Carl von Ossietzky
University of Oldenburg

Bachelor of Engineering Programme
Engineering Physics

Bachelor Thesis

Single-Channel Noise Reduction Using Speech-Distortion Weighted Inter-Frame Wiener Filters

Submitted by
Klaus Brümann

Supervisor
Prof. Dr. Simon Doclo

Co-Supervisor
MSc. Dörte Fischer

Oldenburg, 19th November, 2018

Abstract

Most single-channel speech enhancement algorithms in the short-time Fourier transform (STFT) domain assume that neighbouring speech STFT coefficients are uncorrelated over time and frequency. Consequently, a commonly used approach is to apply a (real-valued) Wiener gain (WG) to each noisy STFT coefficient independently. Alternatively, exploiting the speech correlation between present and past time frames results in a complex-valued filter which is applied to the noisy STFT coefficients. Several single-channel inter-frame speech enhancement algorithms have already been derived, such as the inter-frame Wiener filter (IFWF) and the inter-frame minimum variance distortionless response (IFMVDR) filter. To provide a trade-off between noise reduction and speech distortion, real-valued and complex-valued speech-distortion weighted (SDW)-IFWFs have been derived, which differ in their assumptions about the speech and noise correlation matrices. In this thesis, the influence of the trade-off parameter is analysed for various implementations of the SDW-IFWFs and SDW-WGs in a low-delay filter bank architecture. The influence of the oversampling-factor which determines the frequency resolution of the real-valued SDW-IFWFs is also evaluated within the low-delay framework. Under blind conditions (only the noisy speech signal available), three implementations of a complex-valued SDW-IFWF are compared with two real-valued SDW-IFWFs using practically feasible estimators for the required quantities. Each evaluation is tested on diverse speech and noise material at different signal-to-noise ratios (SNRs), in terms of objective speech quality and intelligibility measures as well as for noise reduction.

Contents

1 Introduction

Modern audio communication systems such as mobile phones, headsets, and hearing aids are often unavoidably affected by undesired background noise. Interfering speakers or ambient traffic are some examples of undesired additive noise which can severely degrade the quality and intelligibility of speech. The aim in this thesis is to use noise reduction techniques to filter out undesired, additive noise from speech recorded by a single microphone in a noisy environment, to improve both the intelligibility and quality of the speech.

Assuming that the speech and noise components are uncorrelated with eachother, various noise reduction filters can be derived in the short-time Fourier transform (STFT) domain. The basic STFT framework consists of a weighted overlap-add (WOLA) filterbank which decomposes a noisy signal into overlapped, windowed time segments and transforms each time segment into the frequency domain via discrete Fourier transform (DFT). The enhanced speech signal in the time domain can then be obtained by transforming the filtered noisy speech coefficients via inverse STFT (ISTFT).

A common assumption made about speech or noise is that neighbouring time frames and frequency bins are uncorrelated with eachother. One such filter which uses this assumption is the single channel Wiener gain (WG) [1], which treats each time-frequency point as uncorrelated with neighbouring points and applies the filter gains to each time-frequency point independently.

In [2] it is shown that it is more accurate to take into account the so-called inter-frame correlation (IFC) of speech and noise, since speech is highly correlated between consecutive time frames and noise signals are also correlated to some degree. This assumption allows complex-valued filters to be derived, such as inter-frame Wiener filters (IFWFs), inter-frame minimum-variance distortionless response (IFMVDR) filters, and speech-distortion weighted- (SDW)-IFWFs (also known as trade-off filters) [2–6], which are applied to coefficient vectors containing past and present frames of the noisy speech to obtain an estimate of the speech STFT coefficients. The IFC matrix of the noisy speech, which is required in all filters which exploit the IFC, contains the autocorrelation estimates of the noisy speech and can be estimated in a few ways, the most common of which is using a first-order recursive smoothing due to its ease of computation, however, in this thesis, other estimators will also be considered. The theory behind filters which exploit the IFC is equivalent to the theory behind the filters found in multi-channel applications [7], where the systems use the correlation between microphones to estimate the filter coefficients, which are then applied to each channel.

The filters which rely on the autocorrelation each have different characteristics. The IFMVDR filter provides noise reduction with a constraint of preserved desired speech, meaning that the correlated speech components are free of distortion. In contrast, the IFWF performs more filtering of noise with the consequence of some distortion of the desired speech, which may affect both the speech quality and intelligibility. It is known that the IFWF can be decomposed into an IFMVDR filter multiplied by a WG. In [4] it is shown that in practice, an IFMVDR with a WG post filter (IFMVDR+WG) implementation filters out more noise than an IFMVDR and distorts speech less than an IFWF, providing a good compromise between both filters. To balance the speech distortion and noise reduction. Complex-valued SDW-IFWFs can be derived with a parameter which adjusts this trade-off. The idea of the single-channel SDW-IFWF in [2] comes from the SDW- multi-channel Wiener filter (MWF) proposed in [8–10] and in [6] it is shown that a real-valued SDW-IFWF can be derived which can be condensed into scalar WGs. The SDW-IFWF is explored here and extended with two more versions. These filters rely on the speech and either noisy speech or noise power spectral density (PSD). The noise PSD can be estimated using a speech presence probability (SPP) estimator [11–13] and the speech PSD can be estimated using the well known *decision-directed* approach to estimate the *a-priori* signal-to-noise ratio (SNR), followed by a power subtraction.

The FIR filters discussed in this thesis tend to be computationally demanding for high numbers of filter coefficients. This is due to the increasingly complex matrix inversions in filters such as the IFMVDR or IFWF. This leads to computational delay, which is undesired in real-time systems and applications such as communication devices. One solution which bypasses matrix inversions uses real-valued scalar gain filters, such as WGs. While scalar gains may be limited in terms of performance due to having no effect on the phase of the STFT coefficients, they are more computationally efficient even in a filterbank with a higher frequency resolution. A drawback of high-resolution filterbanks, however, is that in general they introduce more delay in analysis-synthesis, making them unsuitable for real-time applications such as hearing-aids. In [14], however, it is shown how to maintain low delay when using longer analysis windows for a higher frequency resolution while still maintaining perfect reconstruction. An asymmetrical analysis window is used, which focuses on more recent frames, and the key to maintaining low delay is using a short synthesis window. All filterbanks used in this thesis use a low-delay architechture.

In this thesis, various forms of speech and noise IFC matrix and vector estimation, as well as implementations of noise reduction filters are discussed and evaluated objectively under oracle conditions, i.e. assuming perfect knowledge of the speech and/or noise IFC coefficients, to see what is the best possible performance which can be achieved. The effects of using different numbers of coefficients and a higher resolution filterbank, on the performance of the SDW-IFWFs, are briefly investigated with the aim of pushing the limits of the filter performance under optimal conditions even further. To conclude the evaluation, the best-performing combinations of estimators and filters are implemented under blind conditions, i.e. with access to only a noisy signal, to see what effect the estimation methods of the speech and noise IFCs have on the filter performance.

This thesis is structured as follows: Section 2 introduces the noise reduction problem and contains the details of the STFT filterbank. Section 3 introduces the WG, IFWF, IFMVDR, SDW-IFWFs, and methods of estimating the IFC matrices and vectors. Section 4 specifies the parameters used in the implementations for the experimental evaluation and contains the results of the objective measures applied to the filtered noisy signals. The conclusions and suggestions for further research are included in Section 5.

2 Problem Statement

A noisy speech signal x recorded by a single microphone can be decomposed as

$$x[n] = s[n] + v[n], \tag{2.1}$$

where s is the clean speech signal, v is the noise signal, and n is the sample index.

Since overlapping frames are desired for a higher resolution of the IFC, the STFT filterbank which is used is a WOLA filterbank. The noisy speech STFT coefficient $X_{k,l}$ at time frame l and frequency bin k can be expressed as the Fourier transform of the windowed time signal $x[n]$

$$X_{k,l} = \sum_{n=-\infty}^{\infty} h^K[-n]x[n + lN]e^{\frac{-2\pi j(n+lN)k}{K}} \tag{2.2}$$

where h^K is the analysis window, N is the frame shift, and $j^2 = -1$. K is the number of subband signals in the K-filterbank with the respective frequency index k given as

$$k = -\frac{K}{2} + 1, -\frac{K}{2} + 2, ..., \frac{K}{2}. \tag{2.3}$$

The speech and noise STFT coefficients $S_{k,l}$ and $V_{k,l}$ can also be found using (2.2). Throughout this thesis, it is assumed that the speech and noise signals are uncorrelated with each other. In Section 2.1, the single-frame signal model is introduced and Section 2.2 extends this model to the multi-frame model with the assumption that speech and noise are correlated across consecutive time frames. In Section 2.3 it is shown how the multi-frame model can be used to derive real-valued, transformed filter coefficient vectors which can be condensed into real-valued scalar filter gains and are applied independently as in the single-frame signal model.

2.1 Single-Frame Signal Model

In the single-frame signal model it is assumed that noise and speech are uncorrelated with each other and across time or frequency. Expressing the noisy speech signal in the STFT domain using (2.2), the noisy speech coefficient $X_{k,l}$ can be decomposed into the speech and noise coefficients, $S_{k,l}$ and $V_{k,l}$, respectively, describing the single-frame signal model as follows

$$X_{k,l} = S_{k,l} + V_{k,l}. \tag{2.4}$$

To obtain an estimate of the speech coefficient $\hat{S}_{k,l}$, a real-valued gain $G_{k,l}$ is applied independently to each noisy speech coefficient as follows

$$\hat{S}_{k,l} = G_{k,l} X_{k,l}. \tag{2.5}$$

To obtain the speech signal estimate $\hat{s}[n]$ in the time-domain, the ISTFT is applied, which consits of applying an inverse DFT (IDFT) to each time frame, followed by a WOLA procedure, where the frames are windowed and overlapped according to their frame-shift. The summarized framework of these filters is depicted in Fig. 2.1.

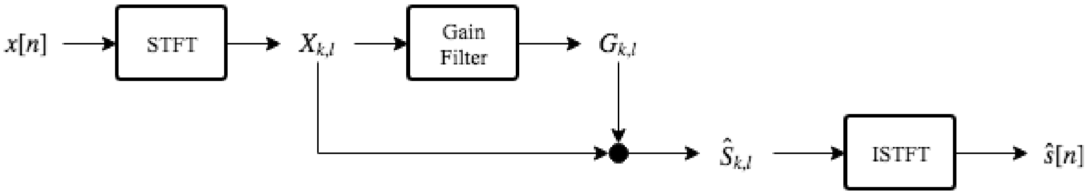

Fig. 2.1. Block diagram of noise reduction using real-valued scalar filter gains.

2.2 Multi-Frame Signal Model

Now, assuming that the speech, noise, and noisy speech signals are correlated over consecutive time frames, it makes sense to observe more than one frame simultaneously when deriving a noise reduction filter. Thus, the present frame and past $M-1$ frames of the noisy speech coefficients $X_{k,l}$ from (2.2) can be stacked in the column-vector $\boldsymbol{x}_{k,l}^{M}$ of length M

$$\boldsymbol{x}_{k,l}^{M} = [X_{k,l}, X_{k,l-1}, ..., X_{k,l-M+1}]^T, \tag{2.6}$$

where ^{T} denotes the transpose operator. The same can be done for the speech and noise to obtain $\boldsymbol{s}_{k,l}^{M}$ and $\boldsymbol{v}_{k,l}^{M}$, respectively, to obtain the multi-frame signal model

$$\boldsymbol{x}_{k,l} = \boldsymbol{s}_{k,l} + \boldsymbol{v}_{k,l} = \boldsymbol{\gamma}_{k,l}^{s} S + \boldsymbol{v}_{k,l} + \boldsymbol{s}'_{k,l}. \tag{2.7}$$

$\boldsymbol{\gamma}_{k,l}^{s}$ is the normalized speech IFC vector which only contains the correlated components of the speech vector $\boldsymbol{s}_{k,l}$ and the uncorrelated speech components are contained in $\boldsymbol{s}'_{k,l}$. The normalized speech IFC vector can be obtained as follows

$$\boldsymbol{\gamma}_{k,l}^{s} = \frac{\mathbf{E}\{\boldsymbol{s}_{k,l} S_{k,l}^{*}\}}{\mathbf{E}\{|S_{k,l}^{2}|\}} = \frac{\mathbf{E}\{\boldsymbol{s}_{k,l} S_{k,l}^{*}\}}{\phi_{k,l}^{s}}, \tag{2.8}$$

where * denotes the complex conjugate and $\phi^s_{k,l}$ is the speech PSD. The noisy speech PSD $\phi^x_{k,l}$ and noise PSD $\phi^v_{k,l}$ can be obtained similarly. The IFC matrices of the speech, noise, and noisy speech, namely $\boldsymbol{R}^s_{k,l}$, $\boldsymbol{R}^v_{k,l}$, and $\boldsymbol{R}^x_{k,l}$, respectively, are defined as follows

$$\boldsymbol{R}^x_{k,l} = \mathbf{E}\{\boldsymbol{x}_{k,l}\boldsymbol{x}^H_{k,l}\}, \tag{2.9}$$

$$\boldsymbol{R}^s_{k,l} = \mathbf{E}\{\boldsymbol{s}_{k,l}\boldsymbol{s}^H_{k,l}\}, \tag{2.10}$$

$$\boldsymbol{R}^v_{k,l} = \mathbf{E}\{\boldsymbol{v}_{k,l}\boldsymbol{v}^H_{k,l}\}, \tag{2.11}$$

where $\mathbf{E}\{\}$ is the expectation operator and ^{H} is the Hermitian transpose operator. As with the past-frames vector, the speech IFC matrix can also be decomposed into an IFC vector containing the correlated speech components $\boldsymbol{R}^{s,\text{corr}}_{k,l}$ and uncorrelated speech components $\boldsymbol{R}^{s'}_{k,l}$. The rank-1 IFC matrix of the correlated speech components is defined as the expectation value of the correlated components

$$\boldsymbol{R}^{s,\text{corr}}_{k,l} = \mathbf{E}\{\boldsymbol{\gamma}^s_{k,l}S_{k,l}(\boldsymbol{\gamma}^s_{k,l}S_{k,l})^H\} = \phi^s_{k,l}\{\boldsymbol{\gamma}^s_{k,l}(\boldsymbol{\gamma}^s_{k,l})^H\}. \tag{2.12}$$

As a result, the noisy speech IFC matrix $\boldsymbol{R}^x_{k,l}$ can be broken down into the speech and noise IFC matrices, where the speech IFC can further be decomposed into its correlated and uncorrelated components

$$\boldsymbol{R}^x_{k,l} = \boldsymbol{R}^s_{k,l} + \boldsymbol{R}^v_{k,l} = \boldsymbol{R}^{s,\text{corr}}_{k,l} + \boldsymbol{R}^v_{k,l} + \boldsymbol{R}^{s'}_{k,l}. \tag{2.13}$$

Examples of speech and noise IFC matrices can be seen in Fig. 2.2.

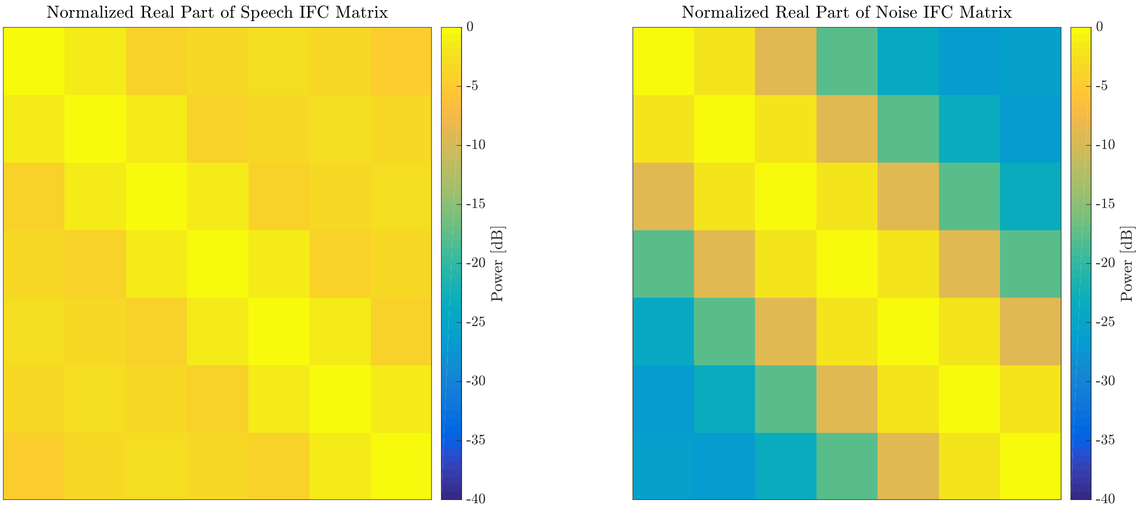

Fig. 2.2. Example of speech and noise IFC matrix estimates.

In Fig. 2.2, it can be seen that speech is correlated over several frames, while the noise only remains weakly correlated within the range of the overlapping frames in the STFT.

A modification which can be made to the multi-frame model is that the uncorrelated speech components are neglected, so that the modified multi-frame model becomes

$$\hat{\boldsymbol{x}}_{k,l} = \boldsymbol{\gamma}^{s}_{k,l} S + \boldsymbol{v}_{k,l}. \tag{2.14}$$

This modification means that the modified noisy speech IFC matrix $\hat{\boldsymbol{R}}^{x}_{k,l}$ is composed of the IFC matrix containing the correlated speech components $\boldsymbol{R}^{s,\text{corr}}_{k,l}$ and the noise IFC matrix $\boldsymbol{R}^{v}_{k,l}$

$$\hat{\boldsymbol{R}}^{x}_{k,l} = \boldsymbol{R}^{s,\text{corr}}_{k,l} + \boldsymbol{R}^{v}_{k,l}. \tag{2.15}$$

It should be noted that the speech, noisy speech and noise PSDs, $\phi^{s}_{k,l}$, $\phi^{x}_{k,l}$, and $\phi^{v}_{k,l}$, are equal to the first element of the respective IFC matrices $\boldsymbol{R}^{s}_{k,l}$, $\boldsymbol{R}^{x}_{k,l}$, and $\boldsymbol{R}^{v}_{k,l}$ as shown

$$\phi^{x}_{k,l} = \mathbf{E}\{|X^{2}_{k,l}|\} = \boldsymbol{e}^{T}_{1} \boldsymbol{R}^{x}_{k,l} \boldsymbol{e}_{1}. \tag{2.16}$$

A common way to estimate the noisy speech PSD $\phi^{x}_{k,l}$ is using the periodogram

$$\hat{\phi}^{x}_{k,l} = |X_{k,l}|^{2}, \tag{2.17}$$

where $\hat{\phi}^{s}_{k,l}$ and $\hat{\phi}^{v}_{k,l}$ can be obtained similarly. The periodogram is often not the best estimate due to its high variance, however, two other common methods are available which aim to produce PSD estimates with less variance, namely, the Welch PSD [15] and the multi-taper PSD [16,17]. The Welch method uses a sliding window to compute an averaged spectrum while the multi-taper method averages a combination pair-wise orthogonal windows. These methods will not be introduced in this thesis, but how to compute them can be found in the literature.

In a blind implementation, the speech PSD can be estimated by first estimating the *a-priori* SNR $\xi_{k,l}$ using the well known decision-directed approach in [18] as follows

$$\hat{\xi}_{k,l} = a \frac{\hat{\phi}^{S}_{k,l-1}}{\hat{\phi}^{v}_{k,l-1}} + (1-a) max\Big(\frac{\phi^{x}_{k,l}}{\hat{\phi}^{v}_{k,l}}\Big), \tag{2.18}$$

where a is a weighting parameter. The *a-priori* SNR $\hat{\xi}_{k,l}$ is then used in a power subtraction [19,20] to obtain the speech PSD estimate $\hat{\phi}^{s}_{k,l}$

$$\hat{\phi}^{s}_{k,l} = \frac{\hat{\xi}_{k,l}}{1+\hat{\xi}_{k,l}} \phi^{x}_{k,l}. \tag{2.19}$$

The normalized IFC vector can be computed using the ML approach from [5], namely

$$\hat{\boldsymbol{\gamma}}^{s}_{k,l} = (1 + \frac{1}{\hat{\xi}_{k,l}})\boldsymbol{\gamma}^{x}_{k,l} - \frac{1}{\hat{\xi}_{k,l}}\hat{\boldsymbol{\mu}}^{v}, \tag{2.20}$$

where $\hat{\boldsymbol{\mu}}^{v}$ is the long-term normalized noise IFC vector which can be estimated using the analysis window h^K, also from [5],

$$\hat{\boldsymbol{\mu}}^{v}[m] = \frac{\sum_n h^K[n]h^K[n+mN]}{\sum_n (h^K[n])^2}. \tag{2.21}$$

To obtain an estimate of the speech coefficient, a complex-valued FIR filter $\boldsymbol{w}_{k,l}$ is applied to the noisy speech frames vector $\boldsymbol{x}_{k,l}$ as follows

$$\hat{S}_{k,l} = \boldsymbol{w}^{H}_{k,l}\boldsymbol{x}_{k,l}. \tag{2.22}$$

Once $\hat{S}_{k,l}$ is obtained, the ISTFT is applied like in Section 2.1 to produce the speech signal estimate $\hat{s}$. The summarized framework is shown in Fig. 2.3.

