter shows just the opposite situation than with the harmonic model. Now, the gain in SNR is higher for the proposed

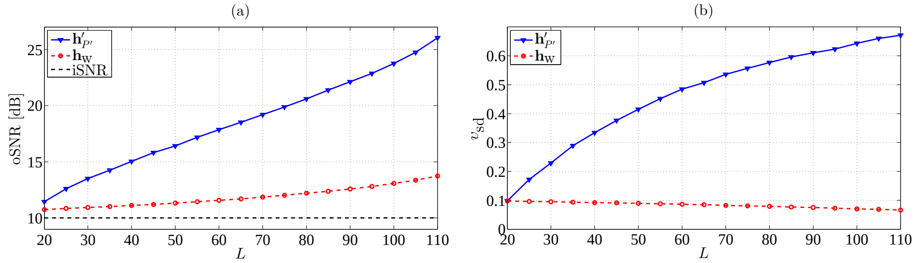


Fig. G.5: Performance as a function of L for a signal with full rank covariance matrix. (a) Output SNR and (b) speech distortion index as a function of the filter length, L, for a signal generated by an AR process simulating unvoiced speech.

method for all filter lengths, but the signal is also more distorted.

After having investigated the filter performance for different filter lengths using synthetic signals, the influence of input SNR and the choice of P' is investigated directly in speech signals. Again, we used signals from the Keele database with $f_\mathrm{s} = 8$ kHz. Excerpts with a length of 20,000 were extracted from different places in the speech signals from two male and two female speakers. Noise was added to give the desired average input SNR and filters with a length $L = 110$ and varying P' were applied. Three different kinds of noise were used, white Gaussian, babble, and car noise, the last two from the AURORA database [27]. The output SNR and signal distortion index are depicted as a function of P' in Fig. G.6. Both the output SNR and the speech distortion index are decreasing with P', as was depicted in Section 4. Thereby, the choice of P' will be a compromise between a high output SNR and a low speech distortion index. In Fig. G.7, the proposed filter is compared, at an input SNR of 10 dB, to the Wiener filter and three filters from [10] ($\mathbf{h}_\mathrm{ls}$,$\mathbf{h}_\mathrm{mv}$,$\mathbf{h}_\mathrm{mls}$), which are subspace-based filters as well. These filters are based on a Hankel representation of the observed signal, which we, from the segment length of 230 samples, construct with a size of 151 times 80. Due to restrictions on the chosen rank (according to P'), this is only varied from 1 to 71. The performance of the Wiener filter is of course independent of P' and it is, therefore, possible to construct a filter that either gives a higher output SNR or a lower speech distortion than the Wiener filter, dependent on the choice of P'. The filters from [10] are dependent on P' as well, but the proposed filter has a broader range of possible combinations of output SNR and speech distortion. At $P' = 1$, a gain in output SNR of approximately 5 dB can be obtained while the speech distortion is comparable. At the other extreme, it is possible to obtain the same output SNR as $\mathbf{h}_\mathrm{ls}$ while the speech distortion index is lowered by approximately 5 dB.

The choice of the value of P' is, however, not dependent on the input SNR, as seen in Fig. G.8, since both the gain in SNR and the speech distortion index

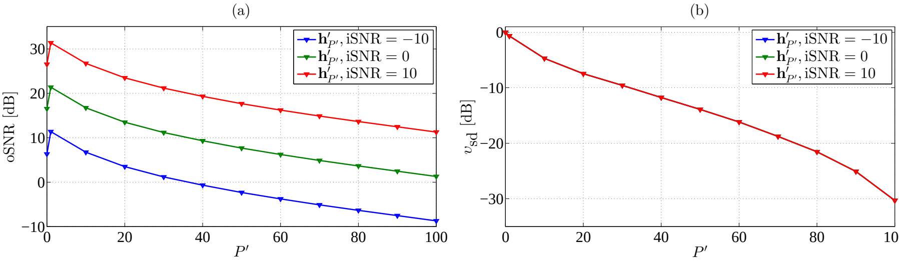


Fig. G.6: Performance as a function of P'. (a) Output SNR and (b) speech distortion index as a function of P' for a speech signal with full rank covariance matrix.

