Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Enhancement and Noise Statistics Estimation for Non-Stationary Voiced Speech

Nørholm, Sidsel Marie; Jensen, Jesper Rindom; Christensen, Mads Græsbøll

Published in:
I E E E Transactions on Audio, Speech and Language Processing

DOI (link to publication from Publisher):
10.1109/TASLP.2016.2514492

Creative Commons License
Unspecified

Publication date:
2016

Document Version
Early version, also known as pre-print

Link to publication from Aalborg University

Citation for published version (APA):
Nørholm, S. M., Jensen, J. R., & Christensen, M. G. (2016). Enhancement and Noise Statistics Estimation for Non-Stationary Voiced Speech. *I E E E Transactions on Audio, Speech and Language Processing*, *24*(4), 645-658. https://doi.org/10.1109/TASLP.2016.2514492

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: March 28, 2024

Enhancement and Noise Statistics Estimation for Non-Stationary Voiced Speech

Sidsel Marie Nørholm, Jesper Rindom Jensen, *Member, IEEE,*
and Mads Græsbøll Christensen, *Senior Member, IEEE*

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

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

I. Introduction

SPEECH enhancement has many applications as in, e.g., mobile phones and hearing aids. Often, the speech enhancement is carried out in a transformed domain, a common one being the frequency domain. Here, the methods based on computational auditory scene analysis (CASA) [2], [3], spectral subtraction [4] and Wiener filtering [5] are well known methods. The CASA methods are based on feature extraction of the speech signal whereas spectral subtraction and Wiener filtering require an estimate of the power spectral density (PSD) of the noise. The PSD can be estimated in different ways [6]–[8], but common to these methods is that they primarily rely on periods without speech to update the noise statistics. In periods of speech, the PSD is mostly given by the previous estimate of the PSD. This update pattern makes the PSD estimates very vulnerable to non-stationary noise. Furthermore, in order to make enhancement in the frequency domain, the data needs to be transformed by use of the Fourier transform. This transform assumes that the signals are stationary within the analysis window which for speech signals is often between 20 ms and 30 ms. It is, however, well known that this assumption of stationary speech does not hold [9], [10], as, e.g., the fundamental frequency and formants vary continuously over time in periods of voiced speech, making the speech signal non-stationary. One example of this is the diphthong where one vowel is followed directly by another with a smooth transition. In [11], [12], it is suggested replacing the standard Fourier transform with a fan-chirp transform in the analysis of non-stationary harmonic signals. The voiced speech parts of a speech signal are often described by a harmonic model, and since voiced speech is the main constituent of speech, it makes good sense to use this transform on speech signals. The voiced speech can also easily be separated from the unvoiced speech by use of voiced/unvoiced detectors [13], [14]. The assumption behind the fan-chirp transform is that the harmonic frequencies of the signal vary linearly over time, and it is shown that spectra obtained using the fan-chirp transform have much more distinct peaks at the positions of the harmonic frequencies. Alternatively, the enhancement can be done directly in the time domain where, e.g., the Wiener filter has also been defined [15]. Most time domain filters also depend on noise statistics in the form of a covariance matrix. These are often obtained by averaging over a small frame of the observed signal, and, therefore, the signal in these frames is also assumed stationary. Also, a common way to filter speech in the time domain is by describing the voiced speech parts by a harmonic model [16]–[18]. The signal based on this model is composed of a set of sinusoids where the frequency of each sinusoid is given by an integer multiple of a fundamental frequency. The fundamental frequency in this model is constant within a frame, and so the voiced speech is assumed stationary. In [17], it is proposed estimating the noise by subtracting an estimate of the desired signal based on the harmonic model, and, from this, make a noise covariance matrix estimate. In doing so, the observed signal only needs to be stationary within the frame of 20 to 30 ms when the noise statistics are estimated and not from one speech free period to the next, as was mostly the case for the PSD. The non-stationarity of speech is considered in [19]–[21] in relation to modelling and parameter estimation. In these papers, a modified version of the harmonic model is used

This work was funded by the Villum Foundation and the Danish Council for Independent Research, grant ID: DFF 1337-00084. Part of this material will be published at INTERSPEECH 2015 [1].

S. M. Nørholm, J. R. Jensen and M. G. Christensen are with the Audio Analysis Lab, AD:MT, Aalborg University, DK-9000 Aalborg, Denmark, e-mail: {smn, jrj, mgc}@create.aau.dk

where a chirp parameter is introduced to allow the frequency of the harmonics to change linearly within each frame. In [19], the first model introduced to describe the speech signal is very flexible, but it is approximated with a Taylor expansion that leads to bigger and bigger deviations from the original model when the harmonic number increases, as mentioned in the paper. In [20], [21], a harmonic chirp model is used to describe the voiced speech, and the parameters of the model are estimated based on maximum likelihood estimation, but using different ways to avoid making a two dimensional search for the fundamental frequency and chirp rate.

In this work we want to explore if there is a benefit of taking the non-stationarity of speech into account when speech enhancement is considered. Therefore, we investigate the harmonic chirp model further in relation to speech enhancement. The linearly constrained minimum variance (LCMV) and the amplitude and phase estimation (APES) filters have previously been derived under the harmonic framework [18], [22], [23]. One objective of this work is to increase the performance of these filters by deriving them according to the harmonic chirp model. Both LCMV and APES filter have the goal of minimising the output noise power from the filter under the constraint that the desired signal is passed undistorted, or equivalently, when the constraint is fulfilled, to maximise the output signal-to-noise ratio (SNR). Therefore, we evaluate the performance of the filters by use of the output SNR and the signal reduction factor which measures the distortion of the desired signal introduced by the filters. Another objective is to investigate the noise covariance matrix that is obtained implicitly when the APES based filter is made in relation to other filters as, e.g., the Wiener filter. The noise covariance matrix estimate is made under the assumption of non-stationary speech when the harmonic chirp model is used. It is generated from the covariance matrix of the observed signal by subtracting the part that conforms to the harmonic chirp model. We propose using this estimate in combination with other filters as well and compare the performance of the Wiener filter using the APES noise covariance matrix to the chirp APES based filter. Alternatively, we suggest estimating the noise covariance matrix based on the earlier mentioned state of the art PSD estimates [7], [8] since more work has been put into noise PSD estimates than estimation of time domain noise statistics. The PSD is related through the Fourier transform to the autocorrelation and, thereby, to the covariance matrix as well.

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

II. Framework

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

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

for discrete time indices $n = 0, ..., N-1$. The desired signal and noise are assumed to be zero mean signals and mutually uncorrelated. Further, we assume that the desired signal is quasi periodic which is a reasonable assumption for voiced speech. Often, voiced speech is described by a harmonic model [18], [24], [25], but here we are using a harmonic chirp model which makes the model capable of handling non-stationary speech.

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

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

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

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

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

$$s(n) = \sum_{l=1}^{L} A_l \cos\left(\theta_l(n)\right) \tag{4}$$

$$= \sum_{l=1}^{L} A_l \cos\left(l\left(\omega_0 n + \frac{k}{2}n^2\right) + \phi_l\right). \tag{5}$$

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

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

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

signal model [16]:

$$s(n) = \sum_{l=1}^{L} a\cos\left(l\left(\omega_0 n + \frac{k}{2}n^2\right)\right) - b\sin\left(l\left(\omega_0 n + \frac{k}{2}n^2\right)\right), \quad (7)$$

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

$$s(n) = \sum_{l=1}^{L} \alpha_l e^{jl\left(\omega_0 n + k/2n^2\right)} + \alpha_l^* e^{-jl\left(\omega_0 n + k/2n^2\right)} = \sum_{l=1}^{L} \alpha_l z^l(n) + \alpha_l^* z^{-l}(n), \quad (8)$$

where

$$z(n) = e^{-j(\omega_0 n + k/2n^2)} \quad (9)$$

and $\alpha_l = \frac{A_l}{2}e^{j\phi}$. Since (7) and (8) are two ways of describing the same signal, it is possible to design optimal filters based on both, but the complex model in (8) gives a more intuitive and simple notation, and, therefore, we will use this model in the following instead of the real model in (7) [16].