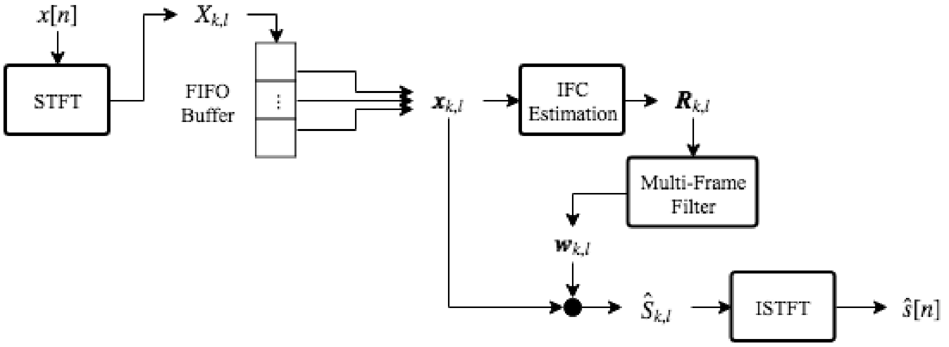

Fig. 2.3. Block diagram of noise reduction where the IFC matrices are used to compute the filter coefficients. The vector of past noisy speech coefficients is stored in a first-in first-out (FIFO) buffer which contains the present and past frames of $X_{k,l}$ defined in (2.6).

2.3 Relation Between Single- and Multi-Frame Signal Models

It is possible to estimate the speech coefficient using scalar gain filters under the assumption that the IFC matrices are Hermitian circulant structured. For this, the causal consecutive-frame vector $\boldsymbol{x}^{M}_{k,l}$ shall be extended to a non-causal vector $\boldsymbol{x}^{2M}_{k,l}$ of length $2M$, defined as

$$\boldsymbol{x}^{2M}_{k,l} = [X_{k,l+M}, X_{k,l+M-1}, ..., X_{k,l}, ..., X_{k,l-M+1}]^T, \tag{2.23}$$

and similarly for $\boldsymbol{s}_{k,l}^{2M}$ and $\boldsymbol{v}_{k,l}^{2M}$. The noisy speech, speech, and noise IFC matrices are defined as in (2.9), (2.10), and (2.11), respectively, but using the consecutive frames vector defined in (2.23).

The STFT in (2.2) can also be applied analogously in a different filterbank with a higher frequency resolution than the K-filterbank. For instance, the F-filterbank with F frequency bands and frequency indices f, calculated similarly to k, using (2.3). In general, $F > K$ and the relation between resolutions of the K- and F-filterbanks is described by the oversampling factor O as follows

$$O = \frac{F}{K} = \frac{2NM}{K}. \tag{2.24}$$

The frequency resolution O depends on the frame shift N since the IFC is sampled at $\frac{1}{N}$ times the sampling frequency f_s which corresponds to a decimation of f_s by factor N. Additionally, given that $2M$ coefficients along consecutive time frames can be transformed into $2M$ coefficients along neighbouring frequencies, the neighbouring frequency bands in the higher resolution F-filterbank also depend on M. An example of how the frequency indices f correspond to the frequency indices k can be seen in Fig. 2.4.

									k													
...	-2				-1				0				1				2				3	...
									f													
...	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	...

Fig. 2.4. Example of filterbank indices k and f for $O = 4$.

Since the IFC matrices are assumed to be Hermitian circulant structured, the correlation vector $\boldsymbol{r}_{k,l}^{x,\text{circ}}$ of length $2M$ which defines the entire Hermitian circulant matrix $\boldsymbol{R}_{k,l}^{x,\text{circ}}$, is defined as

$$\boldsymbol{r}_{k,l}^{x,\text{circ}} = (\boldsymbol{e}_M^{2M})^T \boldsymbol{R}_{k,l}^{x,\text{circ}}. \tag{2.25}$$

$\boldsymbol{e}_M^{2M}$ is the selection vector of length $2M$ and contains all zeros for the coefficients $m' = -M+1, -M+2, \dots, M$, except for a 1 at $m' = 0$, i.e. in the Mth position. To relate the multi-frame model to the single-frame model, the IFC matrix of length $2M$ is assumed to have a Hermitian circulant structure. Hermitian Circulant matrices are a subclass of Hermitian Toeplitz matrices where the additional property holds

$$\boldsymbol{r}[2M - m] = \boldsymbol{r}^*[m] \quad , \qquad m = 0, 1, ..., 2M - 1. \tag{2.26}$$

Thus, the noisy speech IFC matrix $\boldsymbol{R}^{x,\text{circ}}$ has the following structure

$$\boldsymbol{R}^{x,\text{circ}} = \begin{bmatrix} \boldsymbol{r}^x[0] & \boldsymbol{r}^x[2M-1] & \dots & \boldsymbol{r}^x[M] & \dots & \boldsymbol{r}^x[1] \\ \boldsymbol{r}^x[1] & \boldsymbol{r}^x[0] & \ddots & \boldsymbol{r}^x[M+1] & \ddots & \boldsymbol{r}^x[2] \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \boldsymbol{r}^x[M] & \boldsymbol{r}^x[M-1] & \ddots & \boldsymbol{r}^x[0] & \ddots & \boldsymbol{r}^x[M+1] \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \boldsymbol{r}^x[2M-1] & \boldsymbol{r}^x[2M-2] & \dots & \boldsymbol{r}^x[M+1] & \dots & \boldsymbol{r}^x[0] \end{bmatrix}. \tag{2.27}$$

An example of a Hermitian circulant structured speech IFC matrix is shown in Fig. 2.5. To obtain a causal noise IFC matrix estimate $\hat{\boldsymbol{R}}^x_{k,l}$, the Hermitian Toeplitz IFC matrix for the coefficients $m = 0, 1, ..., M-1$ can be extracted from the bottom-right quadrant of the $2M$ x $2M$ circulant IFC matrix, however, it could also be extracted from any M x M block centred along the diagonal.

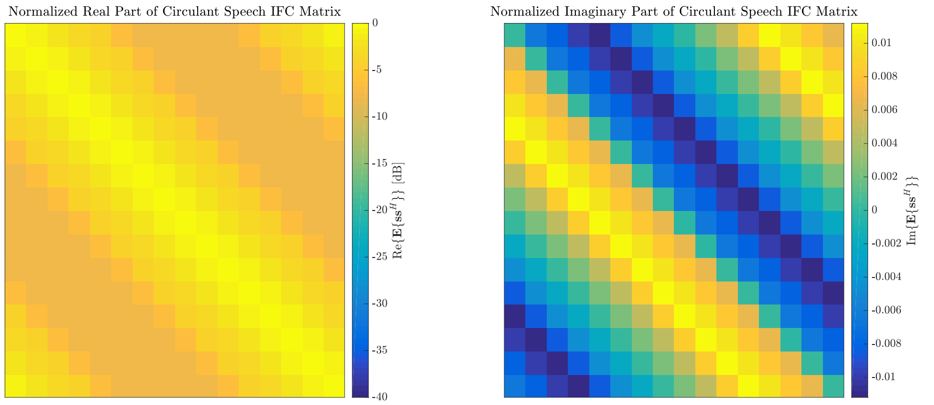

Fig. 2.5. Example of the real and imaginary components of a $2M$x$2M$ Hermitian circulant speech IFC matrix for $M = 8$.

The key transformation used by the filters in this section is based on the assumption that for wide-sense stationary processes, the DFT of the correlation vector $\boldsymbol{r}^x_{k,l}$ results in the PSD $\phi^x_{f,l}$, where the PSD has an O times higher spectral resolution than the coefficients defining the correlation sequence assuming that $2NM > K$, i.e.

$$\phi^x_{Ok+o,l} = \sum_{m=-M+1}^{M} \boldsymbol{r}^{x,\text{circ}}_{k,l}[m] e^{-2\pi(Ok+o)mj} \quad , \qquad o = -\frac{O}{2}+1, -\frac{O}{2}+2, \dots, \frac{O}{2}. \tag{2.28}$$

This means that the frequency subbands f are covered by the indices $Ok + o$. To store the information of the higher resolution PSD in the K-filterbank, the PSD coefficients in the F-filterbank are windowed around the centre-frequencies of the K-filterbank defined by k, as follows

$$\boldsymbol{\phi}^x_{k,l}[o'] = \frac{1}{2M}|\boldsymbol{H}^{F:K}[o']|^2\phi^x_{Ok+o',l} \quad , \qquad o' = -M+1, -M+2, \dots, M. \tag{2.29}$$

where $\boldsymbol{H}^{F:K}[o']$ contains the $2M$ central coefficients ($-M+1$ to M) of the F-point DFT of the analysis window h^K. Examples of $\boldsymbol{H}^{F:K}$ for different oversampling factors can be seen in Figs. 2.6 and 2.7. The windowing preserves the power in the centre coefficients corresponding to a whole frequency band in the K-filterbank and attenuates the power in more distant frequency bands.

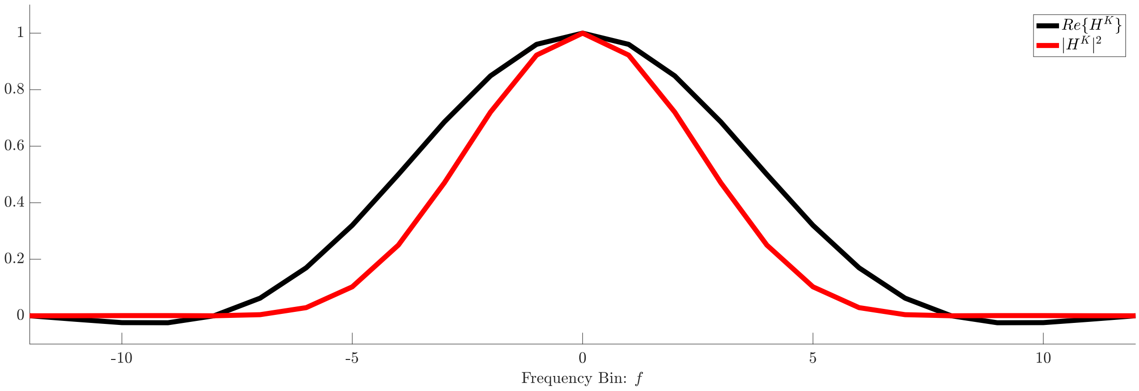

Fig. 2.6. $Re\{\boldsymbol{H}^{F:K}[f]\}$ (Real part of $4K$-point DFT of a Hann Analysis window h^K of length $K = 64$) and $|\boldsymbol{H}^{F:K}[f]|^2$ plotted for frequency coefficients f.

If for example, $O = 2$, then $\boldsymbol{H}^{F:K}$ windows the frequency coefficients f in the F-filterbank more narrowly than for higher values of O, to correspond to the frequency bands in the K-filterbank as shown in Fig. 2.7, since each frequency band f has a larger bandwidth for a smaller value of O.

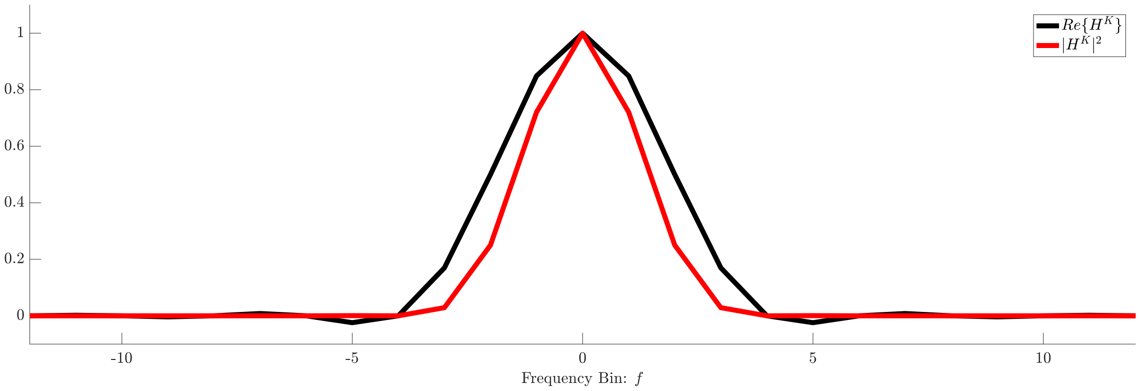

Fig. 2.7. $Re\{\boldsymbol{H}^{F:K}[f]\}$ (Real part of $2K$-point DFT of a Hann Analysis window h^K of length $K = 64$) and $|\boldsymbol{H}^{F:K}[f]|^2$ plotted for frequency coefficients f.

Placing the PSD coefficients from each frequency band k in (2.29) into a diagonal matrix as follows, results in the $2M \times 2M$ diagonal PSD coefficient matrix $\boldsymbol{\Phi}^x_{k,l}$, defined as

$$\boldsymbol{\Phi}^{x}_{k,l} = \boldsymbol{I}^{2M\times 2M} \cdot \boldsymbol{\phi}^{x}_{k,l} = \begin{bmatrix} \phi_{k,l}[-M+1] & 0 & \dots & 0 \\ 0 & \phi_{k,l}[-M+2] & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \phi_{k,l}[M] \end{bmatrix}. \tag{2.30}$$

Using the DFT matrix

$$\boldsymbol{D} = \frac{1}{\sqrt{2M}} \begin{bmatrix} e^{-2\pi j(-M+1)(-M+1)} & \dots & 1 & \dots & e^{-2\pi j(-M+1)(M)} \\ \vdots & \ddots & 1 & \ddots & \vdots \\ 1 & 1 & 1 & 1 & 1 \\ \vdots & \ddots & 1 & \ddots & \vdots \\ e^{-2\pi j(M)(-M+1)} & \dots & 1 & \dots & e^{-2\pi j MM} \end{bmatrix}, \tag{2.31}$$

the Hermitian circulant IFC matrix for a subband K can be estimated as follows

$$\hat{\boldsymbol{R}}^{x,\mathrm{circ}}_{k,l} = \frac{1}{2M} \boldsymbol{D}^{H} \boldsymbol{\Phi}^{x}_{k,l} \boldsymbol{D}. \tag{2.32}$$

This is based on the property from [21], where it is shown that any circulant matrix $\boldsymbol{R}^{x,\mathrm{circ}}_{k,l}$ has eigenvectors

$$\boldsymbol{d}[m] = \frac{1}{\sqrt{2M}} [e^{-2\pi j m(-M+1)}, \dots, e^{-2\pi j m M}]^{T} \quad , \qquad m = -M+1, -M+1, \dots, M \tag{2.33}$$

and corresponding eigenvalues contained in $\boldsymbol{\phi}^{x}_{k,l}$, which define the values along the diagonal of the diagonal eigenmatrix $\boldsymbol{\Phi}^{x}_{k,l}$. Equivalent approximations as in (2.32) can be made for $\boldsymbol{R}^{s,\mathrm{circ}}_{k,l}$ and $\boldsymbol{R}^{v,\mathrm{circ}}_{k,l}$ with (2.32). Using the transformation

$$\boldsymbol{W}_{k,l} = \boldsymbol{D}\boldsymbol{w}_{k,l}, \tag{2.34}$$

the transformed filter coefficient vectors $\boldsymbol{W}_{k,l}$ can be windowed and overlapped into a scalar gain which is applied to the noisy speech $X_{f,l}$ in the F-filterbank. The general formula to obtain $G_{f,l}$ is

$$G_{f,l} = \sum_{o=-\frac{O}{2}+1}^{\frac{O}{2}} \boldsymbol{H}^{F:K}[c - Oo] \boldsymbol{W}_{\mathrm{mod}(f'+o,K)+1-\frac{K}{2},l}[c - Oo] \tag{2.35}$$

where mod() is the modulo operator and $\lfloor\ \rfloor$ is the floor operator. $f' + o$ determines the frequency bands k which are used in the WOLA procedure, where f' is defined as

$$f' = \lfloor \frac{f-1}{O} \rfloor + \frac{K}{2} - 1 \tag{2.36}$$

and $c - Oo$ determines which coefficients of the filter vector $\boldsymbol{W}_{k,l}$ are overlapped with coefficients in adjacent frequency bands determined by f', where c is defined as

$$c = O(\frac{O}{2} - 1) - \mathrm{mod}(O - f, O). \tag{2.37}$$

Summing the overlapped, windowed coefficients in the K-filterbank leads to a constant scalar value in the F-filterbank, since the coefficients in the K-filterbank have a frequency resolution equal to the F-filterbank. For the WOLA procedure to be successful, $2M = O^2$ must hold to obtain the desired scalar gain $G_{f,l}$. In practice, since the DFT is periodic, the modulo function is applied to the frequency bands which go out of range of k, e.g. $\frac{K}{2} + 1$ would become $-\frac{K}{2} + 1$. A visual example of the overlap process of the windows can be seen in Fig. 2.8.

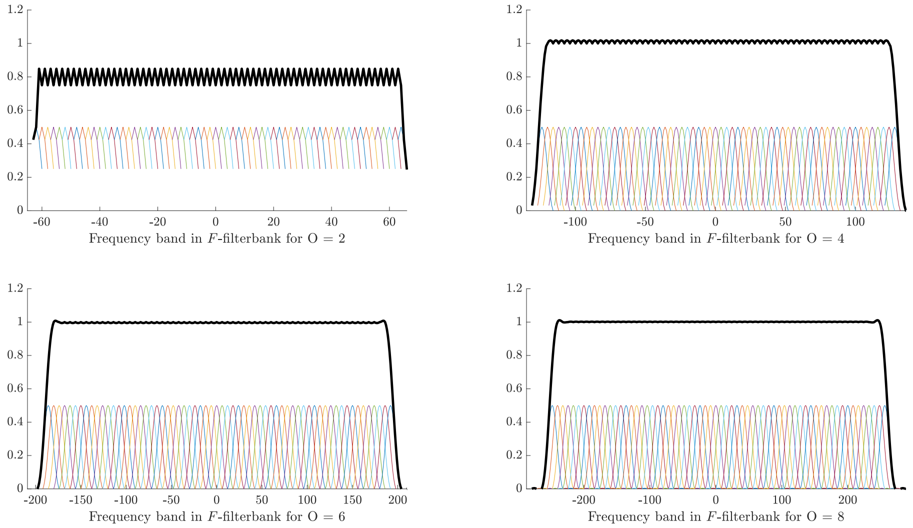

Fig. 2.8. Example of overlapped windows $\boldsymbol{H}^{F:K}$ for $K = 64$ and $O = 2, 4, 6, 8$.

The overlap procedure can be applied as long as $O > 1$, otherwise the F-filterbank has the same or a worse resolution than the K-filterbank. For $O = 2$ it can be seen that the overlap procedure produces a slightly inconsistent value, however, the fluctuations are quite small and as such it can still be used for the overlap procedure. To obtain the speech estimate $s[n]$ in the time domain, the ISTFT is applied to the speech coefficient $\hat{S}_{k,l}$. The summarized framework is shown in Fig. 2.9.

Alternatively, with the help of (2.32), the PSD in the F-filterbank $\phi_{f,l}$ can be used to estimate the Hermitian circulant IFC matrices $\boldsymbol{R}_{k,l}^{\mathrm{circ}}$ in the K-filterbank, to be used in the complex-valued multi-frame filters described in Section 2.2. This procedure is summarized in Fig. 2.10.

Another option would be to use (2.32) to estimate the diagonal matrices containing

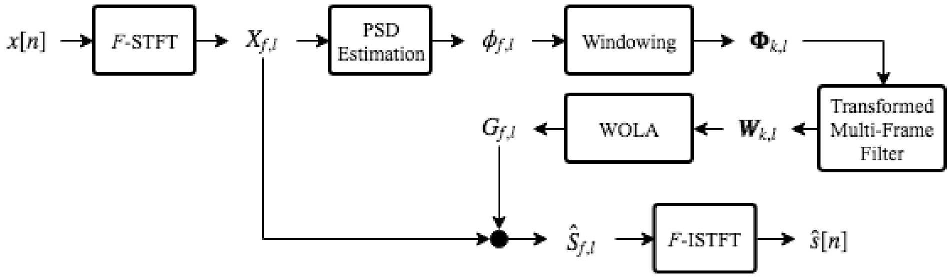

Fig. 2.9. Block diagram of noise reduction using the PSD coefficients to compute the transformed filter coefficients $W_{k,l}$, which are overlapped into real-valued filter gains $G_{f,l}$ in the F-filterbank.