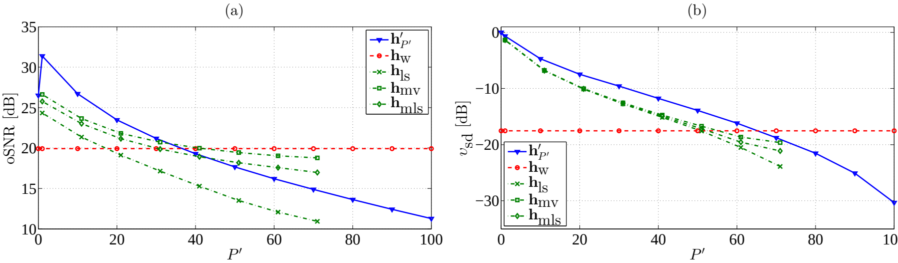


Fig. G.7: Performance as a function of P' compared to other filtering methods. (a) Output SNR and (b) speech distortion index as a function of P' for a speech signal with full rank covariance matrix compared to the Wiener filter and three filters from [10] at an iSNR of 10 dB.

are constant functions of the input SNR, as was also found theoretically in Section 4. This means that it is possible to construct a filter according to the desired combination of gain in SNR and speech distortion, and then this will apply no matter the input SNR. This is not the case for either the Wiener filter or the filters from [10] as is seen in Fig. G.9. For these filters, the gain in SNR is decreasing with input SNR (except for $\mathbf{h}_{\text{ls}}$ which is also constant) as is the speech distortion index.

As a measure of the subjective evaluation, Perceptual Evaluation of Speech Quality (PESQ) scores [28] have been calculated for different filter lengths, different values of P' and different SNRs. The used speech signal contains 40,000 samples from the beginning of the speech signal from the first female speaker in the Keele database. The results are shown in Table G.1 and Table G.2. It is seen that the PESQ scores are increasing with increasing filter length and SNR, even though the effect of going from a filter length of 90 to 110 seems smaller than increasing the length from 70 to 90. The PESQ score is rather low for low values of P', peaks for $P' = 31$ or $P' = 41$, depending on the SNR, and then decreases again for higher values of P'. This is also heard in informal listening tests of the resulting speech signal. At low values of P', the

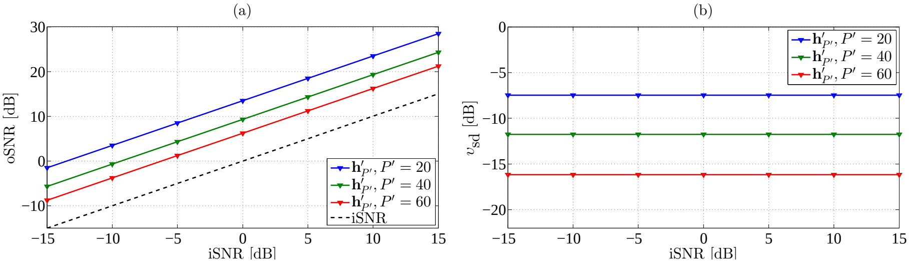


Fig. G.8: Performance as a function of the input SNR. (a) Output SNR and (b) speech distiortion index as a function of the input SNR for a speech signal with full rank covariance matrix.

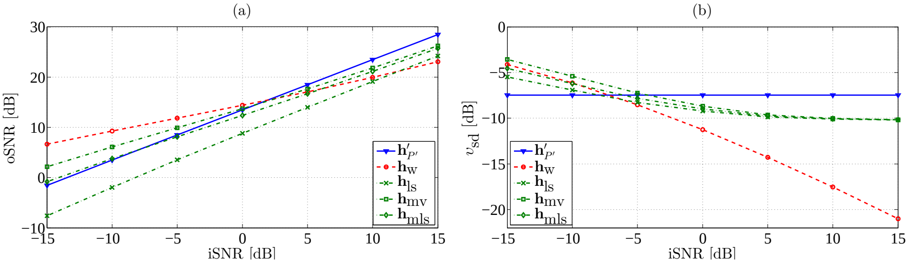


Fig. G.9: Performance as a function of the input SNR compared to other filtering methods. (a) Output SNR and (b) speech distiortion index as a function of the input SNR for a speech signal with full rank covariance matrix compared to the Wiener filter and three filters from [10] at an input SNR of 10dB.

speech signal sounds rather distorted whereas at high levels of P', the signal is noisy, but not very distorted, which also confirms the findings in Fig. G.6. As reflected in the PESQ score, a signal with a compromise between the two is preferred if the purpose is listening directly to the output. In such a context, the performance of the Wiener filter is slightly better than the proposed filter with PESQ scores approximately 0.3 units larger. However, the purpose of noise reduction is sometimes as a pre-processor to, e.g., a speech recognition algorithm. Here, the word error rate increases when the SNR decreases [29, 30], but on the other hand the algorithms are also sensible to distortion of the speech signal [31, 32]. In such cases it might, therefore, be optimal with another relationship between SNR and speech distortion than the one having the best perceptual performance. This optimisation is possible with the proposed filter due to its flexibility.