Defining a subvector of samples

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

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

$$\mathbf{s}(n) = \mathbf{Z}\mathbf{a}, \quad (11)$$

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

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

with

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

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

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

$$\mathbf{x}(n) = \mathbf{s}(n) + \mathbf{v}(n), \quad (14)$$

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

$$\text{iSNR} = \frac{\sigma_s^2}{\sigma_v^2}. \quad (15)$$

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

$$\hat{s}(n) = \mathbf{h}^H\mathbf{x}(n) = \mathbf{h}^H\mathbf{s}(n) + \mathbf{h}^H\mathbf{v}(n), \quad (16)$$

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

$$\sigma_{x,\text{nr}}^2 = \mathbb{E}\{(\mathbf{h}^H\mathbf{x}(n))^2\} = \mathbf{h}^H\mathbf{R}_x\mathbf{h}, \quad (17)$$

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

$$\mathbf{R}_x = \mathbb{E}\{\mathbf{x}(n)\mathbf{x}^H(n)\}. \quad (18)$$

Similar definitions of the variance after noise reduction and the covariance matrix hold for the desired signal and the noise signal. Further, using the signal model in (11), the covariance matrix of the desired signal can be expressed as

$$\mathbf{R}_s = \mathbb{E}\{\mathbf{s}(n)\mathbf{s}^H(n)\} \quad (19)$$
$$= \mathbb{E}\left\{(\mathbf{Z}\mathbf{a})\,(\mathbf{Z}\mathbf{a})^H\right\} \quad (20)$$
$$= \mathbf{Z}\mathbf{P}\mathbf{Z}^H, \quad (21)$$

where

$$\mathbf{P} = \mathbb{E}\{\mathbf{a}\mathbf{a}^H\}. \quad (22)$$

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

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

$$\mathbf{R}_x = \mathbf{R}_s + \mathbf{R}_v. \quad (23)$$

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

$$\text{oSNR}(\mathbf{h}) = \frac{\sigma_{s,\text{nr}}^2}{\sigma_{v,\text{nr}}^2} \quad (24)$$
$$= \frac{\mathbf{h}^H\mathbf{R}_s\mathbf{h}}{\mathbf{h}^H\mathbf{R}_v\mathbf{h}}. \quad (25)$$

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

which is the ratio between the variance of the desired signal before and after noise reduction:

$$\xi_{\text{sr}}(\mathbf{h}) = \frac{\sigma_s^2}{\sigma_{s,\text{nr}}^2} \tag{26}$$

$$= \frac{\sigma_s^2}{\mathbf{h}^H \mathbf{R}_s \mathbf{h}}. \tag{27}$$

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

III. Filters

A. Traditional filters

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

$$\begin{aligned} e(n) &= s(n) - \hat{s}(n) = s(n) - \mathbf{h}^H \mathbf{x}(n) \\ &= s(n) - \mathbf{h}^H \mathbf{s}(n) - \mathbf{h}^H \mathbf{v}(n). \end{aligned} \tag{28}$$

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

$$J(\mathbf{h}) = \mathbb{E}\{e(n)^2\} = \mathbb{E}\{(s(n) - \mathbf{h}^H \mathbf{x}(n))^2\} \tag{29}$$

$$= \mathbb{E}\{ \left(s(n) - \mathbf{h}^H \mathbf{s}(n) - \mathbf{h}^H \mathbf{v}(n)\right)^2 \} \tag{30}$$

Minimisation of $J(\mathbf{h})$ leads to the classical Wiener filter [15]:

$$\mathbf{h}_w = \mathbf{R}_x^{-1} \mathbf{R}_s \mathbf{i}_M, \tag{31}$$

where $\mathbf{i}_M$ is the first column of the $M \times M$ identity matrix. Using (23), the Wiener filter can be rewritten as

$$\mathbf{h}_w = \mathbf{R}_x^{-1} (\mathbf{R}_x - \mathbf{R}_v) \mathbf{i}_M, \tag{32}$$

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

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

$$e_s(n) = s(n) - \mathbf{h}^H \mathbf{s}(n), \tag{33}$$

$$e_v(n) = \mathbf{h}^H \mathbf{v}(n), \tag{34}$$

with the corresponding minimum mean squared errors (MSEs) being

$$J_s(\mathbf{h}) = \mathbb{E}\{e_s(n)^2\} = \mathbb{E}\{(s(n) - \mathbf{h}^H \mathbf{s}(n))^2\} \tag{35}$$

$$J_v(\mathbf{h}) = \mathbb{E}\{e_v(n)^2\} = \mathbb{E}\{(\mathbf{h}^H \mathbf{v}(n))^2\}. \tag{36}$$

These error measures make it possible to, e.g., minimise the noise power output of the filter while constraining the amount of signal distortion the filter introduces [26], i.e.,

$$\min_{\mathbf{h}} J_v(\mathbf{h}) \quad \text{s.t.} \quad J_s(\mathbf{h}) = \beta \sigma_s^2, \tag{37}$$

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

$$\mathbf{h}_\lambda = \left(\mathbf{R}_s + \frac{1}{\lambda} \mathbf{R}_v \right)^{-1} \mathbf{R}_s \mathbf{i}_M, \tag{38}$$

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

B. Parametric filters

The filter in (38) has no control over the distortion of the single harmonics in a voiced speech signal. This is, however, possible by minimisation of $J_v(\mathbf{h})$ under the constraint that the desired signal is passed undistorted, i.e.,

$$\min_{\mathbf{h}} J_v(\mathbf{h}) \quad \text{s.t.} \quad s(n) - \mathbf{h}^H \mathbf{s}(n) = 0. \tag{39}$$

Expressing the signal using the harmonic chirp model in (11), the restriction can be rewritten as

$$s(n) - \mathbf{h}^H \mathbf{s}(n) = 0 \Leftrightarrow \tag{40}$$

$$\mathbf{i}_M^T \mathbf{Z} \mathbf{a} - \mathbf{h}^H \mathbf{Z} \mathbf{a} = 0 \Leftrightarrow \tag{41}$$

$$\mathbf{i}_M^T \mathbf{Z} = \mathbf{h}^H \mathbf{Z} \Leftrightarrow \tag{42}$$

$$\mathbf{b}^H = \mathbf{h}^H \mathbf{Z}, \tag{43}$$

where $\mathbf{b}^H = \mathbf{i}_M^T \mathbf{Z}$ is an $1 \times L$ vector containing the constraints of each harmonic. Using the relation in (17), (39) can be rewritten as

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

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

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

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

The solution to (44) is the linearly constrained minimum variance (LCMV) filter and is given by [22]:

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

The filter reduces to the LCMV filter for harmonic signals when $k = 0$. The covariance matrix of the noise signal is not known and has to be estimated. This is not trivial, but in an optimal situation where the signal model fits perfect, the noise covariance matrix can be replaced by the covariance matrix of the observed signal, $\mathbf{R}_x$, [17], which is easier to estimate, i.e.,