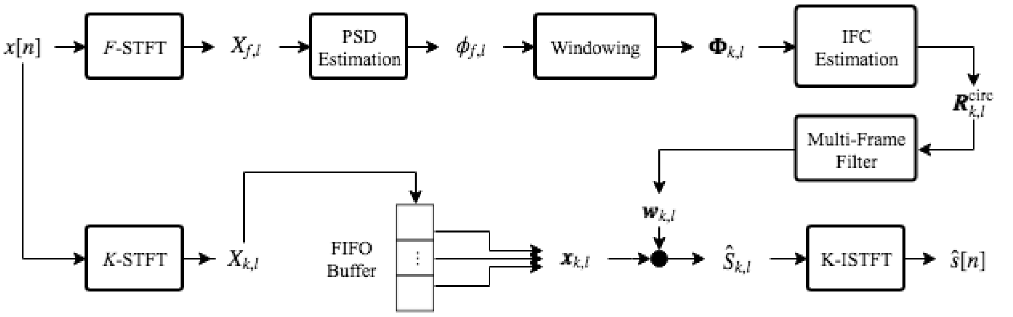

Fig. 2.10. Block diagram of noise reduction where the PSD coefficients are transformed to estimate the circulant IFC matrices in the K-filterbank, which are used to compute complex-valued filters $w_{k,l}$.

the PSD coefficients $\mathbf{\Phi}_{k,l}$ from the circulant IFC matrices $\hat{\boldsymbol{R}}^{\text{circ}}_{k,l}$. This procedure is summarized in Fig. 2.11.

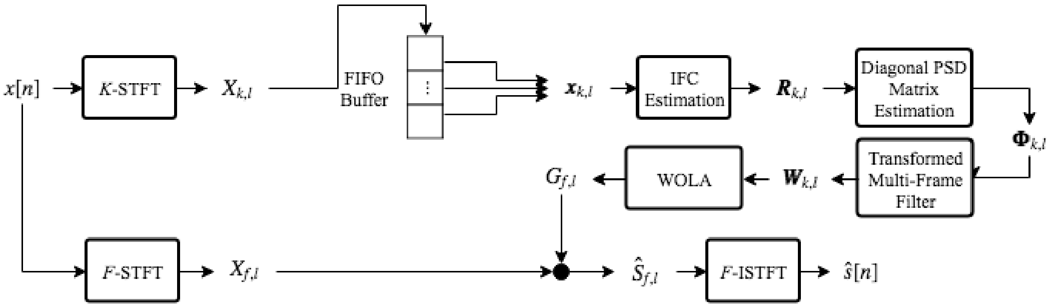

Fig. 2.11. Block diagram of noise reduction where the IFC matrices in the K-filterbank are transformed into PSD coefficients. The PSD coefficients are used to compute the transformed filter coefficients $W_{k,l}$ which are overlapped into real-valued filter gains $G_{f,l}$ in the F-filterbank.

In the rest of this thesis, the frequency bands and time frame index k, f, and l will be omitted for better readability, wherever only the current frame and frequency index are needed.

3 Single-Channel Noise Reduction

This section introduces some variations of Wiener filters, beginning with the WGs and the SDW-WG in the single-frame model in Section 3.1. In Section 3.2, the IFWF and the SDW-IFWFs from [2] are derived and discussed within the multi-frame model. The SDW-IFWFs are also extended to filters which can be condensed into scalar gains by assuming that the IFC matrix is circulant structured, as shown in [6]. Throughout this section it is assumed that noise and speech are uncorrelated with each other and to conclude this section, some practical methods of estimating the IFC matrices for the multi-frame signal model are discussed.

3.1 Single-Frame Filters

This subsection begins with the theory behind the WGs and extends the filter to the SDW-WGs which include a trade-off parameter which balances noise reduction with speech-distortion.

3.1.1 WGs

Assuming that speech and noise are uncorrelated over consecutive time frames, the single-channel WG can be derived by minimizing the mean-square error (MSE) between the filtered noise and the speech at a current time-frequency point as follows

$$J^{\mathrm{WG}} = \mathbf{E}\{|GX - S|^2\}. \tag{3.1}$$

Since the real-valued gains G are scalar values in the cost function, they are inherently limited in terms of how much noise they can filter out since they disregard the phase of the noise coefficient, however, they also benefit from reduced computational complexity, which may be more desired in some applications. Solving (3.1) leads to the WG

$$G^{\mathrm{WG}} = \frac{\xi}{1+\xi}, \tag{3.2}$$

where ξ is defined as the signal-to-noise ratio (SNR) defined by the PSDs of the speech and noise signals

$$\xi = \frac{\phi^s}{\phi^v}. \tag{3.3}$$

To vary the intensity of the noise reduction, a SDW-WG can be derived with a parameter μ which can increase the level of noise reduction with the drawback of increased speech distortion. The corresponding cost function is given as

$$J^{\text{SDW-WG}} = \mathbf{E}\{|GS - S|^2\} + \mu\mathbf{E}\{|GV|^2\}. \tag{3.4}$$

The solution to (3.4) is the SDW-WG

$$G^{\text{SDW-WG}} = \frac{\xi}{\mu + \xi}. \tag{3.5}$$

For $\mu = 0$, the SDW-WG filters out no noise since the gain equals unity, i.e.

$$G^{\text{SDW-WG }(\mu\, =\, 0)} = \frac{\xi}{\xi} = 1. \tag{3.6}$$

For $\mu = 1$, the SDW-WG is equivalent to the WG in (3.2).

3.2 Complex-Valued Multi-Frame Filters

In this section, the WG is extended to the multi-frame signal model by taking into account the correlation of speech and noise across consecutive time frames. The multi-frame equivalent to the WG is the IFWF, which forms the basis for the SDW-IFWF, where the trade-off between noise reduction and speech distortion can be varied. Unlike in the single-frame signal model where the correlation between time frames was neglected, SDW-IFWFs can be derived which aim to preserve the correlated speech components while suppressing the uncorrelated components. It is also shown how an IFWF can be decomposed into an inter-frame minimum power distortionless-response (IFMPDR) filter and a WG, which is shown in [4] to be robust when implemented as an IFMPDR with a WG post-filter.

3.2.1 IFWFs

The IFWF can be derived by minimizing the MSE between the filtered past noisy frames and the current speech frame

$$J^{\text{IFWF}} = \mathbf{E}\{|\boldsymbol{w}^H\boldsymbol{x} - S|^2\}, \tag{3.7}$$

to which the solution is the IFWF filter

$$\boldsymbol{w}^{\text{IFWF}} = (\boldsymbol{R}^x)^{-1}\boldsymbol{R}^s\boldsymbol{e}_1 = (\boldsymbol{R}^s + \boldsymbol{R}^v)^{-1}\boldsymbol{R}^s\boldsymbol{e}_1. \tag{3.8}$$

It is also known that the IFWF can be decomposed into an IFMPDR filter multiplied with a WG, therefore, the derivation of the IFMPDR will be discussed briefly. The IFMPDR aims to minimize the total signal output power while preserving the correlated (desired) speech component

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize:}} \quad & \boldsymbol{w}^H \boldsymbol{R}^x \boldsymbol{w} \\ \text{subject to:} \quad & \boldsymbol{w}^H \boldsymbol{\gamma}^s = 1. \end{aligned} \tag{3.9}$$

The solution to this optimization problem is

$$\boldsymbol{w}^{\text{IFMPDR}} = \frac{(\boldsymbol{R}^x)^{-1}\boldsymbol{\gamma}^s}{(\boldsymbol{\gamma}^s)^H(\boldsymbol{R}^x)^{-1}\boldsymbol{\gamma}^s}. \tag{3.10}$$

After applying the matrix inversion lemma to (3.8), it is seen that the IFWF becomes an IFMVPR filter multiplied by a WG

$$\boldsymbol{w}^{\text{IFWF}} = (\boldsymbol{R}^x)^{-1}\boldsymbol{R}^s\boldsymbol{e}_1 = \frac{(\boldsymbol{R}^x)^{-1}\boldsymbol{\gamma}^s}{(\boldsymbol{\gamma}^s)^H(\boldsymbol{R}^x)^{-1}\boldsymbol{\gamma}^s}\frac{\xi}{\xi+1} = \boldsymbol{w}^{\text{IFMVPR}}G^{\text{WG}}. \tag{3.11}$$

In [4] it is shown that in practice, an implementation of a IFMPDR with WG post-filtering applies more noise reduction than an IFMPDR filter and less speech distortion than an IFWF. Since an IFMPDR filter contains a constraint to preserve speech, it reduces the amount of noise present in the signal without affecting the speech before filtering with a WG. The result is a good compromise between an IFMPDR and an IFWF, however, the main advantage is that this implementation of the IFWF is always robust since the IFMPDR and WG are both inherently robust.

3.2.2 SDW-IFWFs

In [2, 6], SDW-IFWFs are derived with a speech distortion parameter μ which provides a trade-off between noise reduction and speech distortion. For $\mu = 0$, the filters are free of speech distortion, i.e. either no filtering occurs or they become an IFMPDR filter (depending on the implementation). As μ is increased, more noise is reduced, however, the price to pay is increased speech distortion.

The minimization problem which balances speech distortion with noise reduction is defined by the trade-off of the MSE which minimizes speech distortion and the MSE which minimizes noise reduction, depending on the parameter μ as follows

$$J^{\text{SDW-IFWF}} = \mathbf{E}\{|\boldsymbol{w}^H\boldsymbol{s} - S|^2\} + \mu\mathbf{E}\{|\boldsymbol{w}^H\boldsymbol{v}|^2\}. \tag{3.12}$$

The corresponding solution to (3.12) is given as

$$\boldsymbol{w}^{\text{SDW-IFWF}} = (\boldsymbol{R}^s + \mu \boldsymbol{R}^v)^{-1} \boldsymbol{R}^s \boldsymbol{e}_1. \tag{3.13}$$

In this case, when $\mu = 0$, no filtering takes place

$$\boldsymbol{w}^{\text{SDW-IFWF } (\mu\,=\,0)} = \boldsymbol{e}_1 \tag{3.14}$$

and when $\mu = 1$, the SDW-IFWF becomes an IFWF

$$\boldsymbol{w}^{\text{SDW- IFWF } (\mu\,=\,1)} = (\boldsymbol{R}^s + \boldsymbol{R}^v)^{-1} \boldsymbol{R}^s \boldsymbol{e}_1 = (\boldsymbol{R}^x)^{-1} \boldsymbol{R}^s \boldsymbol{e}_1. \tag{3.15}$$

Using the modified multi-frame model in (2.14) where the uncorrelated speech components are neglected, the corresponding cost function to be minimized becomes

$$J^{\text{SDW-IFWF-1}} = \mathbf{E}\{|\boldsymbol{w}^H \boldsymbol{\gamma}^s S - S|^2\} + \mu \mathbf{E}\{|\boldsymbol{w}^H \boldsymbol{v}|^2\}. \tag{3.16}$$

The solution to the modified cost function (3.16) is the SDW-IFWF-1

$$\boldsymbol{w}^{\text{SDW-IFWF-1}} = \frac{(\boldsymbol{R}^v)^{-1} \boldsymbol{\gamma}^s \phi^s}{\mu + (\boldsymbol{\gamma}^s)^H (\boldsymbol{R}^v)^{-1} \boldsymbol{\gamma}^s \phi^s}, \tag{3.17}$$

which, when μ is set to 0, produces an IFMVDR filter

$$\boldsymbol{w}^{\text{SDW-IFWF-1 } (\mu\,=\,0)} = \frac{(\boldsymbol{R}^v)^{-1} \boldsymbol{\gamma}^s}{(\boldsymbol{\gamma}^s)^H (\boldsymbol{R}^v)^{-1} \boldsymbol{\gamma}^s} = \boldsymbol{w}^{\text{IFMVDR}}. \tag{3.18}$$

This filter is closely related to the IFMPDR filter from (3.9) and is the solution to the constrained optimization problem which tries to minimize the noise power while preserving the correlated speech component

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize:}} \quad & \boldsymbol{w}^H \boldsymbol{R}^v \boldsymbol{w} \\ \text{subject to:} \quad & \boldsymbol{w}^H \boldsymbol{\gamma}^s = 1. \end{aligned} \tag{3.19}$$

Since the SDW-IFWF-1 only takes into account the correlated components of the speech signal, this can be ammended by re-estimating the noise IFC to also contain the uncorrelated components of the speech signal as in (2.7), resulting in the cost function

$$J^{\text{SDW-IFWF-X}} = \mathbf{E}\{|\boldsymbol{w}^H \boldsymbol{\gamma}^s S - S|^2\} + \mu \mathbf{E}\{|\boldsymbol{w}^H \boldsymbol{v} + \boldsymbol{s}'|^2\}. \tag{3.20}$$

The solution to (3.20), which takes into account the uncorrelated components of the speech signal, is the SDW-IFWF-X filter

$$\boldsymbol{w}^{\text{SDW-IFWF-X}} = \frac{(\boldsymbol{R}^x)^{-1} \boldsymbol{\gamma}^s \phi^s}{\mu + (1-\mu)(\boldsymbol{\gamma}^s)^H (\boldsymbol{R}^x)^{-1} \boldsymbol{\gamma}^s \phi^s}, \tag{3.21}$$

which is equivalent to the trade-off filter from [2]. When $\mu = 0$, the SDW-IFWF-X produces the IFMPDR filter from (3.10)

$$\boldsymbol{w}^{\text{SDW-IFWF-X } (\mu = 0)} = \frac{(\boldsymbol{R}^x)^{-1}\boldsymbol{\gamma}^s}{(\boldsymbol{\gamma}^s)^H(\boldsymbol{R}^x)^{-1}\boldsymbol{\gamma}^s} = \boldsymbol{w}^{\text{IFMPDR}}. \tag{3.22}$$

When $\mu = 1$, the SDW-IFWF-X becomes equivalent to the IFWF in (3.8).

3.3 Real-Valued Multi-Frame Filters

In this section, real-valued filters are derived, based on the assumptions made in Section 2.3, which can be overlapped and condensed into scalar gains using (2.35). Based on these assumptions, an IFWF and three SDW-IFWFs will be derived. One of the SDW-IFWFs was already proposed in [6]. However, analogously to the complex-valued SDW-IFWFs from Section 3.2, variations of the multi-frame signal model are used to derive two more versions of the SDW-IFWFs.

3.3.1 IFWF-Cs

To derive a real-valued IFWF, the IFC matrices are assumed to be Hermitian circulant such that the cost function for the complex-valued IFWF in (3.7) can be transformed using (2.32) and (2.34), which leads to the modified cost function

$$\begin{aligned} J^{\text{IFWF}} &= \mathbf{E}\{|\boldsymbol{w}^H\boldsymbol{x} - S|^2\} \\ &= \boldsymbol{w}^H\boldsymbol{R}^{x,\text{circ}}\boldsymbol{w} + (\boldsymbol{e}^{2M})^H\boldsymbol{R}^{x,\text{circ}}\boldsymbol{e}^{2M} - \boldsymbol{w}^H\boldsymbol{R}^{x,\text{circ}}\boldsymbol{e}^{2M} - (\boldsymbol{e}^{2M})^H\boldsymbol{R}^{x,\text{circ}}\boldsymbol{w} \\ &= 2M(\boldsymbol{W}^H\boldsymbol{\Phi}^x\boldsymbol{W} + \mathbf{1}^T\boldsymbol{\Phi}^x\mathbf{1} - \boldsymbol{W}^H\boldsymbol{\Phi}^x\mathbf{1} - \mathbf{1}^T\boldsymbol{\Phi}^x\boldsymbol{W}). \end{aligned} \tag{3.23}$$

The solution to the transformed MSE is the real-valued IFWF-C

$$\boldsymbol{W}^{\text{IFWF-C}} = (\boldsymbol{\Phi}^x)^{-1}\boldsymbol{\Phi}^s\mathbf{1}. \tag{3.24}$$

The filter coefficients $\boldsymbol{W}^{\text{IFWF-C}}$ can also be decomposed into the transformed filter coefficients of an IFWF multiplied by a WG, similar to (3.11). The normalized transformed speech IFC vector $\boldsymbol{\Gamma}^s$ is defined as

$$\boldsymbol{\Gamma}^s = \frac{\boldsymbol{\Phi}^s\mathbf{1}}{\mathbf{1}^H\boldsymbol{\Phi}^s\mathbf{1}}. \tag{3.25}$$

where the normalized transformed noisy speech and noise IFC vectors $\boldsymbol{\Gamma}^x$ and $\boldsymbol{\Gamma}^v$, respectively, can be obtained similarly. By transforming (3.11) with (2.35), the decomposed IFWF-C is obtained

$$\boldsymbol{W}^{\text{IFWF-C}} = \left[\frac{(\boldsymbol{\Phi}^x)^{-1}\boldsymbol{\Phi}^s\mathbf{1}}{(\boldsymbol{\Gamma}^s)^H(\boldsymbol{\Phi}^x)^{-1}\boldsymbol{\Phi}^s\mathbf{1}}\right]\cdot\left[(\boldsymbol{W}^{\text{IFWF-C}})^T\boldsymbol{e}_M^{2M}\right] = \boldsymbol{W}^{\text{IFMPDR-C}}G^{\text{WG}} \tag{3.26}$$

where

$$\boldsymbol{W}^{\text{IFMPDR-C}} = \frac{(\boldsymbol{\Phi}^x)^{-1}\boldsymbol{\Phi}^s\mathbf{1}}{(\boldsymbol{\Gamma}^s)^H(\boldsymbol{\Phi}^x)^{-1}\boldsymbol{\Phi}^s\mathbf{1}}. \tag{3.27}$$

is the solution to minimizing the transformed constrained optimization problem

$$\begin{aligned} \underset{\boldsymbol{W}}{\text{minimize:}} \quad & \boldsymbol{W}^H\boldsymbol{\Phi}^x\boldsymbol{W} \\ \text{subject to:} \quad & \boldsymbol{W}^H\boldsymbol{\Gamma}^s = 1. \end{aligned} \tag{3.28}$$

where the aim is to minimize the total output power while preserving the filtered correlated speech component. Since the MPDR-C can be overlapped into a scalar gain, the IFWF-C can be expressed as a multiplication of two gains using (2.35)

$$G^{\text{IFWF-C}} = G^{\text{IFMPDR-C}}G^{\text{WG}}. \tag{3.29}$$

3.3.2 SDW-IFWF-Cs

The three SDW-IFWFs from Section 3.2.2 can be derived similarly to the IFWF in Section 3.3.1. The SDW-IFWF-CX is from [6] and the SDW-IFWF-C and SDW-IFWF-C1 are proposed here to compare their performance.