The effect of choosing different values of P' is visualized in Fig. G.10. Figure G.10(a) shows the spectrogram of a piece of a clean speech signal from the Keele database and in Fig. G.10(b) babble noise was added to give an average input

Table G.1: PESQ-scores at different filter lengths and SNRs for $P' = 31$.

SNR [dB]	$\mathbf{h}_{P'}$, $P' = 31$		
	$L = 70$	$L = 90$	$L = 110$
0	2.160	2.353	2.467
5	2.476	2.656	2.737
10	2.808	2.919	2.920

Table G.2: PESQ-scores for different values of P' and SNR for a filter length of 110.

SNR [dB]	$\mathbf{h}_{\mathrm{w}}$	$\mathbf{h}_{P'}$						
		$P' = 1$	$P' = 11$	$P' = 21$	$P' = 31$	$P' = 41$	$P' = 51$	$P' = 61$
0	2.799	1.051	2.173	2.421	2.467	2.372	2.256	2.159
5	3.086	1.072	2.236	2.580	2.737	2.708	2.610	2.520
10	3.328	1.067	2.274	2.683	2.920	2.999	2.961	2.876

SNR of 10 dB. Fig. G.10(c) and (d) show the spectrograms of the reconstructed speech signal with two different choices of P'. The former is a reconstruction based on $P' = 10$. Definitely, the noise content is reduced when comparing to the noisy speech signal in Fig. G.10(b). However, a high degree of signal distortion has been introduced as well, which can be seen especially in the voiced speech parts, where the distinction between the harmonics is blurred compared to both the clean speech signal and the noisy speech signal. In the latter figure $P' = 70$ and, therefore, both noise reduction and signal distortion are not as prominent as when $P' = 10$. Here, the harmonics are much more well preserved, but, as is seen in the background, it comes with the price of less noise reduction.

A feature of the proposed filter, which is not explored here, is the possibility of choosing different values of P' over time. The optimal value of P' depends on whether the speech is voiced or unvoiced, and how many harmonics there are in the voiced parts. By adapting the value of P' at each time step based on this information, it should be possible to simultaneously achieve a higher SNR and a lower distortion.

6 Conclusions

In this paper, we have presented a new perspective on time-domain single-channel noise reduction based on forming filters from the eigenvectors that diagonalize both the desired and noise signal covariance matrices. These filters are chosen so that they provide an estimate of the noise signal when applied to the observed signal. Then, by subtraction of the noise estimate from the observed signal, an estimate of the desired signal can be obtained. Two cases have been considered, namely one where no distortion is allowed on the desired

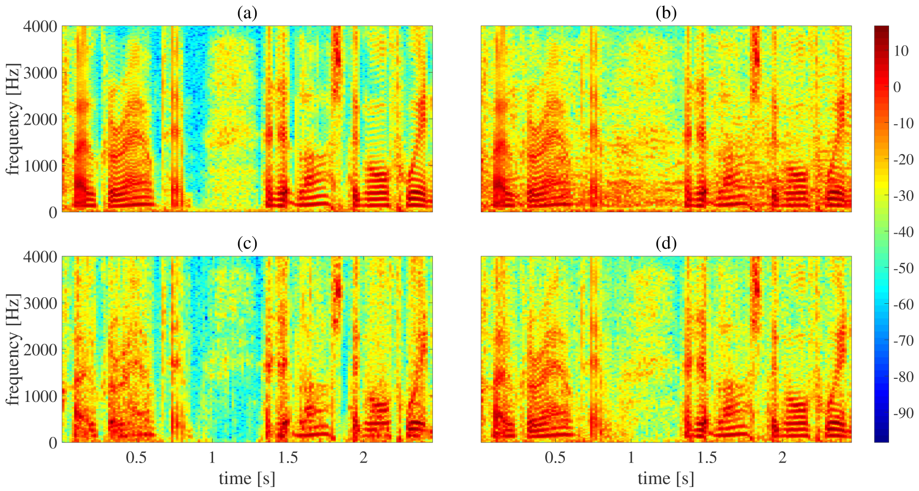


Fig. G.10: Spectra of desired signal, noisy signal and two reconstructions with different choices of P'. (a) Spectrum of a part of a speech signal from the Keele database. (b) Speech signal from (a) contaminated with babble noise to give an average input SNR of 10 dB. (c) Reconstructed speech signal using $P' = 10$. (d) Reconstructed speech signal using $P' = 70$.