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

Another more empirical approach taking its starting point in the MSE is the amplitude and phase estimation (APES) filter [18]. Here, the harmonic chirp model is also assumed and the

expectation is approximated by an average over time, leading to the estimated MSE:

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

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

where

$$\mathbf{w}(n) = \begin{bmatrix} z(n)^1 & z(n)^{-1} & \cdots & z(n)^L & z(n)^{-L}\end{bmatrix}^T. \quad (50)$$

Writing out the terms in the quadratic expression and solving for the amplitudes [18] gives $\widehat{\mathbf{a}} = \mathbf{W}^{-1}\mathbf{G}\mathbf{h}$, and, thereby,

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

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

$$= \mathbf{h}^H\mathbf{Q}\mathbf{h} \quad (53)$$

with

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

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

and

$$\mathbf{Q} = \mathbf{R}_x - \mathbf{G}^H\mathbf{W}^{-1}\mathbf{G}. \quad (56)$$

As with the LCMV filter, the MSE is minimised with a constraint that the desired signal should be passed undistorted, leading to a similar filter [18]:

$$\mathbf{h}_{\text{APES}} = \mathbf{Q}^{-1}\mathbf{Z}(\mathbf{Z}^H\mathbf{Q}^{-1}\mathbf{Z})^{-1}\mathbf{b} \quad (57)$$

IV. Covariance matrix estimates

The covariance matrices used in the derived filters are not known but have to be estimated. The covariance matrix of the observed signal can, e.g., be estimated by use of the sample covariance matrix estimate [22]:

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

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

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

The APES based noise covariance matrix estimate is obtained using a signal driven approach. Alternatively, we suggest taking a noise driven approach and estimate the noise covariance matrix based on noise PSDs. This can be advantageous since several methods exist for estimating the noise power spectral density in the frequency domain, e.g., based on minimum statistics [7] or MMSE [8]. The power spectral density of a signal $g(n)$, $S_g(\omega)$, is related to the autocorrelation, $R_g(\tau)$, and, thereby, also to the covariance matrix of a signal through the Fourier transform [27]

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

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

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

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

The noise covariance matrix is then estimated as:

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

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

V. Performance of parametric filters

The theoretical performance of the LCMV filter in (46) can be found by inserting the expression for the filter in (25) and (27). Moreover, the expression for the covariance matrix of the desired signal introduced in (21) is used. The output power of the desired signal can be expressed as:

$$\mathbf{h}^H\mathbf{R}_s\mathbf{h} = \mathbf{b}^H(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z}\mathbf{P}\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z}(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})^{-1}\mathbf{b} = \mathbf{b}^H\mathbf{P}\mathbf{b} = \mathbf{1}^T\mathbf{P}\mathbf{1} = \sigma_s^2, \quad (62)$$

where $\mathbf{1}$ is a an $L\times 1$ vector of ones. The second last equality sign follows from the facts that $\mathbf{b}$ contains only unit amplitude exponential functions and that $\mathbf{P}$ is a diagonal matrix. The output power of the noise is:

$$\mathbf{h}^H\mathbf{R}_v\mathbf{h} = \mathbf{b}^H(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{R}_v\mathbf{R}_v^{-1}\mathbf{Z}(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})^{-1}\mathbf{b} = \mathbf{b}^H(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})^{-1}\mathbf{b}. \quad (63)$$

The output SNR and signal reduction factor then becomes:

$$\text{oSNR}(\mathbf{h}) = \frac{\sigma_s^2}{\mathbf{b}^H(\mathbf{Z}^H\mathbf{R}_v^{-1}\mathbf{Z})^{-1}\mathbf{b}}, \quad (64)$$

and

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

These expressions for output SNR and signal reduction are made under the assumption that the noise statistics and the

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

VI. Experiments

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

A. Synthetic signal

1) Setup: The LCMV and APES filters based on the harmonic chirp model were tested on a synthetic chirp signal made according to (5) with the same length as the segment length, N. The signal was generated with $L = 10$, $A_l = 1 \, \forall \, l$, random phase, fundamental frequency, and fundamental chirp rate, in the intervals $\phi_l \in [0, 2\pi]$, $f_0 \in [150, 250]$ Hz, $k \in [0, 200]$ Hz^2, respectively. The signal is sampled at 8 kHz and added to white Gaussian noise with a variance calculated to fit the desired input SNR.

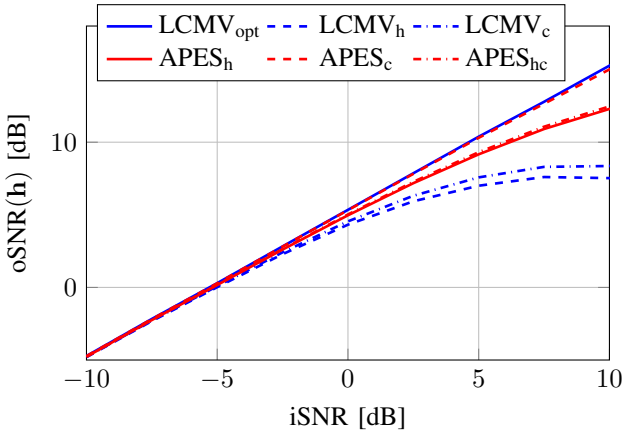
The filters are evaluated as a function of the input SNR, the segment length, N, and the filter length, M. When the parameters are not varied they are set to: iSNR $= 10$ dB, $N = 230$ and $M = 50$. Evaluating M with a fixed N makes it possible to have more elements in the sum in (58) when M is small compared to large, and, thereby, the statistical stability of $\mathbf{R}_x$ would be greater for shorter filters. To avoid this bias and make the conditions as similar as possible for all filter lengths, the same number of elements are used in the sum in (58) independent of the filter length. The fundamental frequency and fundamental chirp rate are assumed known when designing the filters for the synthetic signals. This assumption is made to evaluate the filters without the influence of the performance of a specific parameter estimation method. The results are averaged over 1000 Monte Carlo simulations (MCS). The filters are compared by means of the output SNR in (25) and the signal reduction factor in (27). Using these expressions, the output SNR and signal reduction factor are calculated sample wise based on the N samples used to generate the covariance matrix estimates $\mathbf{R}_s$ and $\mathbf{R}_v$, and afterwards they are averaged over the 1000 MCS.

2) Compared filters: The performance of the chirp based filters is compared to the same filter types based on the harmonic model. This will show whether it is beneficial to expand the traditional harmonic model based on the assumption of stationary speech to a harmonic chirp model where the fundamental frequency is assumed to change linearly within each segment. The LCMV and APES filters derived for the harmonic model can be obtained by setting $k = 0$ in the signal model. A set of six filters are compared in the simulations:

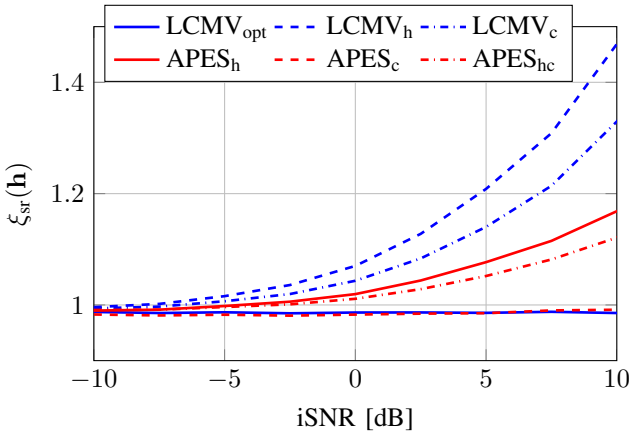
- **LCMV$_{\text{opt}}$**: chirp LCMV filter made according to (46) with $\mathbf{R}_v$ estimated from the clean noise signal. This filter will have the best possible performance a harmonic chirp LCMV filter can have, but can not be made in practice since there is no access to the clean noise signal.
- **LCMV$_{\text{h}}$**: harmonic LCMV filter made according to (47) with $k = 0$.
- **LCMV$_{\text{c}}$**: chirp LCMV filter made according to (47).
- **APES$_{\text{h}}$**: harmonic APES filter made according to (57) with $k = 0$.
- **APES$_{\text{c}}$**: chirp APES filter made according to (57).
- **APES$_{\text{hc}}$**: APES filter made as a combination of the chirp and normal harmonic model with $\mathbf{Z}$ based on the chirp model whereas the estimation of $\mathbf{Q}$ is based on the normal harmonic model. This filter is included to separate the contribution from the modified $\mathbf{Z}$ vector and the modified $\mathbf{Q}$ matrix.