Transforming the cost function in (3.12) using (2.35) yields the following cost function

$$J^{\text{SDW-IFWF-C}} = \mathbf{E}\{|\boldsymbol{W}^H\boldsymbol{s} - S|^2\} + \mu\mathbf{E}\{|\boldsymbol{w}^H\boldsymbol{v}|^2\}. \tag{3.30}$$

By assuming that the IFC matrices are circulant, the cost function in (3.12) can be transformed using (2.35), which is solved by the SDW-IFWF-C

$$\boldsymbol{W}^{\text{SDW-IFWF-C}} = (\boldsymbol{\Phi}^s + \mu\boldsymbol{\Phi}^v)^{-1}\boldsymbol{\Phi}^s\mathbf{1}. \tag{3.31}$$

For $\mu = 0$, the SDW-IFWF-C performs no filtering

$$\boldsymbol{W}^{\text{SDW-IFWF-C }(\mu\,=\,0)} = \mathbf{1} \tag{3.32}$$

and for $\mu = 1$, the SDW-IFWF-C becomes is equivalent to the IFWF-C in (3.24). The rank-1 filter in (3.17) using the circulant IFC can be written as

$$\boldsymbol{W}^{\text{SDW-IFWF-C1}} = \frac{(\boldsymbol{\Phi}^v)^{-1}\boldsymbol{\Phi}^s\mathbf{1}}{\mu + A} \quad , \qquad A = \frac{\mathbf{1}^T\boldsymbol{\Phi}^s(\boldsymbol{\Phi}^v)^{-1}\boldsymbol{\Phi}^s\mathbf{1}}{\mathbf{1}^T\boldsymbol{\Phi}^s\mathbf{1}}. \tag{3.33}$$

For $\mu = 0$ becomes an IFMVDR filter similar to (3.18)

$$\boldsymbol{W}^{\text{SDW-IFWF-C1 }(\mu\,=\,0)} = \frac{(\boldsymbol{\Phi}^v)^{-1}\boldsymbol{\Phi}^s\mathbf{1}}{(\boldsymbol{\Gamma}^s)^H(\boldsymbol{\Phi}^v)^{-1}\boldsymbol{\Phi}^s\mathbf{1}} = \boldsymbol{W}^{\text{IFMVDR-C}} \tag{3.34}$$

which aims to minimize the noise power output while preserving the correlated speech component

$$\begin{aligned} \underset{\boldsymbol{W}}{\text{minimize:}} \quad & \boldsymbol{W}^H\boldsymbol{\Phi}^v\boldsymbol{W} \\ \text{subject to:} \quad & \boldsymbol{W}^H\boldsymbol{\Gamma}^s = 1. \end{aligned} \tag{3.35}$$

Analogously, the modified rank-1 filter (3.21) using the circulant IFC is given as

$$\boldsymbol{W}^{\text{SDW-IFWF-CX}} = \frac{(\boldsymbol{\Phi}^x)^{-1}\boldsymbol{\Phi}^s\mathbf{1}}{\mu + (1-\mu)B} \quad , \qquad B = \frac{\mathbf{1}^T\boldsymbol{\Phi}^s(\boldsymbol{\Phi}^x)^{-1}\boldsymbol{\Phi}^s\mathbf{1}}{\mathbf{1}^T\boldsymbol{\Phi}^s\mathbf{1}}. \tag{3.36}$$

For $\mu = 0$, the SDW-IFWF-CX becomes an MPDR filter equivalent to (3.27)

$$\boldsymbol{W}^{\text{SDW-IFWF-CX }(\mu\,=\,0)} = \frac{(\boldsymbol{\Phi}^x)^{-1}\boldsymbol{\Phi}^s\mathbf{1}}{(\boldsymbol{\Gamma}^s)^H(\boldsymbol{\Phi}^x)^{-1}\boldsymbol{\Phi}^s\mathbf{1}} = \frac{\boldsymbol{W}^{\text{IFWF-C}}}{(\boldsymbol{\Gamma}^s)^H\boldsymbol{W}^{\text{IFWF-C}}} = \boldsymbol{W}^{\text{IFMPDR-C}} \tag{3.37}$$

When $\mu = 1$, the SDW-IFWF-CX becomes equivalent to the IFWF-C in (3.24).

3.4 IFC Matrix Estimation

The filters in Sections 3.2 and 3.3 rely on estimates of the speech and noise IFC matrices. A few standard methods of estimating the IFC matrices will be covered in this section. Since speech is typically considered to be stationary for between 10-50 ms, to obtain a good estimate of the speech IFC matrix, it must be averaged within its range of stationarity. A very common method, due to its ease of computation, is the first-order recursive smoothing (FORS) method. Other methods are the autocorrelation sequence (ACS), the autocovariance method (ACM), and a modified autocovariance method (MACM). The latter three are biased correlation matrix estimates, because the past-frame vectors do not have many coefficients and unbiased estimates would suffer from high variance.

3.4.1 FORS

The speech IFC matrix can be estimated the FORS method, which uses a weighted combination of the past IFC matrix and the new estimate

$$\hat{\boldsymbol{R}}^{x,\text{FORS}}_{k,l} = \lambda \hat{\boldsymbol{R}}^{x}_{k,l-1} + (1-\lambda)\boldsymbol{x}^{M}_{k,l}(\boldsymbol{x}^{M}_{k,l})^{H}, \tag{3.38}$$

and is weighted by the smoothing parameter λ. In [3] it is discussed that in the case of IFMPDR filters (with perfect knowledge of the speech and noise signals), the short-term variation of inherently non-stationary speech signals cannot be captured using a large value of λ, however, using small values of λ can produce singular or ill-conditioned IFC matrices. It should also be considered that if, for instance, only poor estimates of the speech are available, then $\hat{\boldsymbol{R}}^{x}$ may need to be averaged over more values to obtain a more accurate estimate of $\mathbf{E}\{\boldsymbol{x}\boldsymbol{x}^{H}\}$ with less variance, in-which case, a higher value of λ would be desirable. One thing to note about $\hat{\boldsymbol{R}}^{x,FORS}_{k,l}$ is that it is generally not Hermitian Toeplitz structured, which it should be in theory for the perfect IFC matrix estimate when making the assumption that speech is stationary within the range of the IFC matrix. The smoothing time constant can be calculated from the smoothing parameter λ as follows

$$\tau = \frac{-N}{\ln(\lambda) f_s}, \tag{3.39}$$

where f_s is the sampling frequency.

3.4.2 ACS

Hermitian Toeplitz matrices are a class of matrices which can be defined by their first row, for example:

$$\boldsymbol{R}^{\text{Toeplitz}} = \begin{bmatrix} \boldsymbol{r}[0] & \boldsymbol{r}[1] & \dots & \boldsymbol{r}[M-1] \\ \boldsymbol{r}[-1] & \boldsymbol{r}[0] & \dots & \boldsymbol{r}[M-2] \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{r}[-M+1] & \boldsymbol{r}[-M+2] & \dots & \boldsymbol{r}[0] \end{bmatrix}, \tag{3.40}$$

where $\boldsymbol{r}[-m] = \boldsymbol{r}^{*}[m]$.

Assuming that X is wide-sense stationary over a period of at least ρ frames, the coefficients defining the Hermitian Toeplitz structure can be estimated using time lags m and m'

$$\begin{aligned}\hat{\boldsymbol{r}}^{x}_{k,l}[m] &= \frac{1}{\rho}\sum_{m'=0}^{\rho-m} X_{k,l-m'}X^{*}_{k,l-m-m'} \quad , \qquad m = 0,1,...,M-1 \\ &= \frac{1}{\rho}\sum_{m'=0}^{\rho-m} \boldsymbol{x}_{k,l}[m']\boldsymbol{x}^{*}_{k,l}[m'+m],\end{aligned} \tag{3.41}$$

where ρ is the number of frames which the signal is assumed to be stationary and must be larger than M.

The noisy speech ACS IFC matrix can also be found by multiplying two data matrices

$$\boldsymbol{R}^{x,\mathrm{ACS}} = (\boldsymbol{A}^{x,\mathrm{ACS}})^{H}\boldsymbol{A}^{x,\mathrm{ACS}}. \tag{3.42}$$

In the case of the ACS, the data matrix is structured as follows

$$\boldsymbol{A}^{x,\mathrm{ACS}} = \begin{bmatrix} \boldsymbol{x}[0] & \dots & 0 \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}[M-1] & \dots & \boldsymbol{x}[0] \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}[\rho-M] & \dots & \boldsymbol{x}[M-1] \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}[\rho] & \dots & \boldsymbol{x}[\rho-M] \\ \vdots & \ddots & \vdots \\ 0 & \dots & \boldsymbol{x}[\rho] \end{bmatrix}, \tag{3.43}$$

where ρ defines the length of the past-frames vector $\boldsymbol{x}^{\rho}_{k,l}$ and M defines the size of the IFC matrix.

3.4.3 ACM

The ACM estimate is essentially an autocorrelation sequence estimator which subtracts the mean from the stationary random process before computing the autocorrelation sequence

$$\boldsymbol{R}^{x,\mathrm{ACM}} = \mathbf{E}\{|(\boldsymbol{x}-\bar{\boldsymbol{x}})(\boldsymbol{x}-\bar{\boldsymbol{x}})^{H}|^{2}\}, \tag{3.44}$$

where $\bar{\boldsymbol{x}}$ is the mean of $\boldsymbol{x}$. To estimate the noisy speech IFC matrix $\boldsymbol{R}^{x,\mathrm{ACM}}$ using the ACM, a similar matrix multiplication of data matrices is applied to (3.42), where the ACM data matrix $\boldsymbol{A}^{x,\mathrm{ACM}}$ uses a submatrix of $\boldsymbol{A}^{x,\mathrm{ACS}}$

$$\boldsymbol{A}^{x,\mathrm{ACM}} = \begin{bmatrix} \boldsymbol{x}[M] & \dots & \boldsymbol{x}[0] \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}[\rho - M + 1] & \dots & \boldsymbol{x}[M] \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}[\rho - 1] & \dots & \boldsymbol{x}[\rho - M + 1] \end{bmatrix}. \tag{3.45}$$

3.4.4 MACM

The MACM data matrix $\boldsymbol{A}^{x,\mathrm{MACM}}$ [22] is a block matrix containing the ACM data matrix with the conjugate column-reversed ACM data matrix stacked beneath. To estimate the noisy speech MACM IFC matrix $\boldsymbol{R}^{x,\mathrm{MACM}}$, (3.42) is used with the MACM data matrix.

$$\boldsymbol{A}^{x,\mathrm{MACM}} = \begin{bmatrix} \boldsymbol{x}[M] & \dots & \boldsymbol{x}[0] \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}[\rho - M + 1] & \dots & \boldsymbol{x}[M] \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}[\rho - 1] & \dots & \boldsymbol{x}[\rho - M + 1] \\ \boldsymbol{x}^*[0] & \dots & \boldsymbol{x}^*[M] \\ \vdots & \iddots & \vdots \\ \boldsymbol{x}^*[M] & \dots & \boldsymbol{x}^*[\rho - M + 1] \\ \vdots & \iddots & \vdots \\ \boldsymbol{x}^*[\rho - M + 1] & \dots & \boldsymbol{x}^*[\rho - 1] \end{bmatrix}. \tag{3.46}$$

4 Evaluation

In this section, a wide range of filter implementations (a filter with a given IFC matrix or PSD estimator) are tested under oracle conditions (knowledge of all the required quantities). The parameters of a given implementation e.g. μ and λ, are varied with the aim of finding the best performing combination out of the variations which are implementable, the best performing multi-frame filter implementations are then tested under blind conditions (knowledge of only the noisy speech signal).

Some quantities will remain fixed, such as the sampling frequency $f_s = 16000$ Hz and the frame length of the K- filterbank which is chosen to be $K = 64$ samples. The value defining the length of the IFC coefficient vector was set to $M = 8$ and the frame shift was set to $N = 16$, resulting in an overlap of 48 samples and therefore a total analysis-synthesis delay of 3 ms. A Hann window of length 64 was used as the analysis and synthesis window in the K-filterbank, as well as the synthesis window in the F-filterbank. The asymmetrical analysis window used in the F-filterbank consisted of the first half of a Hann window of length $F - K/2$ (where F is obtained from (2.24)) concatenated with the second half of a Hann window of length $K/2$. The short analysis windows of length K mean that with a frame shift of N, the analysis-synthesis delay τ_{delay} is only 3 ms based on the relation

$$\tau_{delay} = \frac{K - N}{f_s}. \tag{4.1}$$

Using a synthesis window which is 256 samples long would result in a delay of 15 ms, which is too long for many real-time applications. The resulting analysis and synthesis windows can be seen in Fig. 4.1 and are summarized in Table 1.

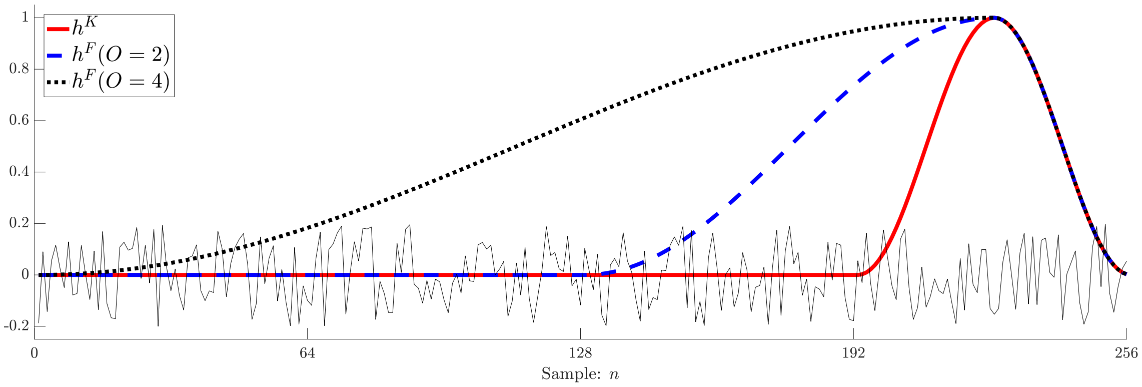

Fig. 4.1. Analysis and Synthesis windows in the K- and F-filterbanks.

Table 1: Analysis and synthesis windows in the K- and F-filterbanks.

	Analysis Window	Synthesis Window
K-filterbank	h^K	h^K
F-filterbank	h^F	h^K

To make the estimation of the inverse of the noisy speech IFC matrix $(\hat{\boldsymbol{R}}^x)^{-1}$ more robust, it was regularized as follows

$$(\hat{\boldsymbol{R}}^x)^{-1} = \left\{ \boldsymbol{R}^x + \delta \frac{\mathrm{tr}(|\boldsymbol{R}^x|)}{M} \boldsymbol{I}_{M\mathrm{x}M} \right\}^{-1} \tag{4.2}$$

where $\delta = 0.04$ was used as the regularization parameter and $\boldsymbol{I}_{M\mathrm{x}M}$ is the identity matrix of dimensions MxM. The same regularization was also applied to the speech and noise IFC matrices $\boldsymbol{R}^s$ and $\boldsymbol{R}^v$, respectively.

The speech quality improvement was evaluated using the perceptual evaluation of speech quality (PESQ) score [23], where the PESQ improvement is calculated over the reference score of the speech and the noisy speech. The speech intelligibility improvement is calculated in the same way but using the short term objective intelligibility (STOI) improvement [24]. The segmental SNR (seg. SNR) improvement is also included as an objective measure for how much noise reduction is applied, and is calculated as the difference between the seg. SNR of the estimated speech and the noisy speech

$$\Delta\text{seg. SNR} = \frac{10}{|\boldsymbol{M}_{sp}|} \sum_{m \in \boldsymbol{M}_{sp}} \log_{10} \frac{\sum_{n=0}^{N-1} s^2[n+mR]}{\sum_{n=0}^{N-1} (s[n+mR] - \hat{s}[n+mR])^2} - \text{seg. SNR}^x, \tag{4.3}$$

where $\boldsymbol{M}_{sp}$ is the set of frames in which speech is present and the seg. SNR of the noisy speech is defined as

$$\text{seg. SNR}^x = \frac{10}{|\boldsymbol{M}_{sp}|} \sum_{m \in \boldsymbol{M}_{sp}} \log_{10} \frac{\sum_{n=0}^{N-1} x^2[n+mR]}{\sum_{n=0}^{N-1} (x[n+mR] - \hat{x}[n+mR])^2}. \tag{4.4}$$

Each implementation was tested on 10 different speech signals from the TIMIT speech corpus [25] with a duration of around 3 seconds (5 male and 5 female speakers). Each speech signal was tested in combination with one of 3 different types of additive noise (SSN, babble noise, and traffic noise) from the NOISEX database [26] and at 3 different SNRs (10 dB, 5 dB, and 0 dB), all-in-all, corresponding to almost 5 minutes of test data. The scores for each implementation were averaged across all 90 tests. The value of each evaluation is plotted as a red circle with the number next to it, indicating the respective maximal PESQ, STOI, or seg. SNR score.

Section 4.1 contains a comprehensive evaluation, of all filters introduced in Section 3 and selected methods of estimating the IFC matrices or PSD coefficients under the assumption of oracle knowledge (knowledge of all required quantities). In Section 4.2, the effect of varying the oversampling factor O is investigated in the real-valued multi-frame filters and in Section 4.3 the best performing implementable multi-frame filters from Section 4.1 are implemented and their performance is evaluated under blind conditions.

4.1 Evaluation with Oracle Knowledge

Within this section, different filter implementations are tested with various methods of estimating the IFC or PSD coefficients and then compared with direct access to the speech and noise components. In Section 4.1.1 the single-frame filters were evaluated, comparing the periodogram, FORS, Multi-taper, and Welch PSD estimates in the K- and F-filterbanks. In Section 4.1.2, the SDW-IFWFs are evaluated in the K-filterbank, comparing the FORS, ACS, ACM, and MACM IFC estimates. In section 4.1.3, the SDW-IFWF-Cs were evaluated in the F-filterbank with both the periodogram and Welch PSD estimates. Furthermore, SDW-IFWFs were evaluated where the IFC matrices in the K-filterbank were obtained using the periodogram PSD estimate in (2.32) and another implementation which was tested was using the ACS IFC matrix estimate to obtain the PSD coefficients to use in the SDW-IFWF-Cs.

It should be noted that the IFC or PSD estimates and parameters were applied the same for speech, noise, and noisy speech, e.g. if the FORS IFC estimate was used with $\lambda = 0.5$ for speech, then the noisy speech and noise IFC estimates would be estimated in the same way. In this section, all filters in the F-filterbank used an oversampling factor of $O = 4$, meaning that the length of the F-filterbank was $F = 256$ samples.

4.1.1 Single-Frame Filters

In this section, the SDW-WGs from (3.5) were tested using Periodogram, the FORS PSD estimate, Welch's method, and the multi-taper method to estimate the PSDs in both the K- and F-filterbank as shown in Table 2.

Table 2: SDW-WGs to be tested with the corresponding PSD estimation methods.

Filter Method	**Quantities**		**PSD Estimation Method**			
SDW-WG	$\phi^s_{k,l}$	$\phi^v_{k,l}$	(2.16) & (3.38)	(2.17)	Multi-taper PSD	Welch PSD
SDW-WG	$\phi^s_{f,l}$	$\phi^v_{f,l}$				

Figs. 4.2, 4.3, 4.4 show the results of the PESQ, STOI, and segmental SNR improvements which the SDW-WGs achieved in the $K-$filterbank. The smoothing parameter λ which is used in the FORS PSD estimate is converted into the smoothing time constant τ using (3.39). The averaging window length is varied in the Welch PSD and the number of windows used in the multi-taper PSD are varied to see which values produce the best performance.

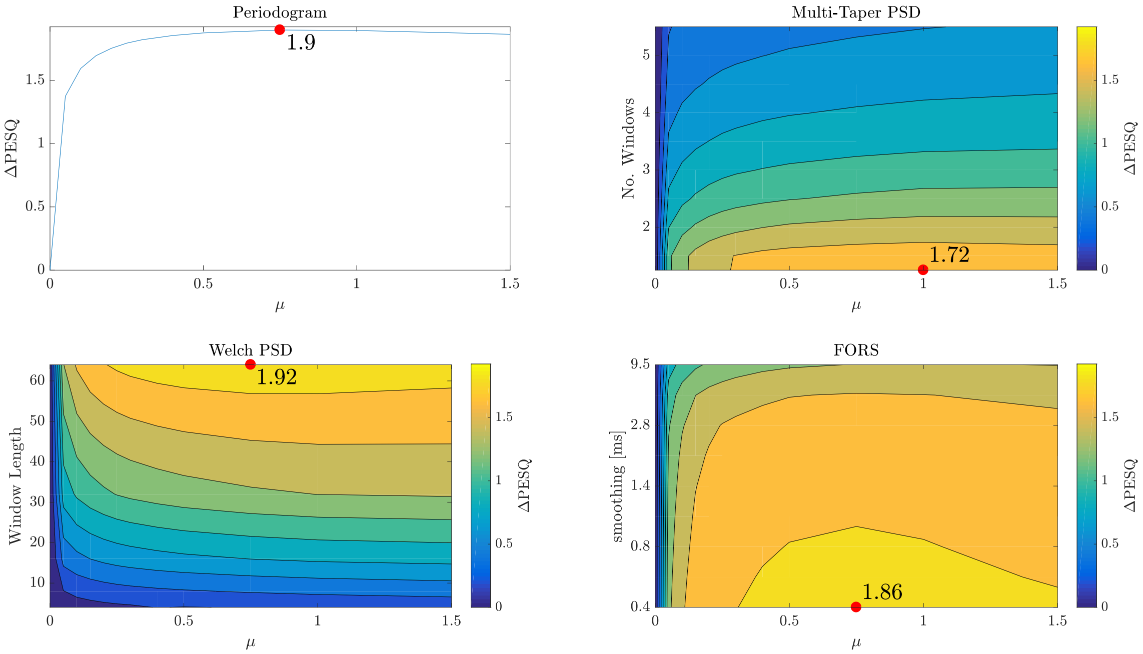

Fig. 4.2. ΔPESQ scores for SDW-WG in the $K-$filterbank.