signal and one where distortion is allowed. The former case applies to signals that have a rank that can be assumed to be less than the rank of the observed signal covariance matrix, which is, for example, the case for voiced speech. The latter case applies to desired signals that have a full-rank covariance matrix. In this case, the only way to achieve noise reduction is by also allowing for distortion on the desired signal. The amount of distortion introduced depends on a parameter corresponding to the rank of an implicit approximation of the desired signal covariance matrix. As such, it is relatively easy to control the trade-off between noise reduction and speech distortion. Experiments on real and synthetic signals have confirmed these principles and demonstrated how it is, in fact, possible to achieve higher output signal-to-noise ratio or a lower signal distortion index with the proposed method than with the classical Wiener filter. Moreover, the results show that only a small loss in output signal-to-noise ratio is incurred when no distortion can be accepted, as long as the filter is not too short. The results also show that when distortion is allowed on the desired signal, the amount of distortion is independent of the input signal-to-noise ratio. The presented perspective is promising in that it unifies the ideas behind subspace methods and optimal filtering, two methodologies that have traditionally been seen as quite different.

References

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

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

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

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

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

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

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

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

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

[10] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP J. on Advances in Signal Processing*, vol. 2007, no. 1, p. 24, Jun. 2007.

[11] S. Rangachari and P. Loizou, "A noise estimation algorithm for highly nonstationary environments," *Speech Communication*, vol. 28, pp. 220–231, 2006.

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

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

[14] R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, “Low complexity DFT-domain noise PSD tracking using high-resolution periodograms,” *EURASIP J. on Advances in Signal Processing*, vol. 2009(1), p. 15, 2009.

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

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

[17] J. Benesty, M. Souden, and J. Chen, “A perspective on multichannel noise reduction in the time domain,” *Applied Acoustics*, vol. 74, no. 3, pp. 343–355, Mar. 2013.

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

[19] M. G. Christensen and A. Jakobsson, “Optimal filter designs for separating and enhancing periodic signals,” *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

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

[21] ——, “Non-causal time-domain filters for single-channel noise reduction,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 5, pp. 1526–1541, Jul. 2012.

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

[23] J. N. Franklin, *Matrix Theory.* Prentice-Hall, 1968.

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

[25] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals.* New York: Wiley, 2000.

[26] F. Plante, G. F. Meyer, and W. A. Ainsworth, “A pitch extraction reference database,” in *Proc. Eurospeech*, Sep. 1995, pp. 837–840.

[27] D. Pearce and H. G. Hirsch, “The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *Proc. Int. Conf. Spoken Language Process.*, Oct 2000.

[28] Y. Hu and P. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[29] X. Cui and A. Alwan, “Noise robust speech recognition using feature compensation based on plynomial regression of utterance snr,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 6, pp. 1161–1172, nov 2005.

[30] R. P. Lippmann, “Speech recognition by machines and humans,” *Speech Communication*, vol. 22, no. 1, pp. 1 – 15, 1997.

[31] J. M. Huerta and R. M. Stern, “Distortion-class weighted acoustic modeling for robust speech recognition under GSM RP-LTP coding,” in *Proc. of the robust methods for speech recognition in adverse conditions*, 1999.

[32] T. Takiguchi and Y. Ariki, “PCA-based speech enhancement for distorted speech recognition,” *Journal of Multimedia*, vol. 2, no. 5, pp. 13–18, Sep 2007.

SUMMARY

This thesis deals with speech enhancement, i.e., noise reduction in speech signals. This has applications in, e.g., hearing aids and teleconference systems. We consider a signal-driven approach to speech enhancement where a model of the speech is assumed and filters are generated based on this model. The basic model used in this thesis is the harmonic model which is a commonly used model for describing the voiced part of the speech signal. We show that it can be beneficial to extend the model to take inharmonicities or the non-stationarity of speech into account. Extending the model introduces extra parameters and we suggest methods to estimate these extra parameters and derive filters based on the extended models.

ISSN (online): 2246-1248
ISBN (online): 978-87-7112-404-0

AALBORG **UNIVERSITETSFORLAG**