3) Evaluation: The output SNR and signal reduction factor as a function of the input SNR are shown in Fig. 1. At an input SNR of -10 dB all filters perform equally well, but as the input SNR is increased the difference in performance between the filters is increased. As expected, the LCMV$_{\text{opt}}$ sets an upper bound for the performance with a similar gain in SNR at all considered levels of input SNR and no distortion of the desired signal. The harmonic chirp APES based filter, APES$_{\text{c}}$, has similar performance to the optimal LCMV filter. The difference between the two filters, APES$_{\text{h}}$ and APES$_{\text{hc}}$, is only minor. They deviate from the LCMV$_{\text{opt}}$ around 0 dB input SNR and at an input SNR of 10 dB the gain in SNR is around 3 dB less than for the optimal LCMV filter. They also introduce some distorion of the desired signal, with APES$_{\text{h}}$ distorting the desired signal slightly more than APES$_{\text{hc}}$. These two filters have the same noise covariance matrix estimate but different versions of the $\mathbf{Z}$ matrix, as is also the case for the two LCMV filters, LCMV$_{\text{h}}$ and LCMV$_{\text{c}}$, based on the covariance matrix of the observed signal. LCMV$_{\text{h}}$ and LCMV$_{\text{c}}$ have the worst performance of the compared filters, but show the same tendencies as APES$_{\text{h}}$ and APES$_{\text{hc}}$. The difference between the two filters is mainly a smaller signal distortion for the chirp based filter, but here also with a slight difference

in the output SNRs of the two filters. This shows, at least for relatively short filter lengths of $M = 50$, that the major change in performance comes from changing the covariance matrix, from the covariance matrix of the observed signal to the harmonic APES covariance matrix and further again to the harmonic chirp APES covariance matrix. Changing $\mathbf{Z}$ has a minor role but still has an influence, primarily with respect to the distortion of the desired signal.



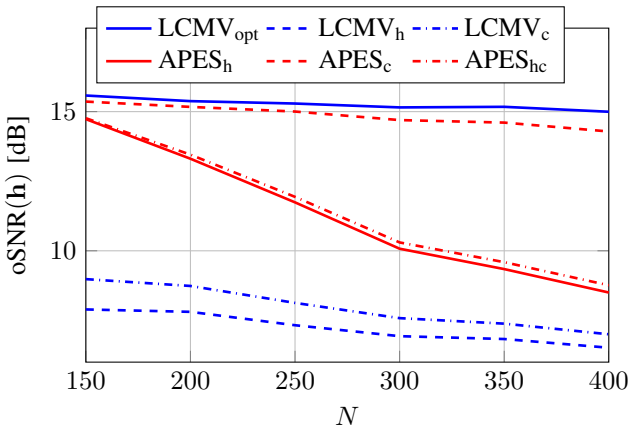
(a) Output SNR



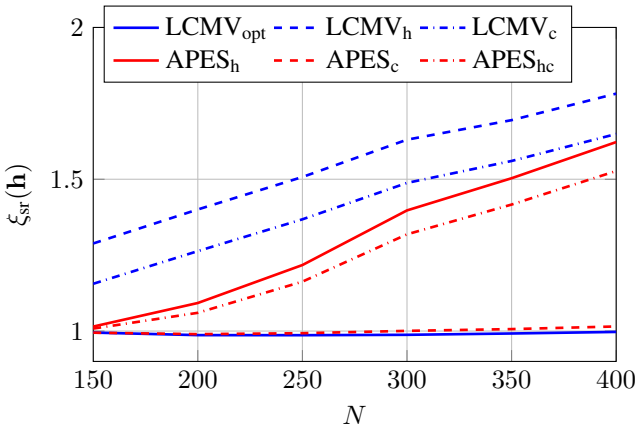
(b) Signal reduction factor

Fig. 1: Output SNR (a) and signal reduction factor (b) as a function of the input SNR for synthetic chirp signals.

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



(a) Output SNR

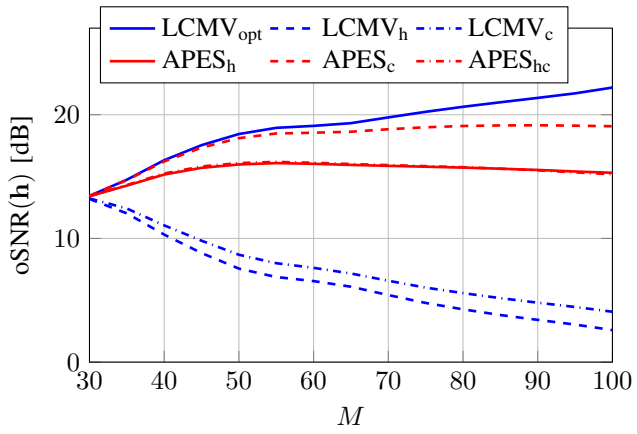


(b) Signal reduction factor

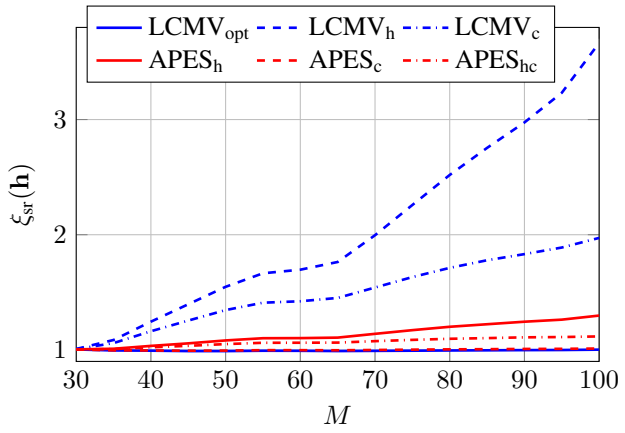
Fig. 2: Output SNR (a) and signal reduction factor (b) as a function of the number of samples N for synthetic chirp signals.

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

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



(a) Output SNR



(b) Signal reduction factor

Fig. 3: Output SNR (a) and signal reduction factor (b) as a function of the filter length M for synthetic chirp signals.