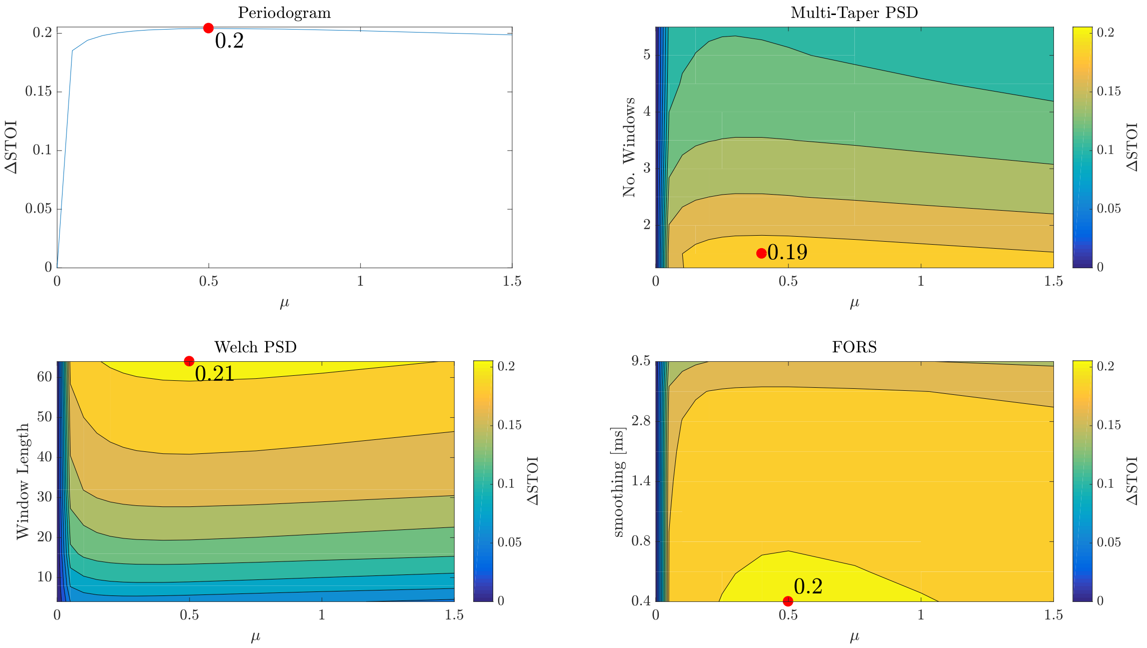

Fig. 4.3. ΔSTOI scores for SDW-WG in the $K-$filterbank.

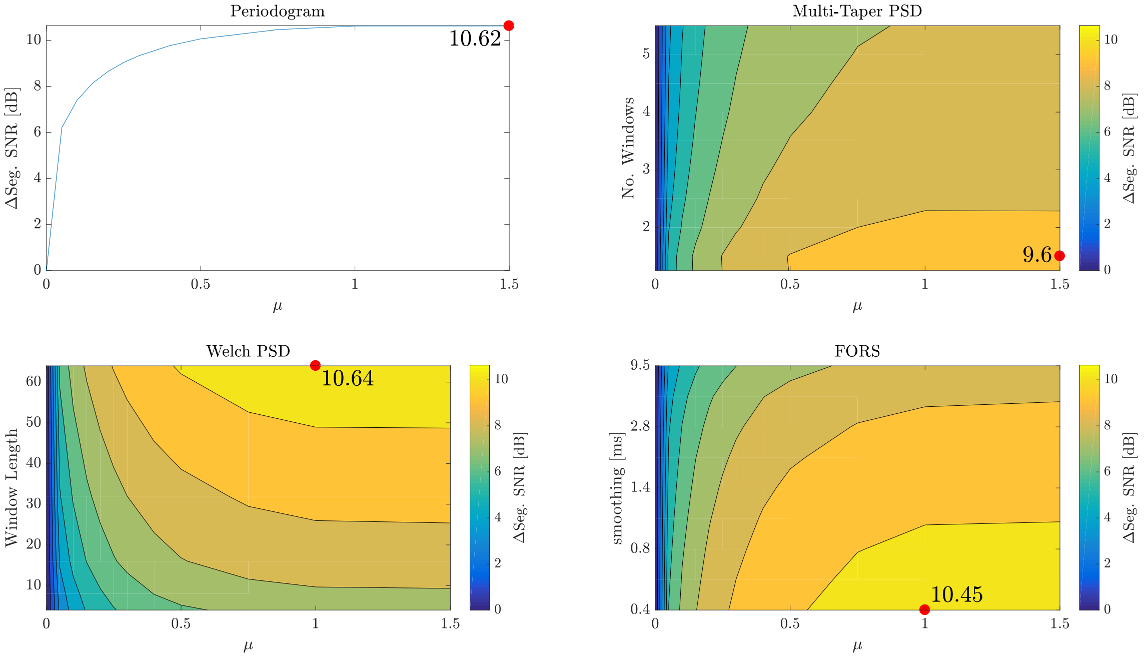

Fig. 4.4. ΔSeg. SNR scores for SDW-WG in the $K-$filterbank.

The Periodogram, Welch PSD with a window length of 64, and FORS with $\lambda = 0.1$ (0.4 ms smoothing), all perform similarly. For the periodogram, Welch, and FORS PSD estimates, the optimal value of μ is 0.75. The STOI improvement favours lower values of μ while the seg. SNR improvement favours higher values, the PESQ improvement provides a good medium. The Welch PSD estimate produces worse results when shorter windows are used and the FORS estimate produces worse results when λ is increased. The multi-taper PSD method also produces good results, especially when less windows are used, however, it is outperformed by all of the other estimators.

Figs. 4.5, 4.6, 4.7 show the results of the SDW-WGs in the $F-$filterbank.

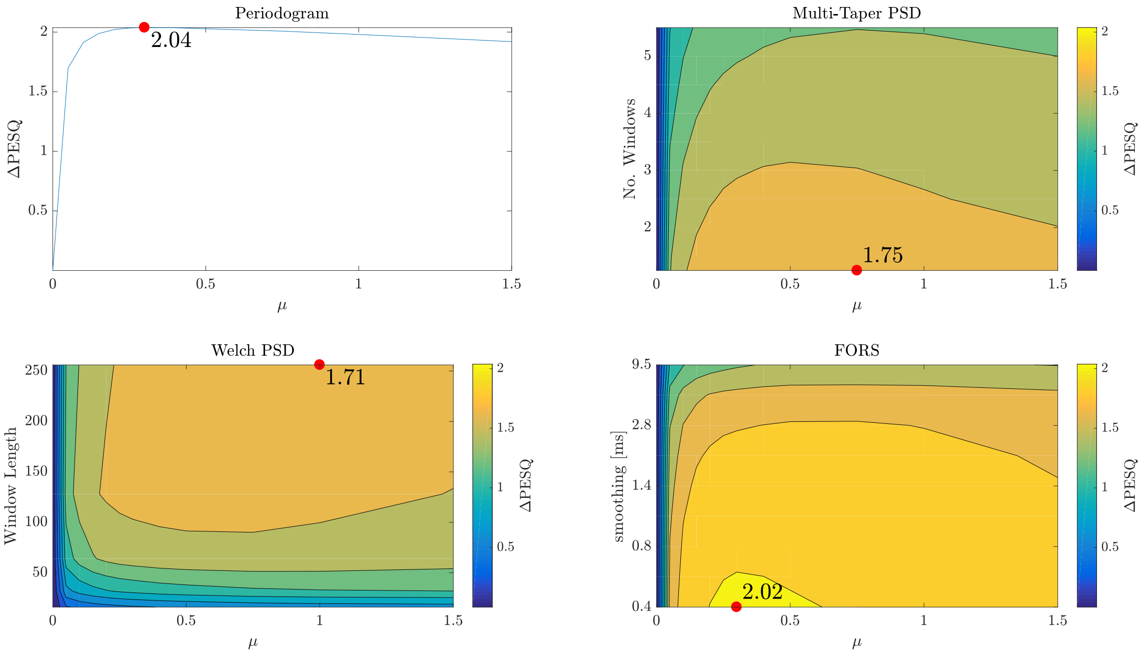

Fig. 4.5. ΔPESQ scores for SDW-WG in the $F-$filterbank.

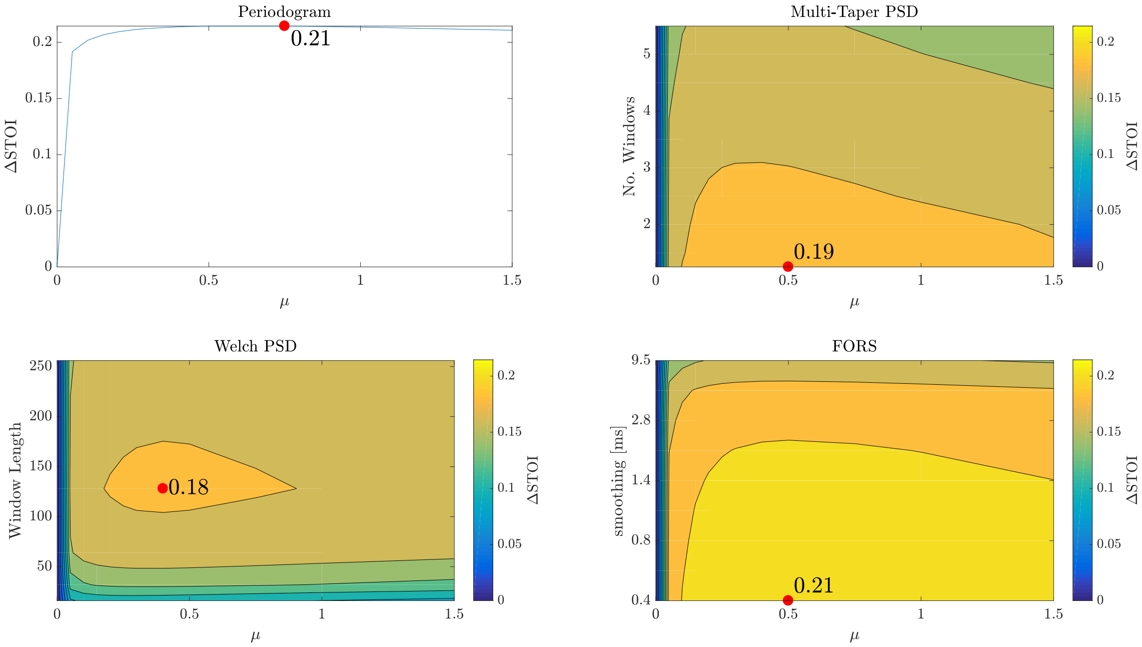

Fig. 4.6. ΔSTOI scores for SDW-WG in the $F-$filterbank.

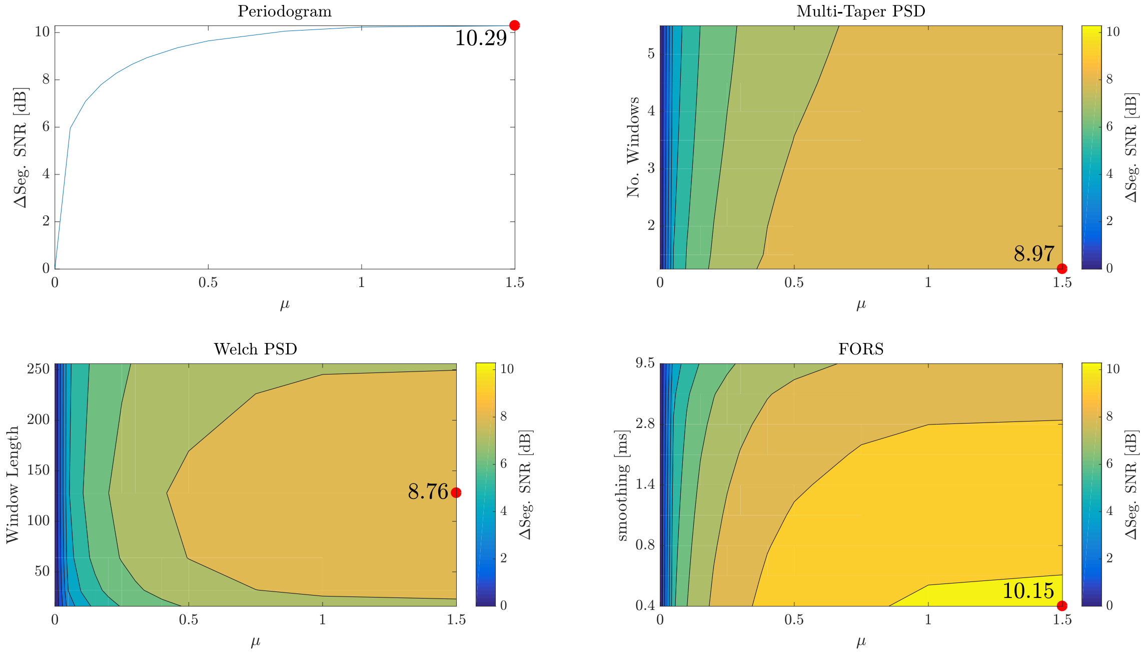

Fig. 4.7. ΔSeg. SNR scores for SDW-WG in the $F-$filterbank.

The periodogram and FORS with $\lambda = 0.1$ (0.4 ms smoothing) perform the best. While $\mu = 0.3$ produces the best PESQ improvement with these methods, $\mu = 0.5-0.75$ produces the best STOI improvement, and the best seg. SNR improvement is obtained for $\mu = 1.5$. Taking all measures into account, $\mu = 0.5$ produces the best overall results. As in the K-filterbank, the results of the FORS PSD estimate worsen for increasing values of λ. The multi-taper and Welch PSD perform worse than the periodogram and FORS for all values of μ, especially in terms of PESQ improvement.

Out of the SDW-WGs, the best results were achieved in the F-filterbank using the periodogram and FORS with $\lambda = 0.1$, however, the Welch PSD, as well as the periodogram and FORS with $\lambda = 0.1$ in the K-filterbank, also produced good results. The main drawback of the SDW-WG is that it provides no filtering for $\mu = 0$.

4.1.2 Multi-Frame Filters

In this section, the three SDW-IFWFs from Section 3.2.2 were tested in the K-filterbank. The FORS, ACS, ACM, and MACM estimators were tested for each filter shown in Table 3.

Table 3: SDW-IFWF filters to be tested with the given IFC matrix estimation methods from Section 3.4.

Filter Method	**Quantities**		**IFC Matrix Estimation Method**			
SDW-IFWF	$\boldsymbol{R}^{s}_{k,l}$	$\boldsymbol{R}^{v}_{k,l}$	FORS	ACS	ACM	MACM
SDW-IFWF-1	$\boldsymbol{R}^{s}_{k,l}$	$\boldsymbol{R}^{v}_{k,l}$	FORS	ACS	ACM	MACM
SDW-IFWF-X	$\boldsymbol{R}^{s}_{k,l}$	$\boldsymbol{R}^{x}_{k,l}$	FORS	ACS	ACM	MACM

Figs. 4.8, 4.9, and 4.10 show the results of the SDW-IFWF filter.

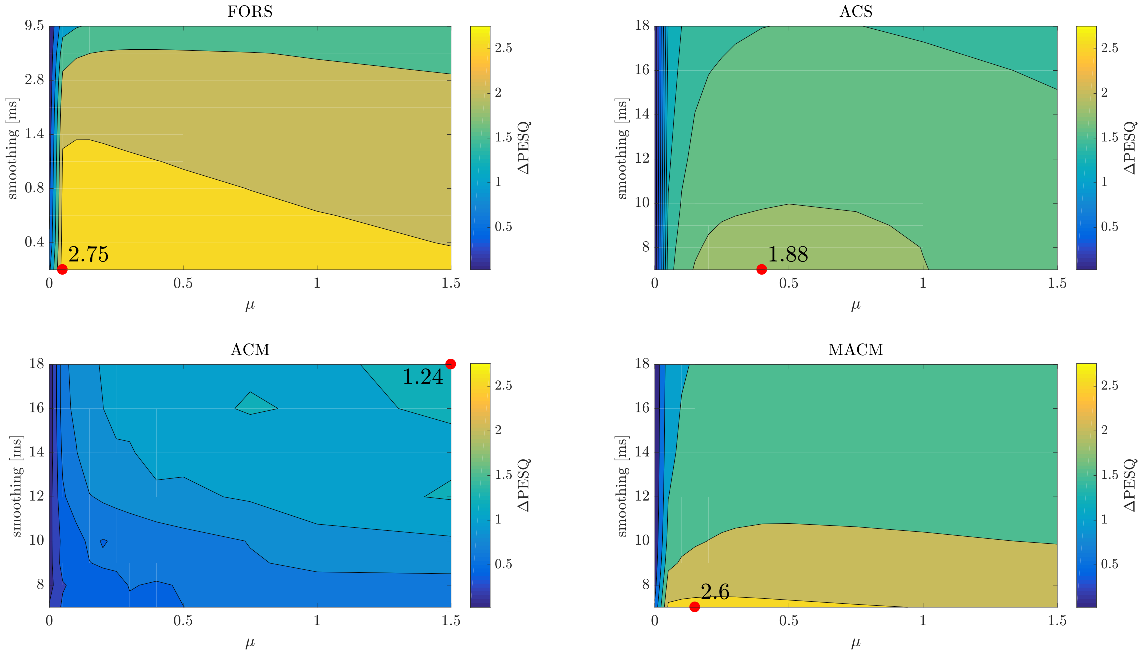

Fig. 4.8. ΔPESQ scores for the SDW-IFWF in the K−filterbank.

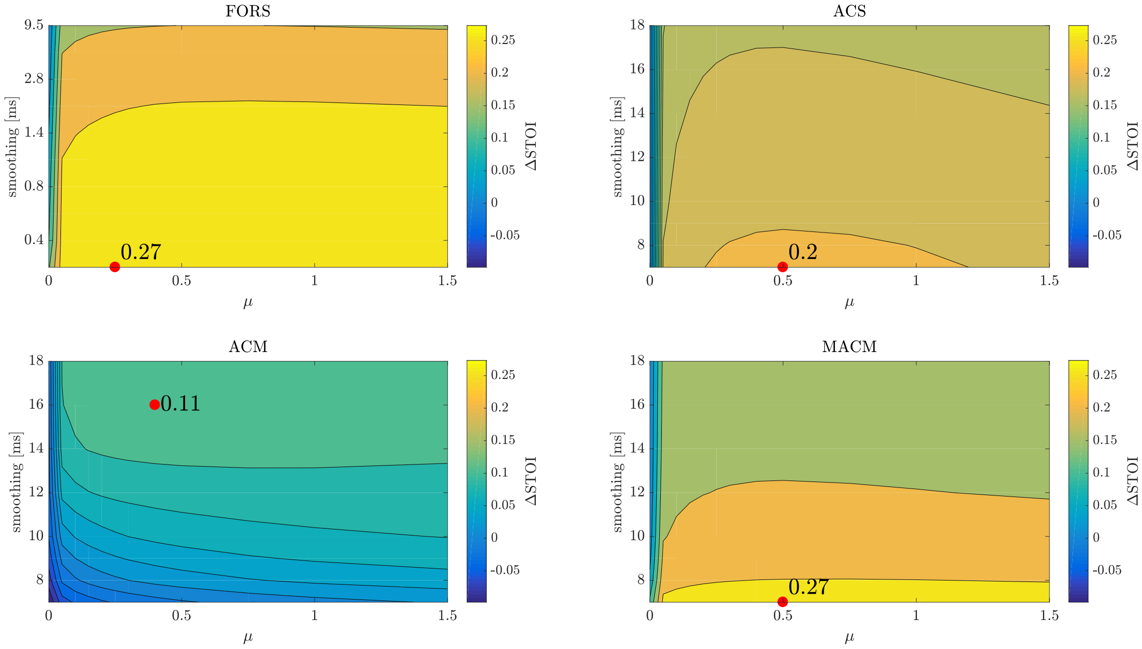

Fig. 4.9. ΔSTOI scores for the SDW-IFWF in the K−filterbank.

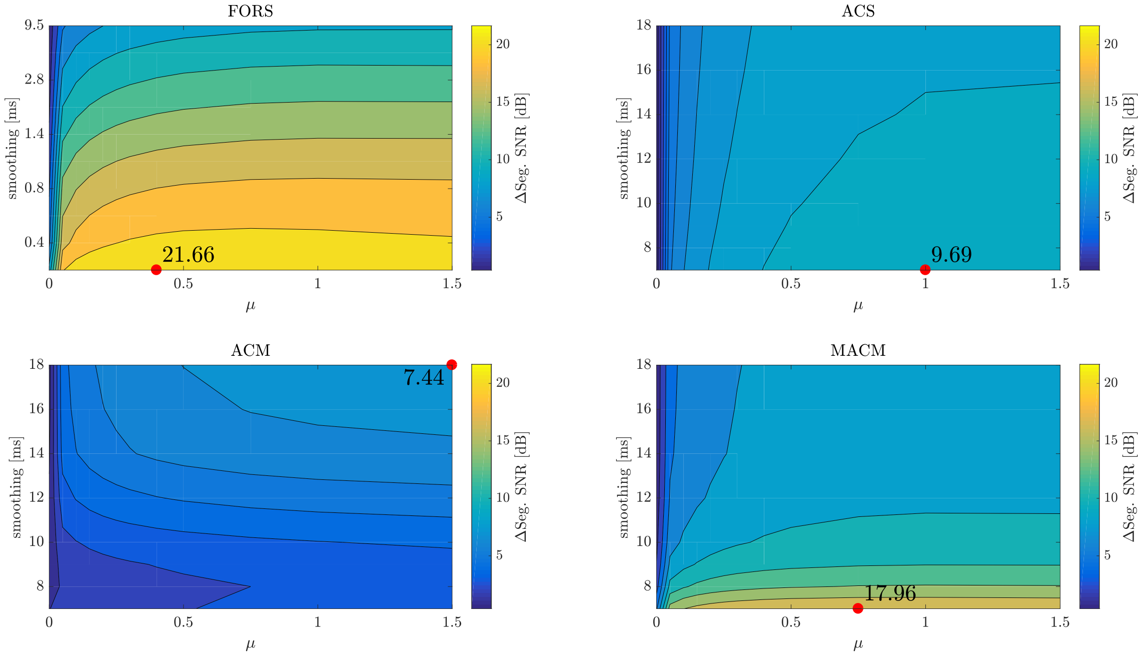

Fig. 4.10. ΔSeg. SNR scores for the SDW-IFWF in the K−filterbank.

The FORS with $\lambda = 0$ (no smoothing) and MACM with $\rho = 7$ (7 ms smoothing) were the best performing IFC matrix estimators for the SDW-IFWF. Taking into account the PESQ, STOI, and seg. SNR improvements, the FORS method performed the best at $\mu = 0.3$ and the MACM performed the best at $\mu = 0.5$. With FORS, the ACS, and the MACM, the shorter the smoothing, the better the performance of the SDW-IFWF under oracle conditions. The ACS performed worse for this filter than the FORS and the MACM estimates, and the ACM yielded the worst results.

Figs. 4.11, 4.12, and 4.13 show the results of the SDW-IFWF-1 filter.

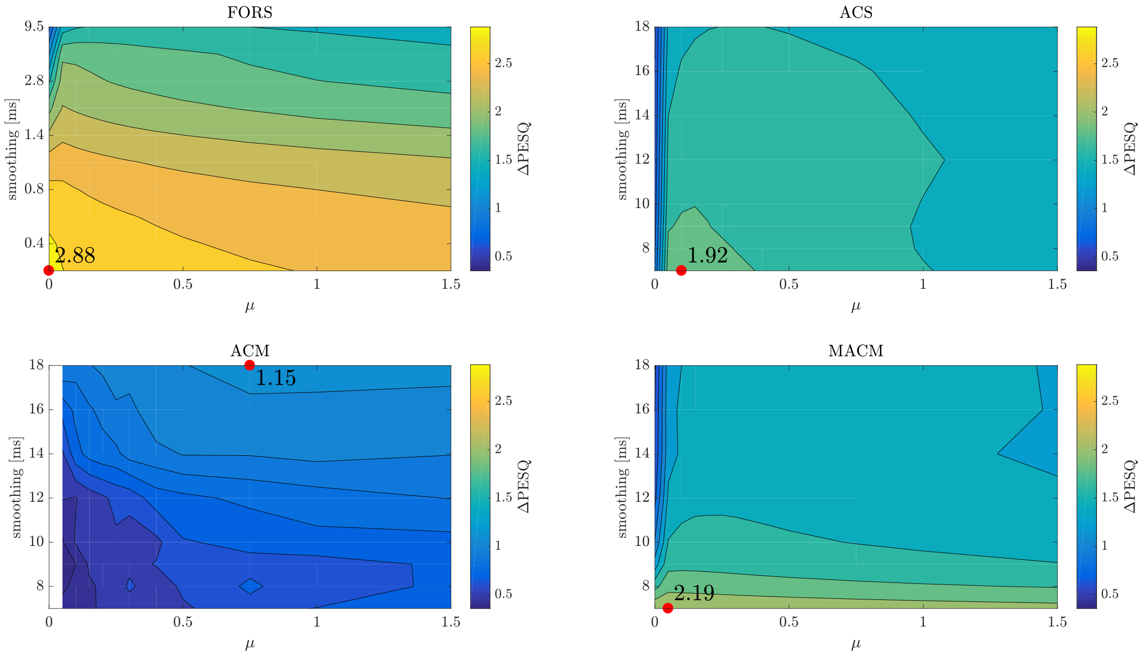

Fig. 4.11. ΔPESQ scores for the SDW-IFWF-1 in the K−filterbank.

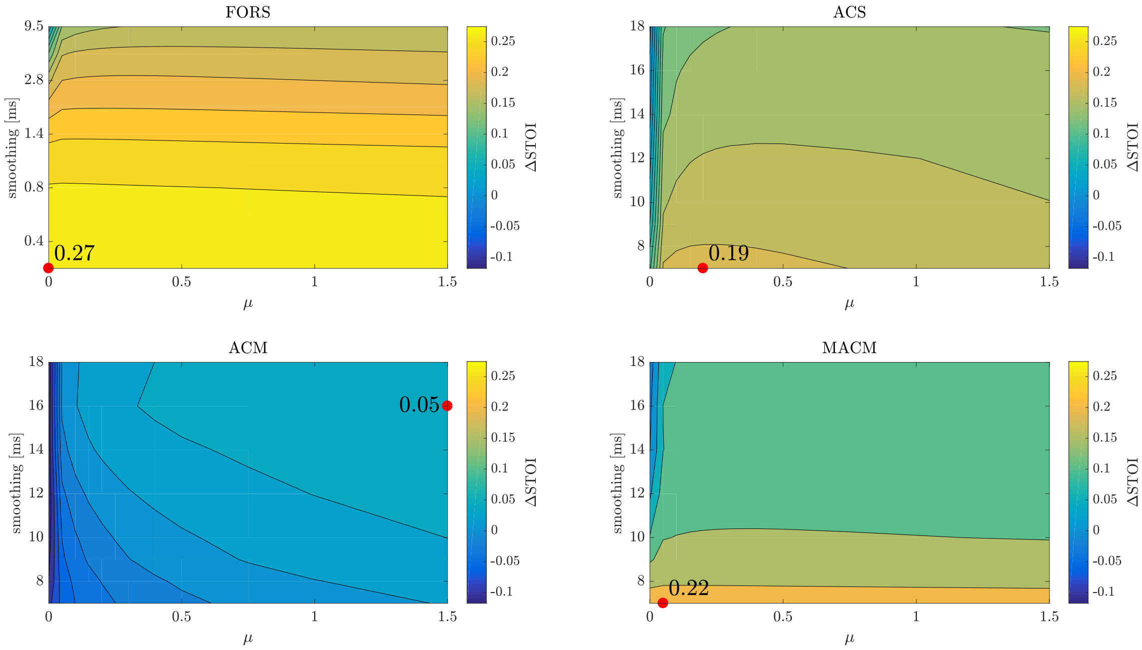

Fig. 4.12. ΔSTOI scores for the SDW-IFWF-1 in the K−filterbank.

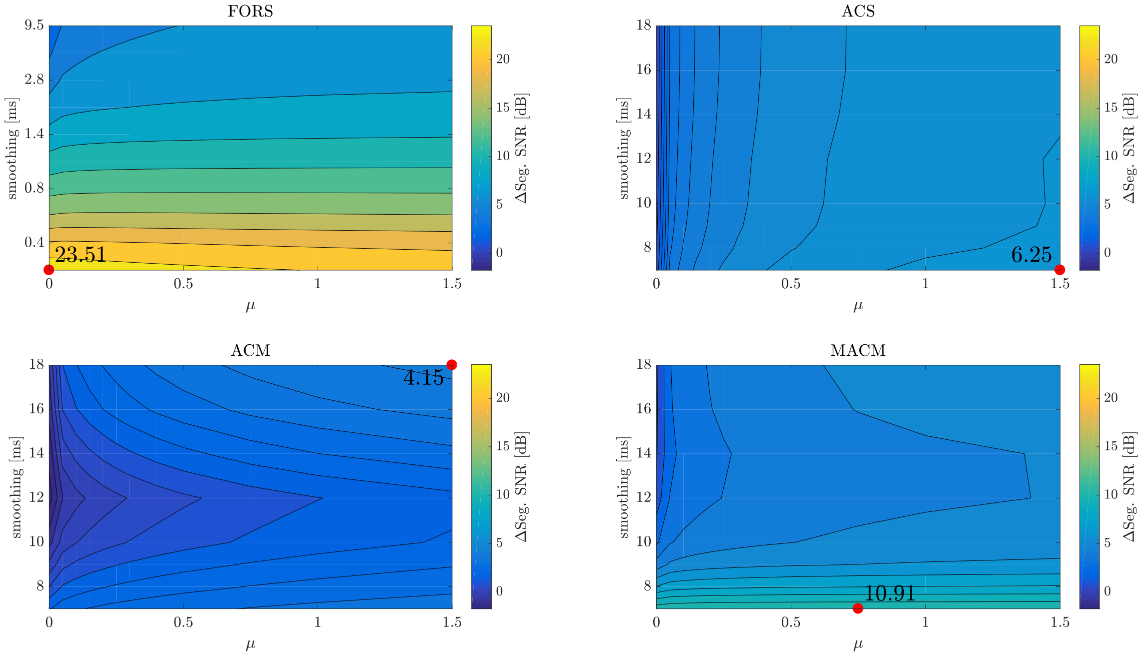

Fig. 4.13. ΔSeg. SNR scores for the SDW-IFWF-1 in the K−filterbank.

The FORS estimate with $\lambda = 0$ (no smoothing) performs the best in all tests. Although the STOI improvement shows similar scores for all tested values of μ, the PESQ and seg. SNR improvements clearly indicate maxima at $\mu = 0$. In the case of the FORS, ACS, and MACM estimates, the longer the smoothing, the worse the performance. The MACM performs slightly better than the ACS in all tests and the ACM performs the worst overall.

Figs. 4.14, 4.15, and 4.16 show the results of the implementations of the SDW-IFWF-X filter.

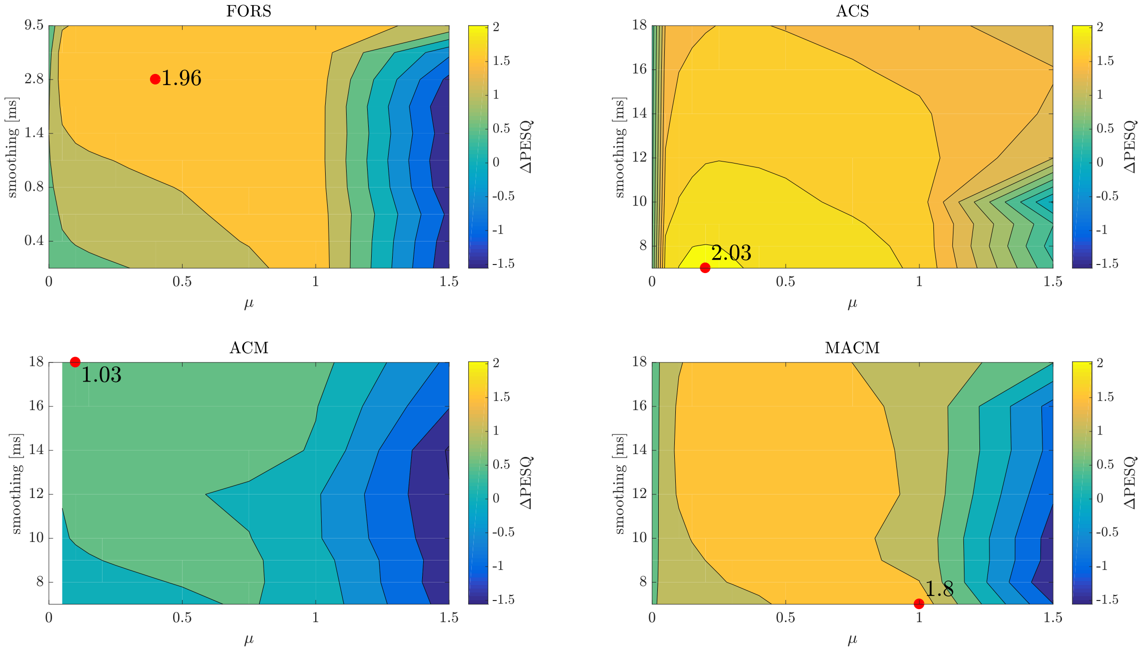

Fig. 4.14. ΔPESQ scores for the SDW-IFWF-X in the K−filterbank.

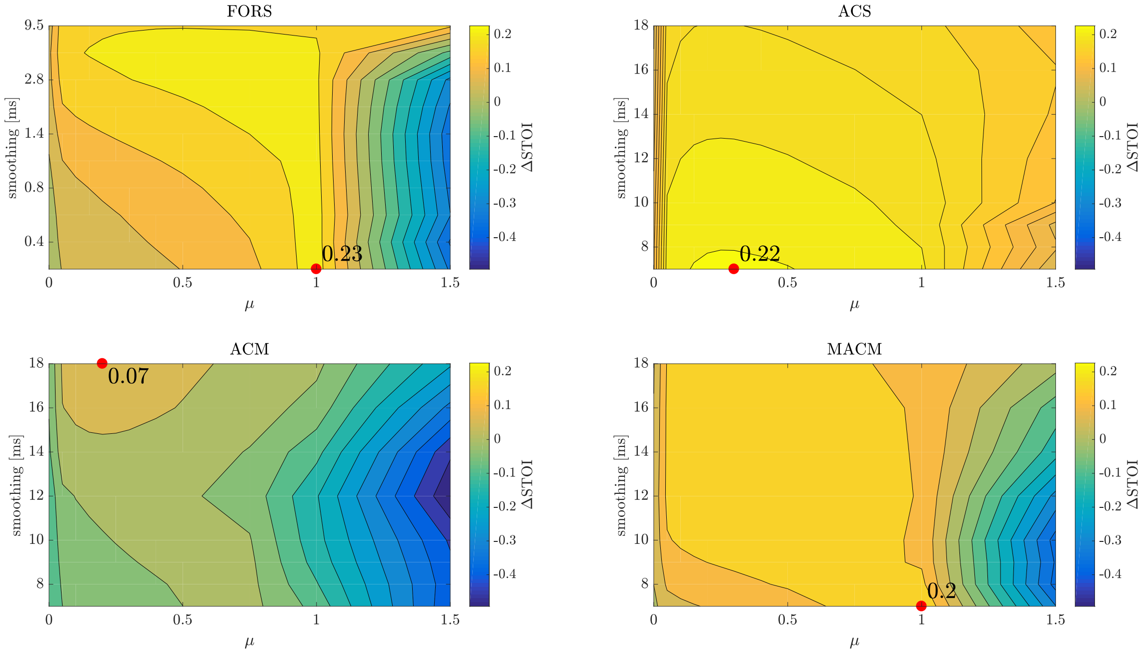

Fig. 4.15. ΔSTOI scores for the SDW-IFWF-X in the K−filterbank.

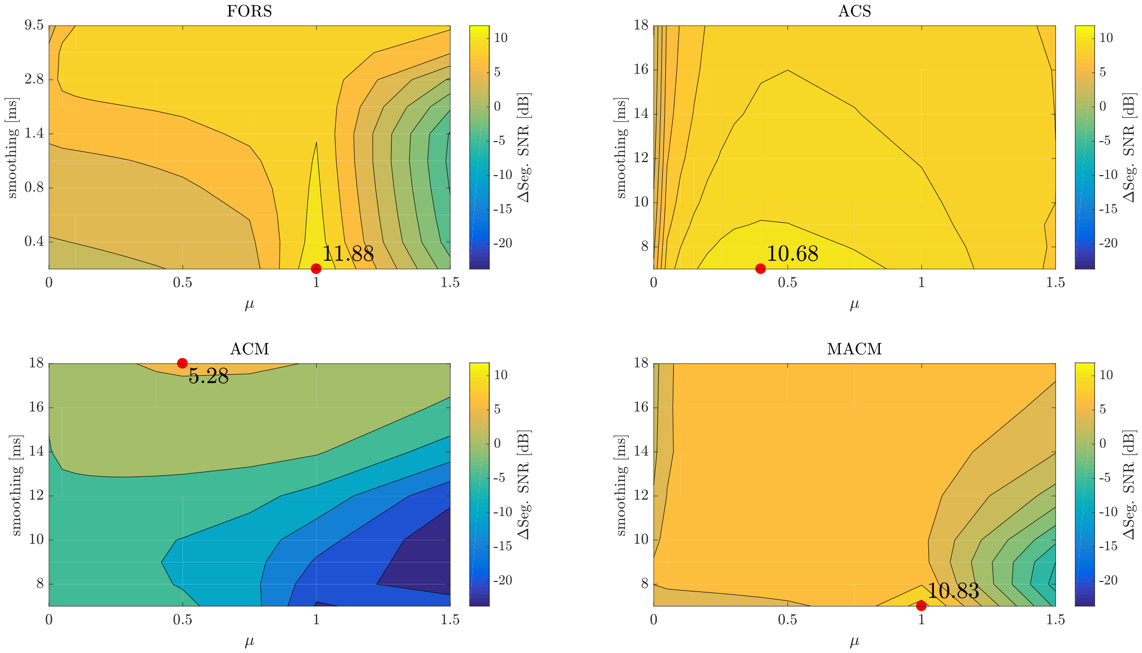

Fig. 4.16. ΔSeg. SNR scores for the SDW-IFWF-X in the K−filterbank.

In contrast to the SDW-IFWF and SDW-IFWF-1, the smoothing plays more of a role in the SDW-IFWF-X. The ACS, FORS, and MACF IFC estimates provide similar levels of maximal performance with respect to all scores. In the case of the FORS and MACM, the best performing value of the tradeoff parameter is $\mu = 1$, however, lower values of μ perform better with longer smoothing. In the case of the ACS, $\mu = 0.3$ produces the best overall results, where longer smoothing shows a clear decline in performance. The ACM IFC estimate delivers poor PESQ, STOI, and seg. SNR reduction in all tests.

The best performing filters of this section are the SDW-IFWF-1 using FORS with $\lambda = 0$ (no smoothing), closely followed by the SDW-IFWF using either FORS with $\lambda = 0$ (no smoothing) or MACM with $\rho = 7$ (7ms smoothing). The SDW-IFWF-X is the only implementable variation under blind conditions and produced good results for a wide range of smoothing lengths and using the FORS, ACS, and MACM IFC estimates.

4.1.3 Multi-Frame Filters Using Circulant IFC Assumption

In this section, the three SDW-IFWF-Cs in Section 3.3 were tested in the F-filterbank, where the PSD was estimated using either the periodogram or the Welch PSD as shown in Table 4.

Table 4: SDW-IFWF-C filters to be tested and corresponding PSD coefficient estimation methods.

Filter Method	Quantities		PSD Diagonal Matrix Estimation Method	
SDW-IFWF-C	$\boldsymbol{\Phi}^s_{k,l}$	$\boldsymbol{\Phi}^v_{k,l}$	(2.17)	Welch PSD
SDW-IFWF-C1	$\boldsymbol{\Phi}^s_{k,l}$	$\boldsymbol{\Phi}^v_{k,l}$		
SDW-IFWF-CX	$\boldsymbol{\Phi}^s_{k,l}$	$\boldsymbol{\Phi}^x_{k,l}$		

The Periodogram in the F-filterbank was also used to estimate the diagonal matrices containing the PSD coefficients of the neighbouring frequency bands of k, which were transformed into IFC matrices using (2.32) and then used in the SDW-IFWFs from Section 3.2.2 in the $K-$filterbank, shown in Table 5.

Table 5: SDW-IFWF filters to be tested using the transformed IFC matrices.