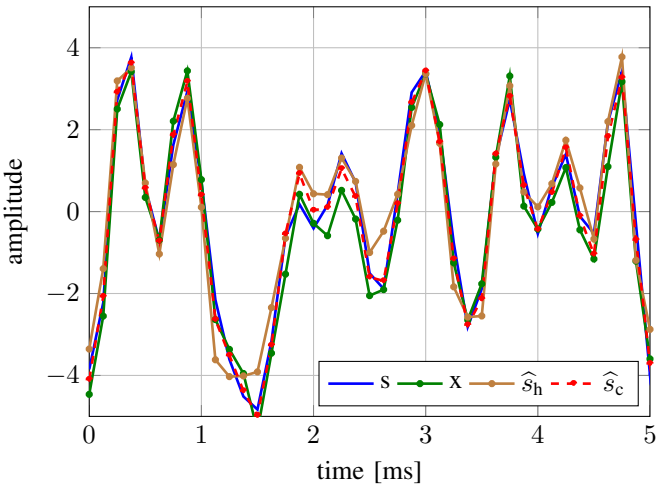


Fig. 4: Reconstructed signal using APES$_h$ and APES$_c$ filters compared to the clean and noisy signals. The noise is white Gaussian and the input SNR is 10 dB.

is.

B. Speech signals

1) Setup: The speech signals used are the 30 sentences included in the NOIZEUS database [28]. Three male and three female speakers produced the 30 Harvard sentences contained in the database. The signals are sampled at 8 kHz and corrupted with noise from the AURORA database [29]. In the first part of this evaluation of speech signals, where the chirp model is compared to the harmonic model, the parameters of the speech signals are estimated from the clean signals. This is done to be able to compare the results for speech signals with the simulations on synthetic data where the parameters were assumed known. In the second part, where the LCMV and Wiener filters are compared, results based on parameters estimated from the noisy signals are shown. The model order and a preliminary fundamental frequency are estimated for every 50 samples using a nonlinear least squares (NLS) estimator [22] with the lower and upper limit for the fundamental frequency given by 80 Hz and 400 Hz, respectively. This is followed by a smoothing [30] and joint estimation of the fundamental frequency and chirp parameter for each sample using the iterative NLS estimator described in [20]. Since the filters are independent of one another, and the fundamental frequency and chirp rate are estimated with reference to the sample being estimated, $\mathbf{Z}$ is also defined with reference to this sample, i.e., the time index in $\mathbf{Z}$ is going from 0 to $M+1$ in each filter. The filter length is increased to $M = 70$ because the real speech signals in many frames have more harmonics than the 10 used to create the synthetic signals. Therefore, a filter with more degrees of freedom is preferred. A good compromise between filter length and segment length for the LCMV and APES filters would according to [31] be $N = 4M$, but this would lead to quite long segments with the given filter length and, as a compromise, the segment length is again set to $N = 230$. The voiced periods are picked out using a generalised likelihood ratio test [32], [33]. Alternatively, the MAP criteria [22] or other voiced/unvoiced detectors can be used [13], [14]. In some cases where unvoiced speech is mistakenly assigned as voiced, the filters become numerically unstable, and these samples are, therefore, excluded from the evaluation. If the filter is not unstable, the unvoiced speech assigned as voiced is processed as if it was voiced speech. This is expected to give a slight decrease in the performance since it is not possible to obtain noise reduction without signal distortion when using the harmonic model in periods of unvoiced speech. In the first part, where the LCMV filters are compared, white Gaussian noise is used and the output SNR and signal reduction factor are calculated using (25) and (27) to facilitate the comparison with the results for the synthetic signal. As was the case for the synthetic signals, the performance measures are calculated sample wise and afterwards averaged over the entire speech signal and the NOIZEUS speech corpus. When the LCMV and APES filters are compared to the Wiener filter, babble noise is used, where the noisy signals are taken from the NOIZEUS speech corpus. The noise levels in the NOIZEUS speech corpus range from 0 dB to 15 dB. The babble noise is chosen because it is one of the most difficult noise types

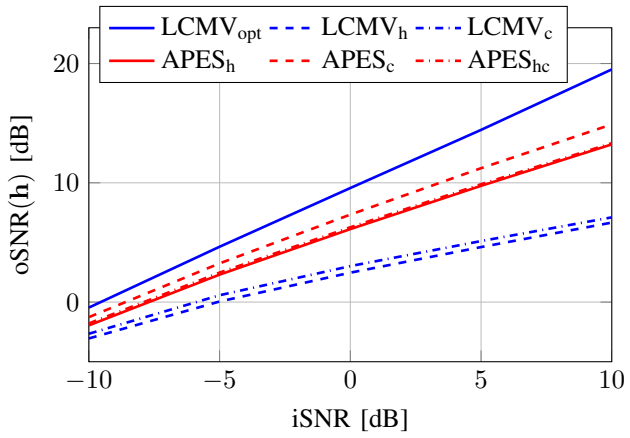
to remove. Results are shown both when the parameters are estimated from the clean signal and when the parameters are estimated from the noisy signals. Since the filters are made based on different ways to estimate the covariance matrices the filters are here compared by means of the output SNR in (24) and the signal reduction factor in (26). Before calculating the variance, the voiced speech parts have been concatenated. This way there will only be one value of the output SNR and signal reduction factor per speech signal which is then averaged over the speech corpus.

2) Compared filters: In the first part of the simulations with real speech, the same filters used for the synthetic signals are compared. In the second part, the LCMV and APES based filters are compared to the Wiener filter. This is done for two different choices of covariance matrices, the first one using the APES derivation, the other using (61) based on the MMSE criterion [8] for finding the PSD. Filters based on the PSD using MMSE and minimum statistics perform almost equally well, and, therefore, only one type of these filters is shown. Further, flexible Wiener filters with two different values of λ are included in the comparisons, leading to six filters:

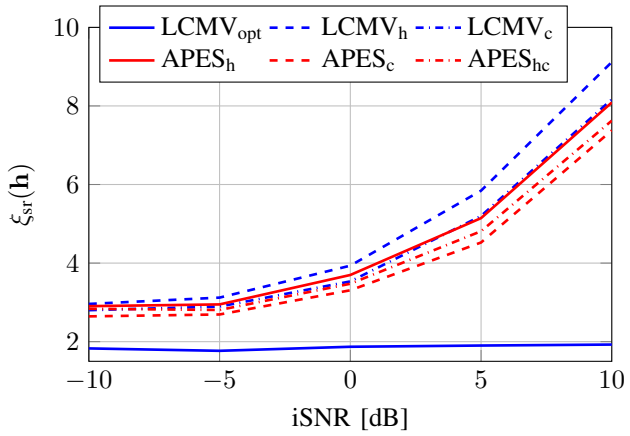
- $\mathbf{APES_c}$: chirp APES filter made according to (57).
- $\mathbf{LCMV_{MMSE}}$: chirp LCMV filter made according to (46) with $\mathbf{R}_v$ estimated from (61) using MMSE.
- $\mathbf{W_c}$: Wiener filter made according to (31) with $\mathbf{R}_s$ estimated using the APES principle as $\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G}$.
- $\mathbf{W_{MMSE}}$: Wiener filter made according to (32) with $\mathbf{R}_v$ estimated from (61) using MMSE.
- $\mathbf{W}_{\lambda=0.2}$: Trade-off Wiener filter made according to (38) with $\lambda = 0.2$ and $\mathbf{R}_s$ estimated using the APES principle as $\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G}$.
- $\mathbf{W}_{\lambda=5}$: Trade-off Wiener filter made according to (38) with $\lambda = 5$ and $\mathbf{R}_s$ estimated using the APES principle as $\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G}$.

Note that all filters except $\mathbf{W_{MMSE}}$ are in some way based on the harmonic chirp model. The $\mathbf{APES_c}$ through both the modified Fourier vector $\mathbf{Z}$ and the covariance matrix estimate $\mathbf{Q}$. The LCMV through $\mathbf{Z}$ and the three Wiener filters $\mathrm{W_c}$, $\mathrm{W}_{\lambda=0.2}$ and $\mathrm{W}_{\lambda=5}$ through the used covariance matrix.

3) Evaluation: In Fig. 5, the output SNR and signal reduction factor are shown as a function of the input SNR. The output SNR and signal reduction factor are calculated using (25) and (27) as was also the case for the synthetic signals. It is seen that the tendencies are the same as for the synthetic signal. $\mathrm{APES_c}$ does not follow the optimal LCMV filter as closely as it did for the synthetic signal, but this is not surprising since the synthetic signals were made according to the harmonic chirp model, and the parameters were assumed known. For the speech signals, the parameters are estimated, and the model does not fit perfectly since the fundamental frequency will not be completely linear in any considered piece within a speech signal. Even though the performance of the $\mathrm{APES_c}$ filter deviates more from the optimal LCMV filter than it did considering synthetic signals, it still has a better performance than the other considered filters. This means that the harmonic chirp model is better at describing the voiced parts of a speech signal and increased performance can be obtained by replacing the traditional harmonic filters with chirp filters.



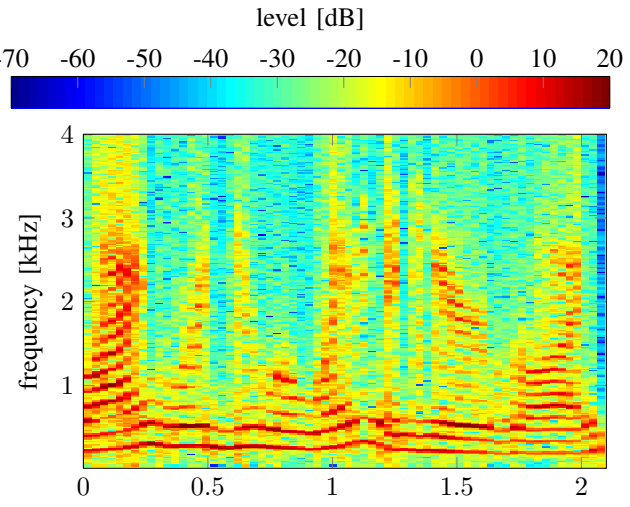
(a) Output SNR



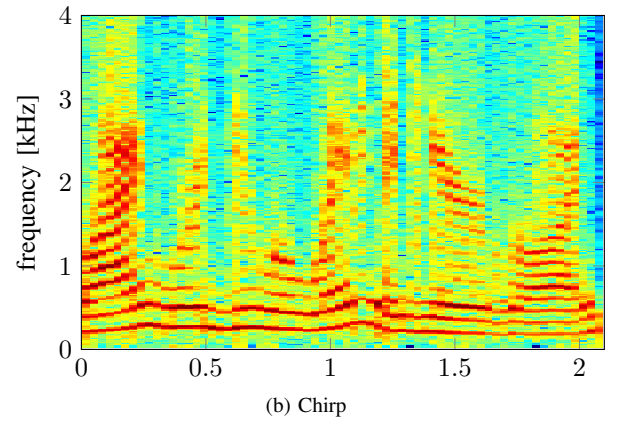
(b) Signal reduction factor

Fig. 5: Output SNR (a) and signal reduction factor (b) as a function of the input SNR, average over NOIZEUS speech corpus added white noise. Parameters estimated from clean speech signals.

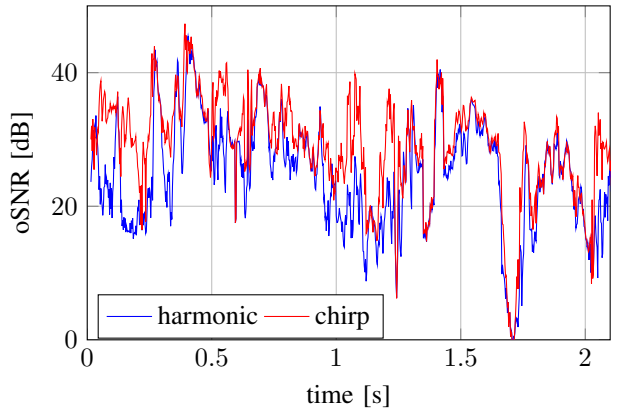
As an example, the speech signal 'Why were you away a year, Roy?' uttered by a female speaker is filtered. The signal has the advantage that it only contains voiced speech, and the entire signal can, therefore, be filtered by the proposed methods. The signal is sampled at 8 kHz, the segment length is 230, the filter length is 70, and the parameters are estimated in the same way as the previous speech signals. The noise is white Gaussian and the input SNR is 10 dB. The spectrograms of the filtered speech signal using $\mathrm{APES_h}$ and $\mathrm{APES_c}$ are shown in Fig. 6 together with the output SNR over time. It is seen that the output SNR of the chirp filter is larger or equal to the output SNR of the harmonic filter. The difference is most pronounced in the first 0.25 seconds and between 1 and 1.25 seconds where the fundamental frequency is changing the most. Here, it is also seen in the spectrograms that the harmonics look slightly cleaner when the chirp filter is used. The Perceptual Evaluation of Speech Quality (PESQ) score [34] for the speech filtered with the harmonic filter is 2.21



(a) Harmonic



(b) Chirp



(c) Output SNR over time

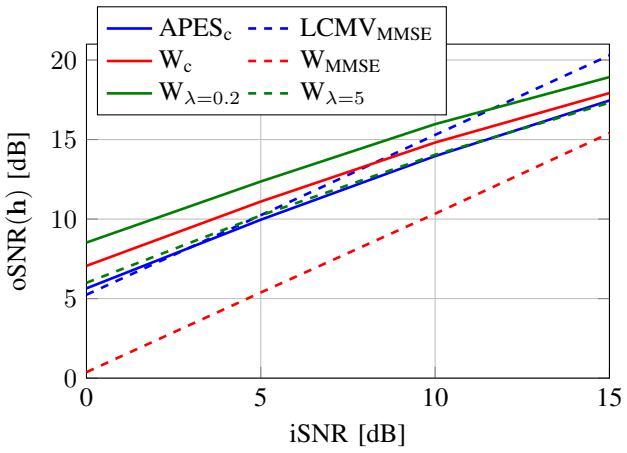
Fig. 6: Spectrograms of speech signal after filtering with (a) traditional harmonic filter and (b) harmonic chirp filter. In (c) the output SNR over time is shown. The input SNR is 10 dB and the noise is white Gaussian. The clean signal can be seen in Fig. 10.

whereas the chirp filter gives a PESQ score of 2.32 and the noisy signal gives a PESQ score of 1.57. The speech signals related to this comparison and the comparison in Fig. 10 can be found at http://www.create.aau.dk/smn.

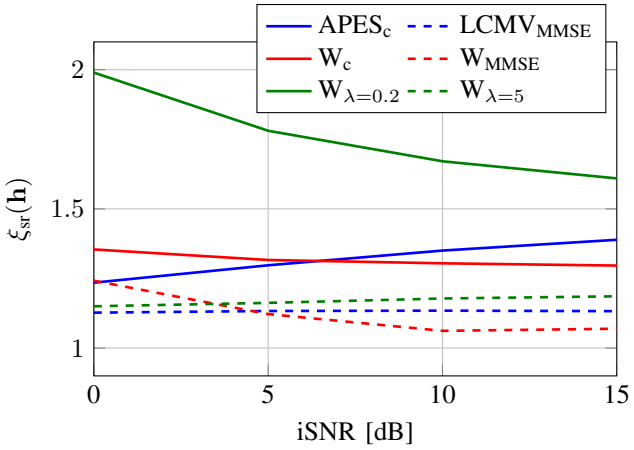
The increased performance of the harmonic chirp filters relative to the harmonic filters should of course be viewed in light of an increased computational complexity since the joint estimation of the fundamental frequency and chirp rate is based on a search in a two-dimensional space. However, [20] describes how to find the parameters iteratively which will decrease the complexity relative to a two-dimensional grid search, and the initial fundamental frequency estimate used in the algorithm is only estimated for every 50 samples in this work which seems to be sufficient for giving good estimates.