Filter Method	Quantities		PSD Diagonal Matrix Estimation Method
SDW-IFWF	$\boldsymbol{\Phi}^s_{k,l}$	$\boldsymbol{\Phi}^v_{k,l}$	Transformed Periodogram (2.17) & (2.32)
SDW-IFWF-1	$\boldsymbol{\Phi}^s_{k,l}$	$\boldsymbol{\Phi}^v_{k,l}$	
SDW-IFWF-X	$\boldsymbol{\Phi}^s_{k,l}$	$\boldsymbol{\Phi}^x_{k,l}$	

In addition, the ACS was also used to estimate the IFC matrices in the K-filterbank, then transformed into diagonal PSD matrices and used in the SDW-IFWF-Cs, shown in Table 6.

Table 6: SDW-IFWF-C filters to be tested using the transformed diagonal matrices containing the PSD coefficients.

Filter Method	Quantities		IFC Matrix Estimation Method
SDW-IFWF-C	$\boldsymbol{R}^s_{k,l}$	$\boldsymbol{R}^v_{k,l}$	Transformed ACS (3.41) & (2.32)
SDW-IFWF-C1	$\boldsymbol{R}^s_{k,l}$	$\boldsymbol{R}^v_{k,l}$	
SDW-IFWF-CX	$\boldsymbol{R}^s_{k,l}$	$\boldsymbol{R}^x_{k,l}$	

Figs. 4.17, 4.18, and 4.19 show the results of the implementations of the SDW-IFWF filter and the SDW-IFWF-C described in Tables 4, 5, and 6.

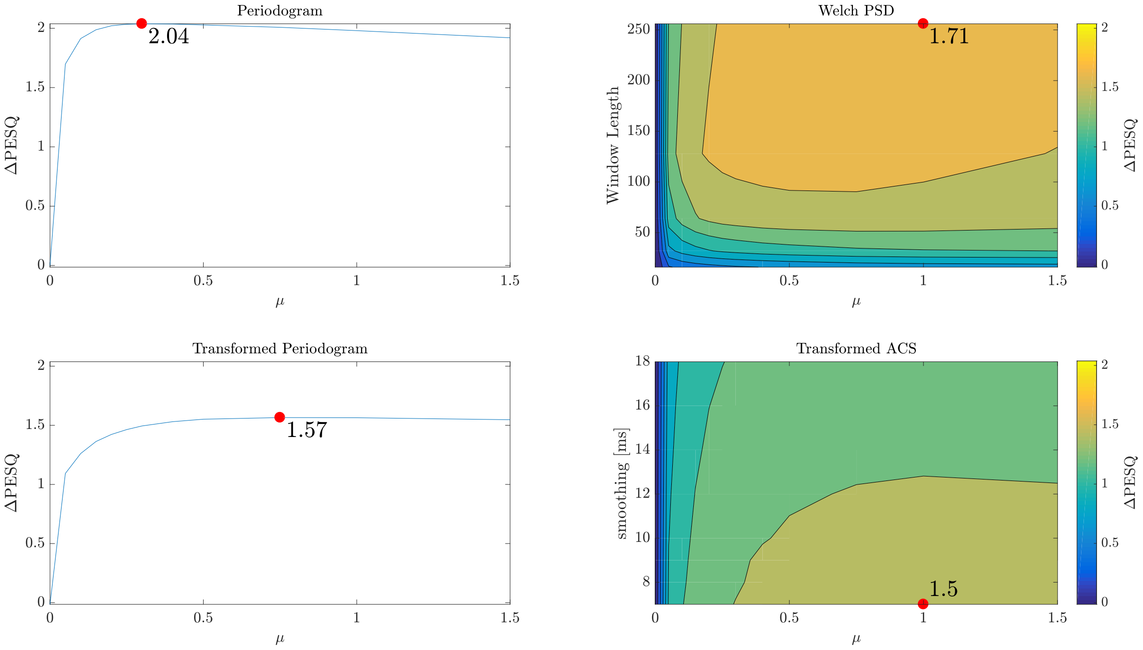

Fig. 4.17. ΔPESQ scores for the SDW-IFWF filter and the SDW-IFWF-C described in Tables 4, 5, and 6.

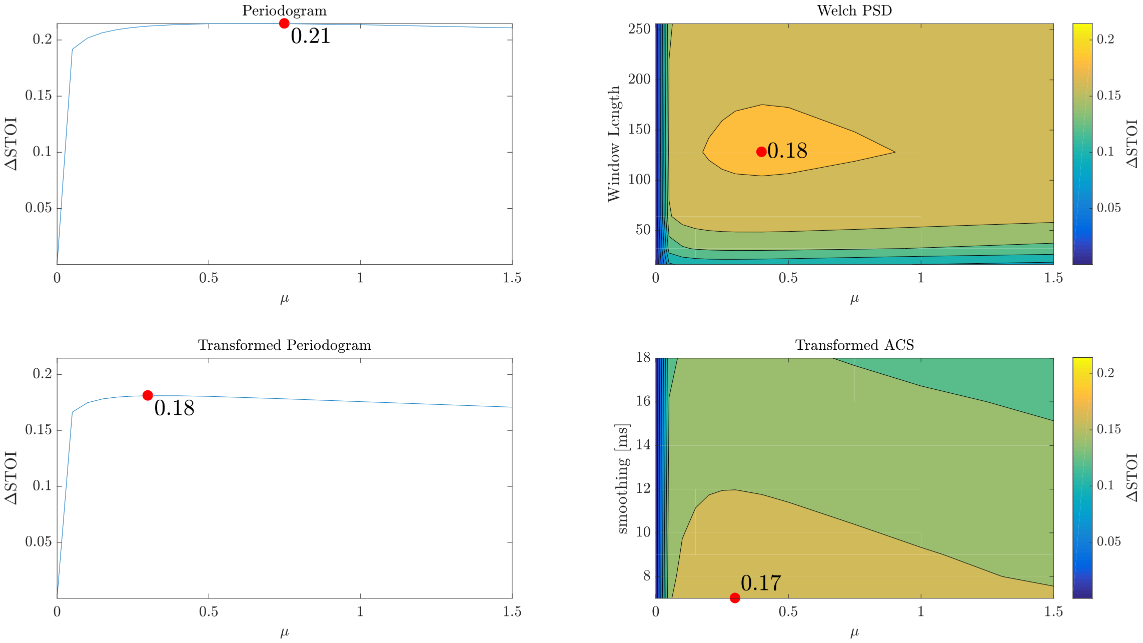

Fig. 4.18. ΔSTOI scores for the SDW-IFWF filter and the SDW-IFWF-C described in Tables 4, 5, and 6.

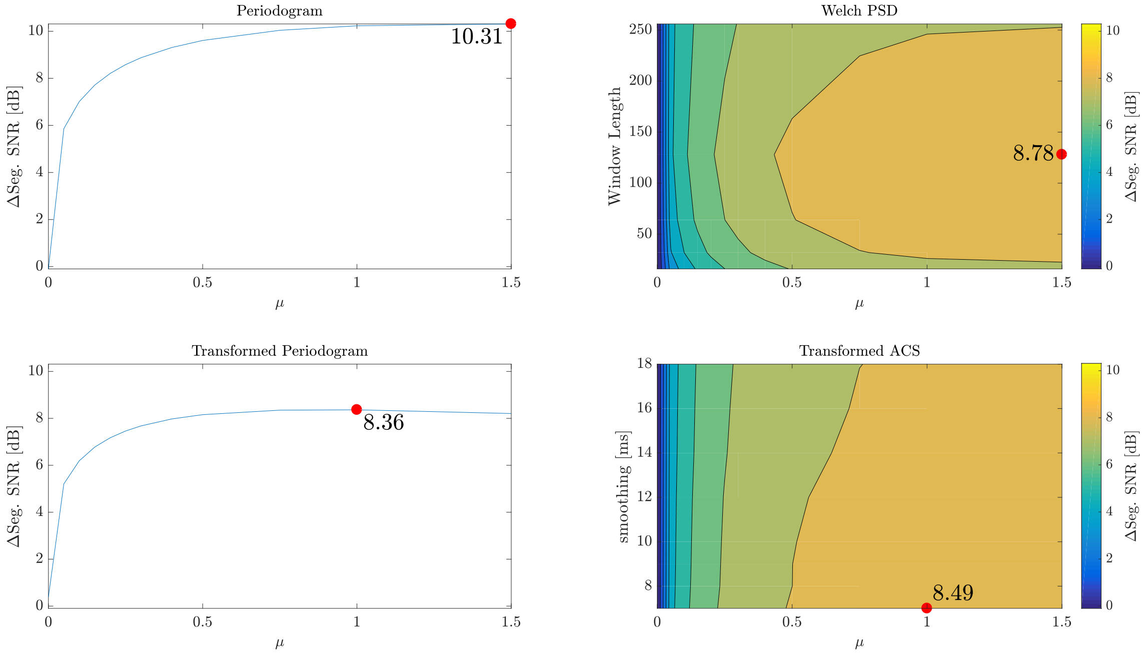

Fig. 4.19. ΔSeg. SNR scores for the SDW-IFWF filter and the SDW-IFWF-C described in Tables 4, 5, and 6.

The best performing SDW-IFWF-C in all tests was the periodogram. The PESQ improvement indicates that $\mu = 0.3$ produces the best score, while the best STOI and seg. SNR improvement are obtained at $\mu = 0.75$ and $\mu = 1.5$, respectively. As such, $\mu = 0.75$ is a good compromise, since the PESQ improvement declines noticeably for higher values while the STOI and seg. SNR improvements stay fairly even and the seg. SNR improvement declines quickly for lower values. The Welch PSD provides slightly worse results than the periodogram in all tests, with the best results produced by the window length 128 and $\mu = 0.75$. The transformed periodogram and transformed ACS also provide noise reduction and perform very similarly to eachother, but do not perform as well as the methods where the PSD or IFC matrices are estimated directly.

Figs. 4.20, 4.21, and 4.22 show the results of the implementations of the SDW-IFWF-1 filter and the SDW-IFWF-C1 described in Tables 4, 5, and 6.

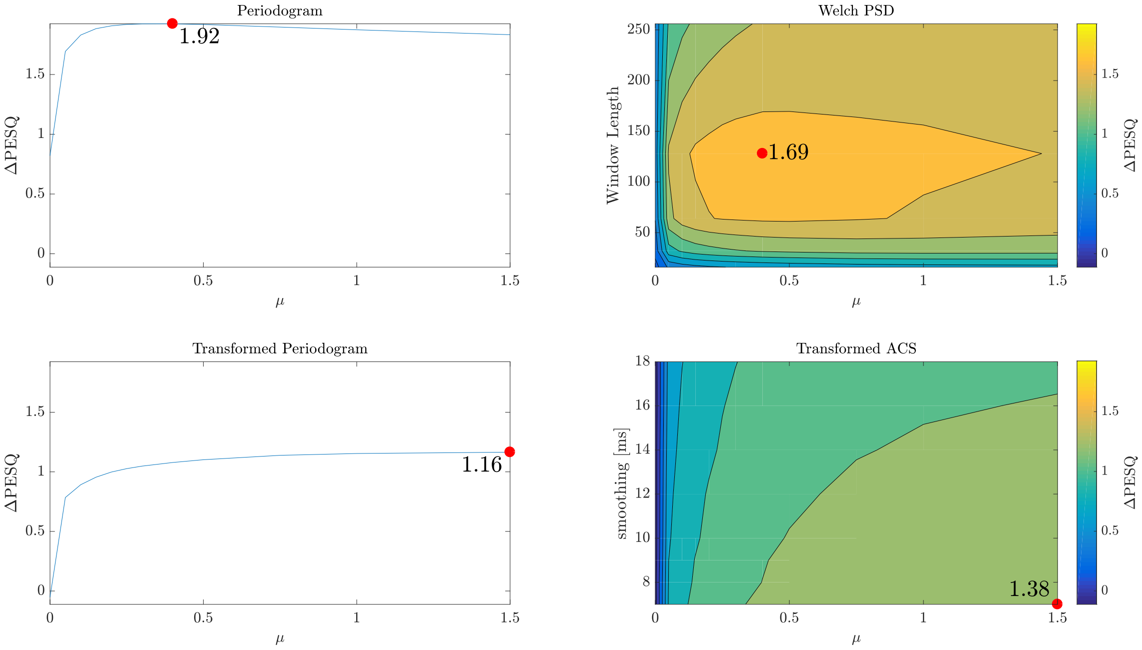

Fig. 4.20. ΔPESQ scores for the SDW-IFWF-1 filter and the SDW-IFWF-C1 described in Tables 4, 5, and 6.

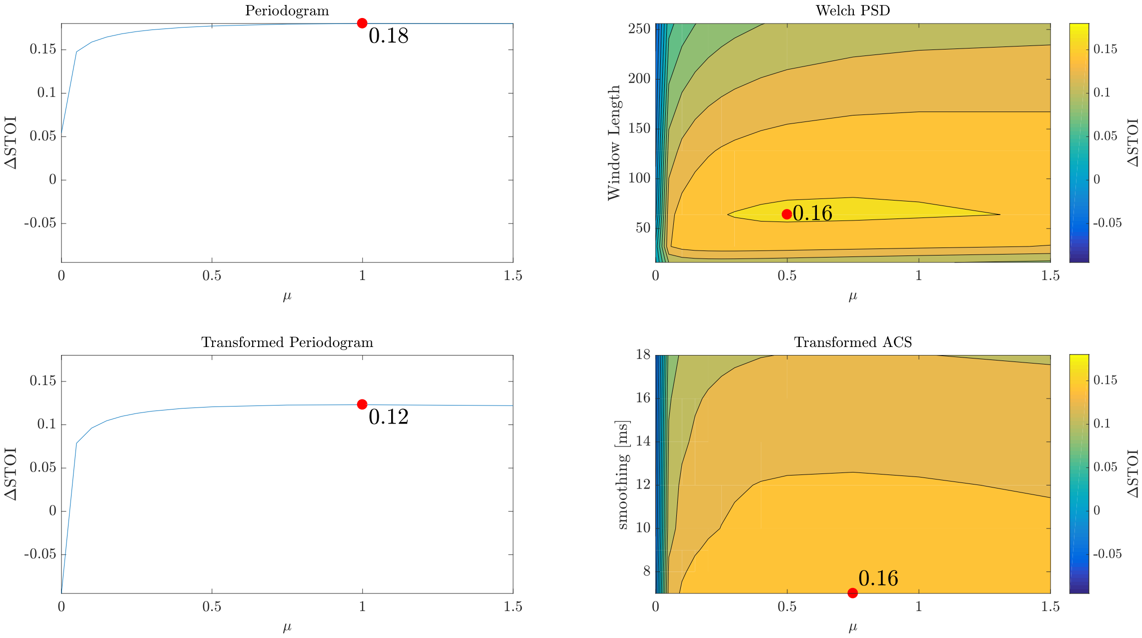

Fig. 4.21. ΔSTOI scores for the SDW-IFWF-1 filter and the SDW-IFWF-C1 described in Tables 4, 5, and 6.

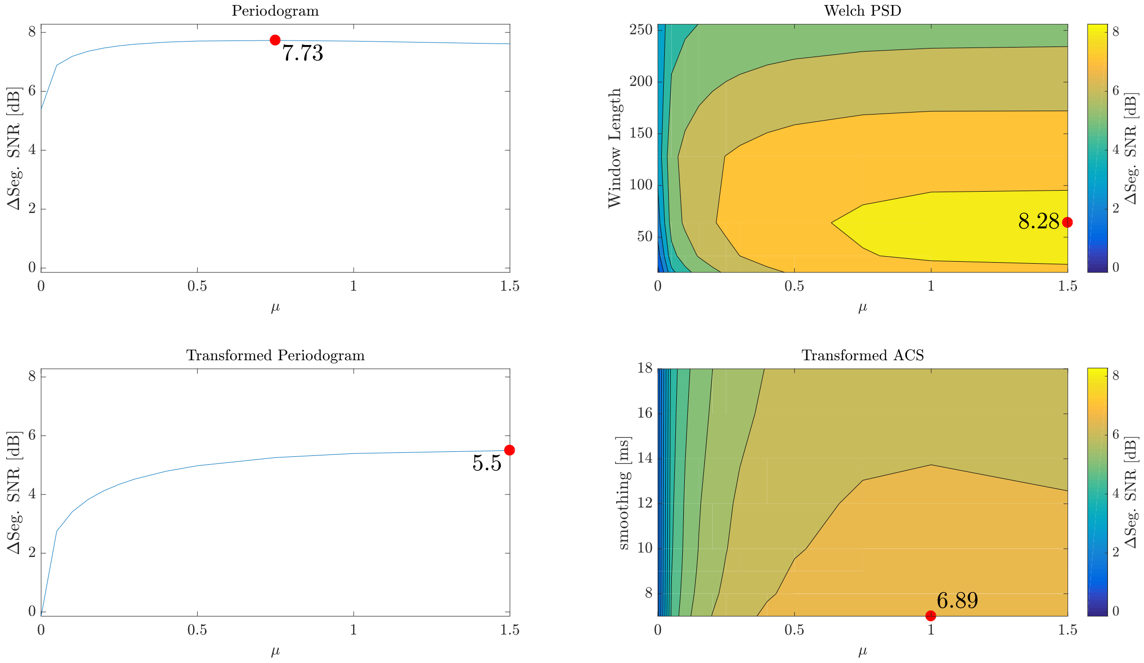

Fig. 4.22. ΔSeg. SNR scores for the SDW-IFWF-1 filter and the SDW-IFWF-C1 described in Tables 4, 5, and 6.

In the case of the SDW-IFWF-C1, the best PESQ, STOI, and seg. SNR improvements were produced by the periodogram. However, the overall performance was marginally worse than the SDW-IFWF-C in all tests, except for at $\mu = 0$, where the SDW-IFWF-C1 provides noise reduction while the SDW-IFWF-C provides no filtering. Nevertheless, taking into account all tests, the best performing value of μ in the case of the periodogram was 0.75. When using the transformed periodogram or transformed ACS, effectively no noise reduction occurs at $\mu = 0$. The Welch PSD estimate performs slightly worse than the periodogram and reaches its best performance with a window length of 64 and $\mu = 0.75$.

Figs. 4.23, 4.24, and 4.25 show the results of the SDW-IFWF-X filter and the SDW-IFWF-CX described in Tables 4, 5, and 6.

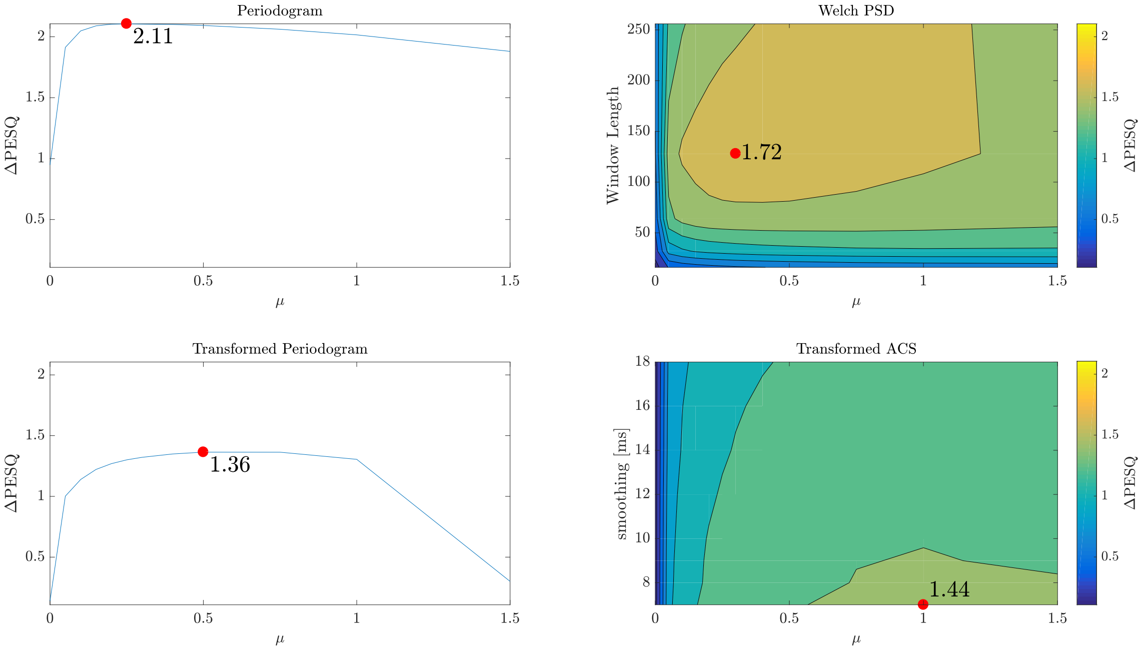

Fig. 4.23. ΔPESQ scores for the SDW-IFWF-X filter and the SDW-IFWF-CX described in Tables 4, 5, and 6.