Now we turn to alternative combinations of filters and covariance matrices. Here, the output SNR and signal reduction factor are calculated according to (24) and (26). This ensures that no filter is favoured in the way the performance is calculated since the covariance matrices based on the sample covariance principle and the PSD are made in two fundamentally different ways. In Fig. 7a it is seen that five of the six filters work very similar. The Wiener filter in combination with the PSD noise covariance matrix perform significantly worse than the rest when it comes to output SNR. However, the PSD covariance matrix works quite well in combination with the LCMV filter. This filter is one of the better filters at higher input SNRs with respect to output SNR, and it has a low level of distortion at all input SNRs as is seen in Fig. 7b. This can probably be explained by looking at the filters in (31) and (46). The Wiener filter is dependent on two covariance matrices, and the relative levels of these two matrices are, therefore, important for the look of the filter. The LCMV based filters are only dependent on one covariance matrix, and in some way the denominator of the LCMV can be seen as a normalisation which makes the filter independent of the absolute size of the covariance matrix used. The trade-off Wiener filter with $\lambda = 0.2$ gives a higher output SNR than the Wiener filter but at the same time it also gives rise to a higher signal distortion. The flexible Wiener filter with $\lambda = 5.0$ works in the opposite way. It gives a lower output SNR, but also a lower degree of signal distortion. In Fig. 8, the parameters are estimated from the noisy signals whereas the voiced/unvoiced detection is based on the clean signal. The output SNR for the signal dependent filters is decreased a few dBs at low input SNRs whereas it is very similar at high input SNRs. This makes sense since the estimation of parameters is more difficult at low SNRs than at high SNRs. The Wiener filter dependent on the PSD has the same performance in the two situations. In Fig. 9, also the voiced/unvoiced detection is made based on the noisy signal. The overall performance of all filters is slightly decreased compared to making the detection based on the clean signal, but the tendency between the filters is very similar. This suggests that more unvoiced periods are assigned as voiced speech where the voiced signal model will not apply, and thus the performance will decrease slightly.

As an example, the speech signal 'Why were you away a year, Roy?' is again filtered, now in the presence of babble noise at an input SNR of 10 dB. The filters used for this comparison are the APES_c, $\text{LCMV}_{\text{MMSE}}$, W_c and W_{MMSE} and the spectrograms of the resulting signals are shown in Fig. 10 along with spectrograms of the clean and the noisy signal. From this figure, it seems like the Wiener filter in combination with the APES covariance matrix removes the

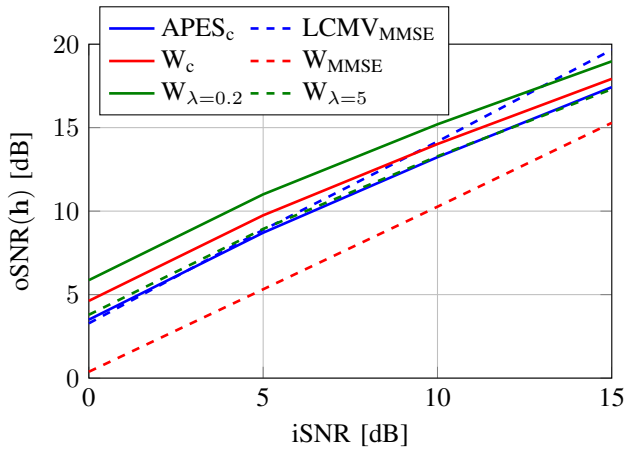


(a) Output SNR

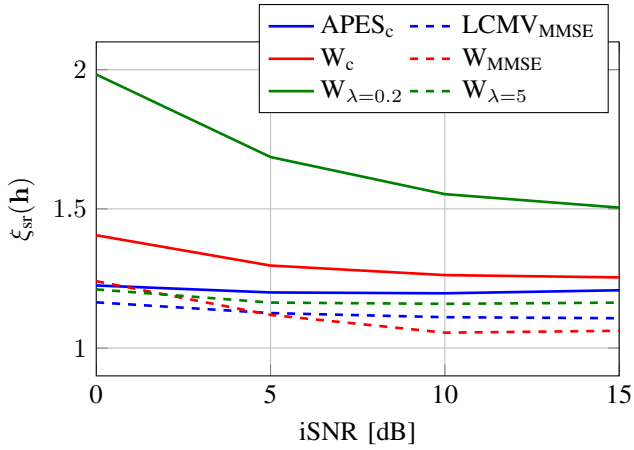


(b) Signal reduction factor

Fig. 7: Output SNR (a) and signal reduction factor (b) as a function of the input SNR, averaged over NOIZEUS corpus with babble noise. Parameters estimated from clean speech signals. Voiced/unvoiced detection based on clean signal.



(a) Output SNR



(b) Signal reduction factor

Fig. 8: Output SNR (a) and signal reduction factor (b) as a function of the input SNR, averaged over NOIZEUS corpus with babble noise. Parameters estimated from noisy speech signals. Voiced/unvoiced detection based on clean signal.

most noise between the harmonics whereas the APES filter and the LCMV filter remove the noise slightly less, both between the harmonics and outside the range of the speech signal. The Wiener filter in combination with the PSD noise covariance matrix seems to perform no noise reduction and the harmonics are even more difficult to distinguish than in the noisy signal. These observations are in line with the curves of output SNR when looking at an input SNR of 10 dB where the W_{MMSE} performs worse than the noisy signal, the $APES_c$ and $LCMV_{MMSE}$ perform almost equally well and the W_c performs the best. The PESQ scores for the four filtered signals are, $APES_c$: 2.09, $LCMV_{MMSE}$: 2.25, W_c: 2.18 and W_{MMSE}: 1.54. It is interesting to see that the $LCMV_{MMSE}$ gives rise to the highest PESQ score since this was not clear from the spectrograms, but this filter gives a lower signal reduction factor than the $APES_c$ 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. 6, the PESQ score of the filtered signals decreased whereas the PESQ score of the noisy signal increased. This difference is mainly due to the different noise types while the fact that the parameters in Fig. 6 were estimated from the clean signal only contributes slightly. Since babble noise is noise made up from several speakers speaking at the same time, it is distributed in the same frequency range as the speech signal. This makes it more difficult to estimate the relevant parameters and also more difficult to filter out the noise afterwards. However, prewhitening of the noisy signal can help mediate this problem [35] with the noise statistics found using one of the methods in [36].

VII. Conclusion

In this paper, the non-stationarity of voiced speech is taken into account in speech enhancement. This is done by describing the speech by a harmonic chirp model instead of the traditional harmonic model. The chirp used is a linear chirp which allows the fundamental frequency to vary linearly within each segment, and, therefore, the speech signal is not

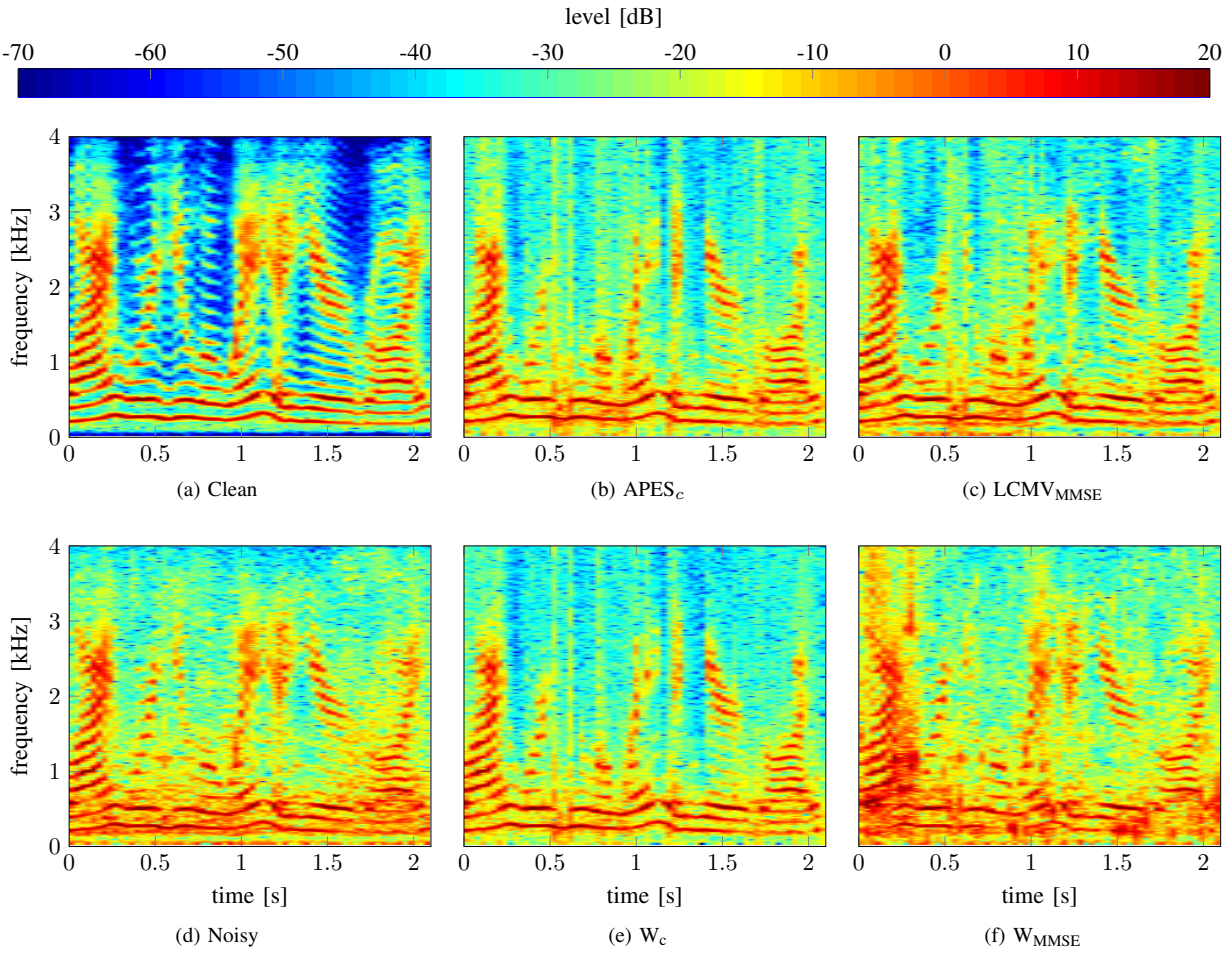


Fig. 10: Spectrograms of clean, noisy and filtered speech. Babble noise is mixed with speech at an input SNR of 10 dB.

assumed stationary within a segment. Versions of the linearly constraint minimum variance (LCMV) filter and amplitude and phase estimation (APES) filter are derived in the framework of harmonic chirp signals. As an implicit part of the APES filter, a noise covariance matrix estimate is derived. This makes the APES filter an estimate of the optimal LCMV filter which maximises the output SNR under the constraint that the desired signal is passed undistorted. APES gives a noise covariance matrix estimate which only assumes the noise signal to be stationary in frames of 20-30 ms as opposed to methods based on power spectral densities (PSDs) which primarily update the noise statistics in periods of unvoiced speech. It is shown through simulations on synthetic and speech signals that the chirp filters give rise to a higher output SNR and a lower signal distortion than their harmonic counterparts, and, therefore, the chirp model describes voiced speech better than the traditional harmonic model. We suggest also using the APES noise covariance matrix estimate in other filters as, e.g., the Wiener filter, and we compare it to a noise covariance matrix estimate based on the PSD. The APES noise covariance matrix estimate is shown to work well in combination with the Wiener and trade-off Wiener filters, whereas the PSD based noise covariance matrix estimate works well in combination with the LCMV filter. All chirp based Wiener and LCMV filters outperform the Wiener filter in combination with the PSD noise covariance matrix estimate.

REFERENCES

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

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

[3] E. W. Healy, S. E. Yoho, J. Chen, Y. Wang, and D. Wang, "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," *J. Acoust. Soc. Am.*, vol. 138, no. 3, pp. 1660–1669, 2015.

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

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

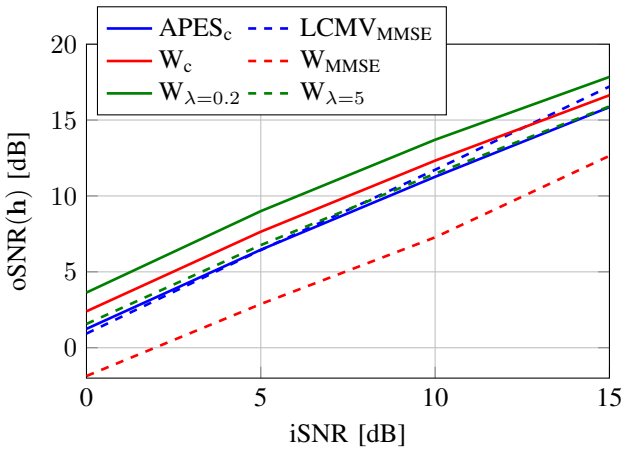
[6] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, Jan. 1999.

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

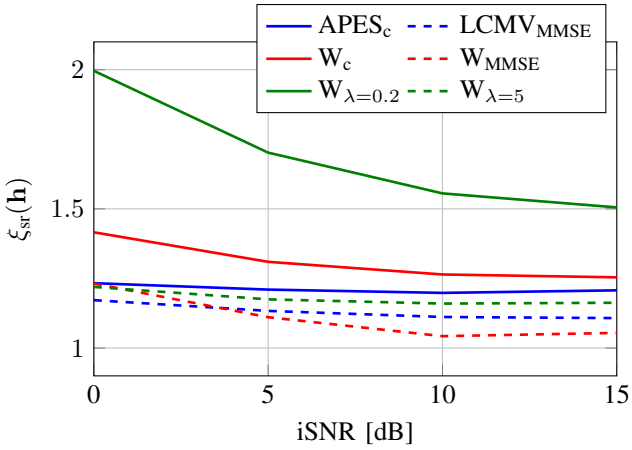
[8] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

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

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



(a) Output SNR



(b) Signal reduction factor

Fig. 9: Output SNR (a) and signal reduction factor (b) as a function of the input SNR, averaged over NOIZEUS corpus with babble noise. Parameters estimated from noisy speech signals. Voiced/unvoiced detection based on noisy signal.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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