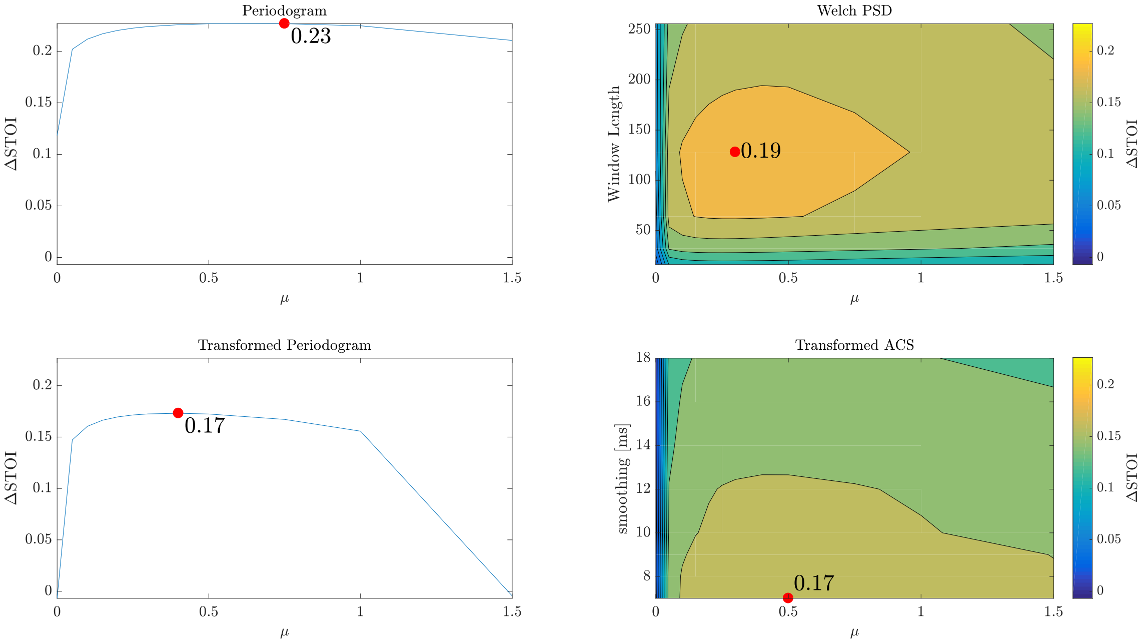

Fig. 4.24. ΔSTOI scores for the SDW-IFWF-X filter and the SDW-IFWF-CX described in Tables 4, 5, and 6.

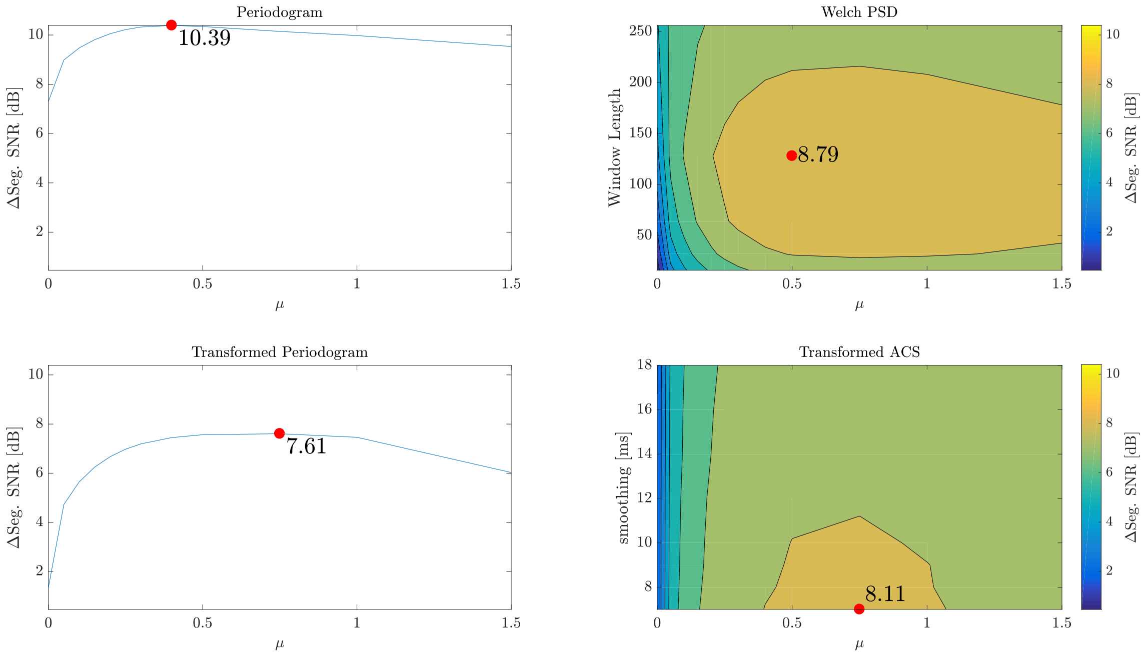

Fig. 4.25. ΔSeg. SNR scores for the SDW-IFWF-X filter and the SDW-IFWF-CX described in Tables 4, 5, and 6.

The periodogram provides the best PESQ, STOI, and seg. SNR improvements. Taking the three scores into account, the optimal value for μ is 0.4, since lower values produce a notable reduction in seg. SNR improvement while higher values produce a notable reduction in PESQ improvement. The Welch PSD estimate provides slightly worse scores than the periodogram and achieves its best performance using a window length of 128 and $\mu = 0.3$.

The SDW-IFWF-C and SDW-IFWF-CX using the periodogram provide the best maximal performance in all tests. Taking the PESQ, STOI and seg. SNR scores into account across all tested values of the speech distortion parameter μ, the SDW-IFWF-C performs better than the SDW-IFWF-CX for $\mu > 0.5$, however, the opposite is the case for $\mu < 0.5$. Looking at $\mu = 0$, the SDW-IFWF-X provides the most noise reduction in all tests, followed by the SDW-IFWF-1, and then the SDW-IFWF which provides no reduction.

Overall, the best performing filters in this section are the SDW-IFWF-C and the SDW-IFWF-CX using the periodogram, albeit the SDW-IFWF-C1 using the periodogram provides more noise reduction than the SDW-IFWF-C at $\mu = 0$. The goal of the implementations including the transformed IFC and PSD coefficients was not to obtain the best performing filter, but just to show that the circulant IFC matrix can be used to estimate the PSD coefficients in practice (and vice-verse), to show that the filters derived in Section 3.3 can also be applied in practice.

4.1.4 Comparison of Implementations

In this Section, the best performing implementations from Sections 4.1.1, 4.1.2, and 4.1.3 are cherry-picked and compared to get an overview of the best performance of these filters.

Fig. 4.26 shows the results of the best performing real-valued filter gains from Sections 4.1.1 and 4.1.3, in terms of PESQ, STOI, and seg. SNR improvement.

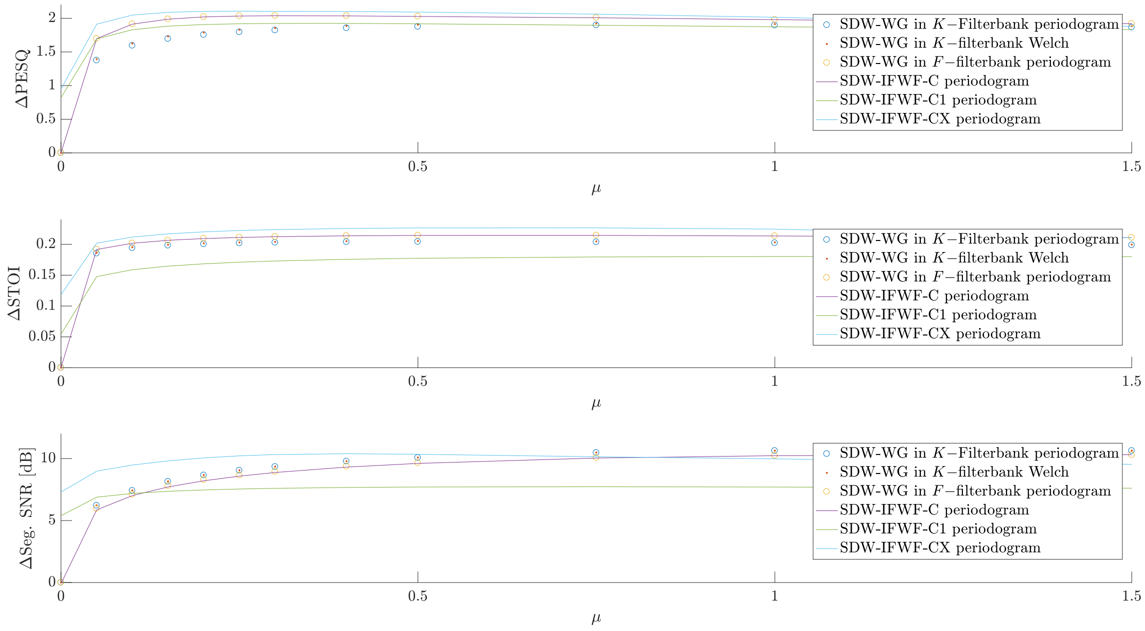

Fig. 4.26. Comparison of the scalar gain filter performance using the periodogram to estimate the PSD.

The SDW-WG in the F-filterbank produces very similar results to the SDW-IFWF-C. The SDW-IFWF-CX performs better than all other real-valued filter gains when the trade-off parameter $\mu < 1$, especially as $\mu = 0$, where it is the only filter gain which can apply noise reduction at this value. The filter performance is relatively even for higher values of μ, where the seg. SNR improvement of the SDW-IFWF-CX drops below other implementations for $\mu > 1$. The PESQ scores indicate that the filter performance is quite similar, with the exception of the real-valued gains in the K-filterbank, which perform notably worse for $\mu < 0.75$, however, the performance is similar when $\mu \geq 0.75$. The STOI and seg. SNR improvements show that the SDW-IFWF-C1 provides noise reduction at $\mu = 0$ unlike the SDW-WGs and the SDW-IFWF-Cs, however, it performs the worst out of all filters for higher values of

μ.

Fig. 4.27 shows the results of the best performing implementations of the SDW-IFWFs and SDW-IFWF-1 from Section 4.1.2 in terms of PESQ, STOI, and seg. SNR improvement. Neither of these filters is implementable under blind conditions.

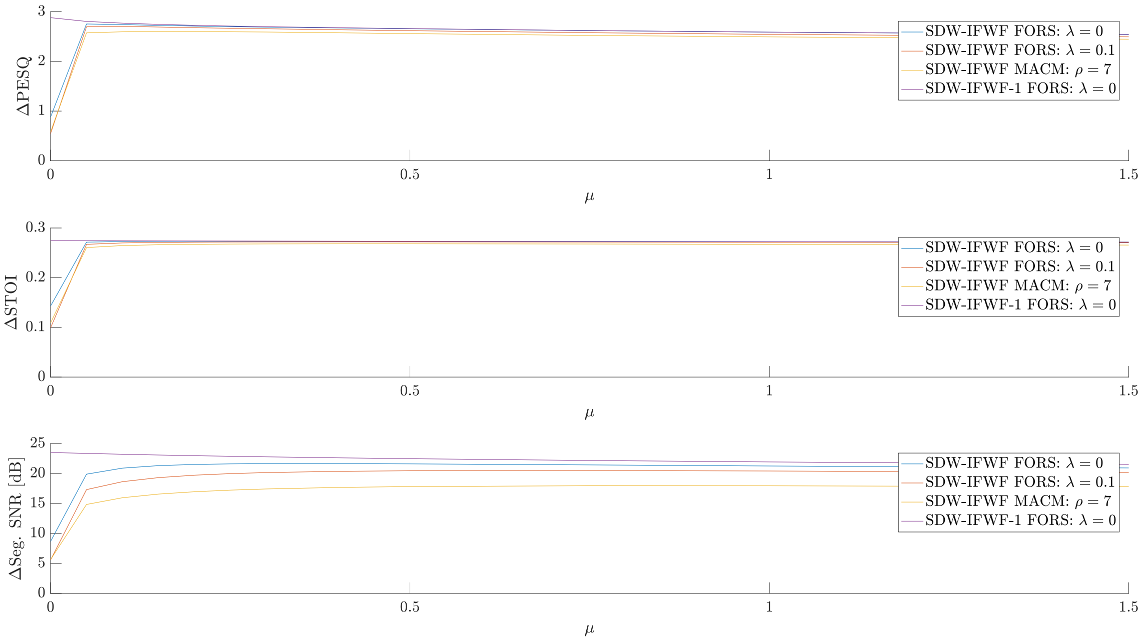

Fig. 4.27. Comparison of the best performing IFC matrix estimators in the SDW-IFWF.

The SDW-IFWF-1 using FORS with $\lambda = 0$ is the best filter in all tests. Out of the SDW-IFWFs, the FORS implementation with $\lambda = 0$ provides the best performance in terms of seg. SNR improvement. The performance of the filters in terms of PESQ and STOI, on the other hand, is very similar. Informal listening tests even showed that the MACM implementation sounded more natural and closer to the actual clean speech signal than the implementations using FORS.

Fig. 4.28 shows the results of the best performing implementations of the SDW-IFWF-Xs from Section 4.1.2, which can be implemented under blind conditions.

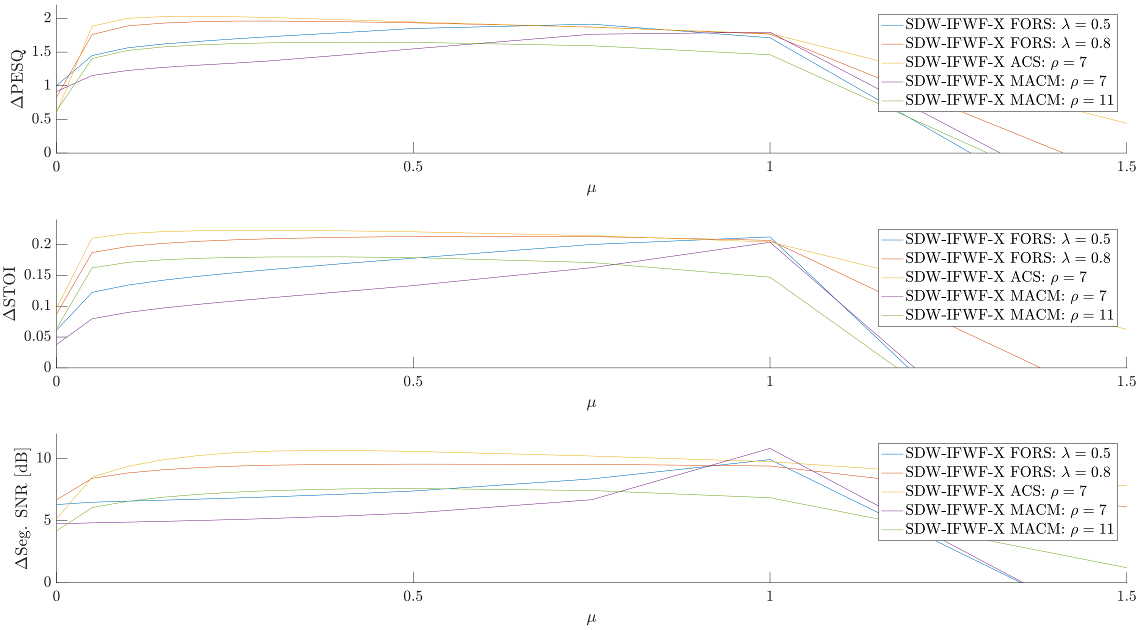

Fig. 4.28. Comparison of the best performing IFC matrix estimators in the SDW-IFWF-1 and SDW-IFWF-Xs.

Out of the SDW-IFWF-Xs, the ACS IFC matrix estimate with $\rho = 7$ (7 ms smoothing) performed the best overall in all tests, closely followed by FORS with $\lambda = 0.8$ (4.5 ms smoothing). The MACM with $\rho = 7$ (7 ms smoothing) and FORS with $\lambda = 0.5$ (1.4 ms smoothing) both peak at $\mu = 1$ in all tests, where they perform the best when taking into account the PESQ, STOI, and seg. SNR improvements together. However, they perform worse than the other IFC estimation methods for other values of μ.

Comparing the best performing SDW-IFWF-1 in Fig. 4.27 with the best performing SDW-IFWF-X in Fig. 4.28, the SDW-IFWF-1 provides significantly more noise reduction. This is surprising since in [27] it was shown that the MPDR (SDW-IFWF-X at $\mu = 0$) performed better than the MVDR (SDW-IFWF-1 at $\mu = 0$) under oracle conditions, however, only the recursive smoothing factor $\lambda = 0.88$ (7.88 ms smoothing) was taken into account. When looking at Figs. 4.11, 4.12, and 4.13, similar PESQ and seg. SNR improvements are produced as in [27] when looking at the scores which were achieved for $\lambda = 0.88$.

4.2 Influence of Oversampling Factor and IFC Length on Filter Performance

Here, the effect of increasing the frequency resolution compared to the K-filterbank was investigated using the oversampling factor O for the values 2, 4, 6, 8, 10. The

SDW-IFWF-C, SDW-IFWF-C1, and SDW-IFWF-CX from Section 3.3.2 were all tested under the assumption of oracle knowledge. In each case, the periodogram was used to estimate the PSD, since it was seen in the previous subsection that the periodogram produced the best results in the F-filterbank. Using the PSDs, the filter coefficients $\boldsymbol{W}$ were computed according to each variation of the real-valued multi-frame filter and overlapped into scalar gains $G_{f,l}$ using (2.35).

Although an increase in O results in a higher spectral resolution, and thus allowing for a better determination of the harmonic structure, it comes with the trade-off of a reduced temporal resolution due to the longer analysis windows. This temporal smearing as a result of increasing O can be seen in Fig. 4.29.

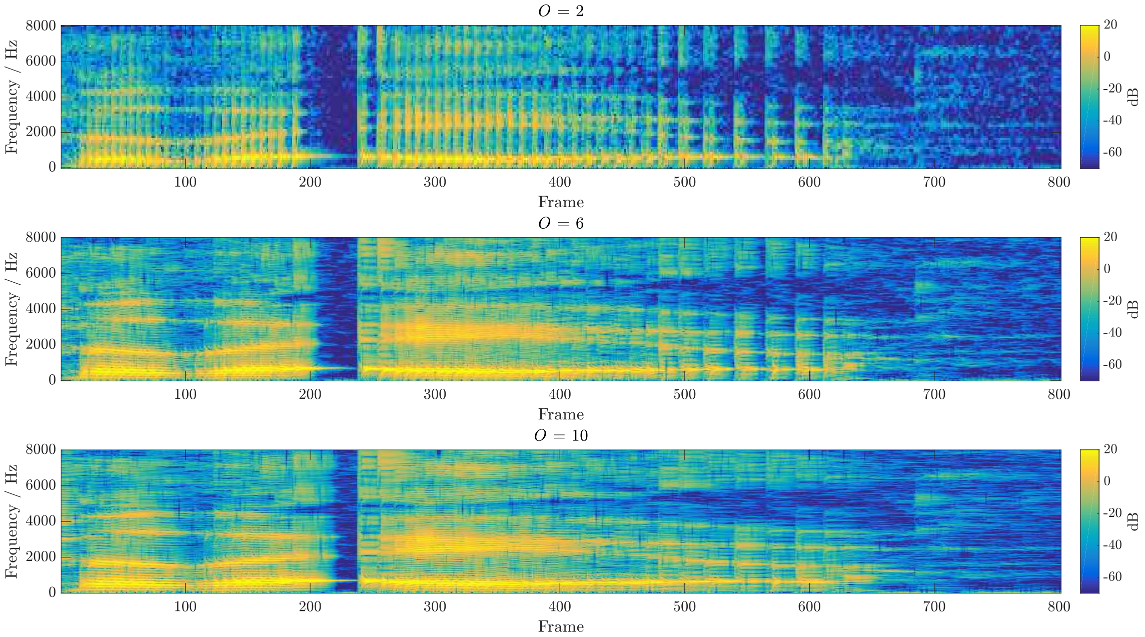

Fig. 4.29. Examples of the influence of the oversampling factor O on the enhanced speech spectrogram using the SDW-IFWF-CX for $\mu = 0.25$.

The results of the PESQ, STOI, and seg. SNR improvements are presented in Figs. 4.30, 4.31, and 4.32.

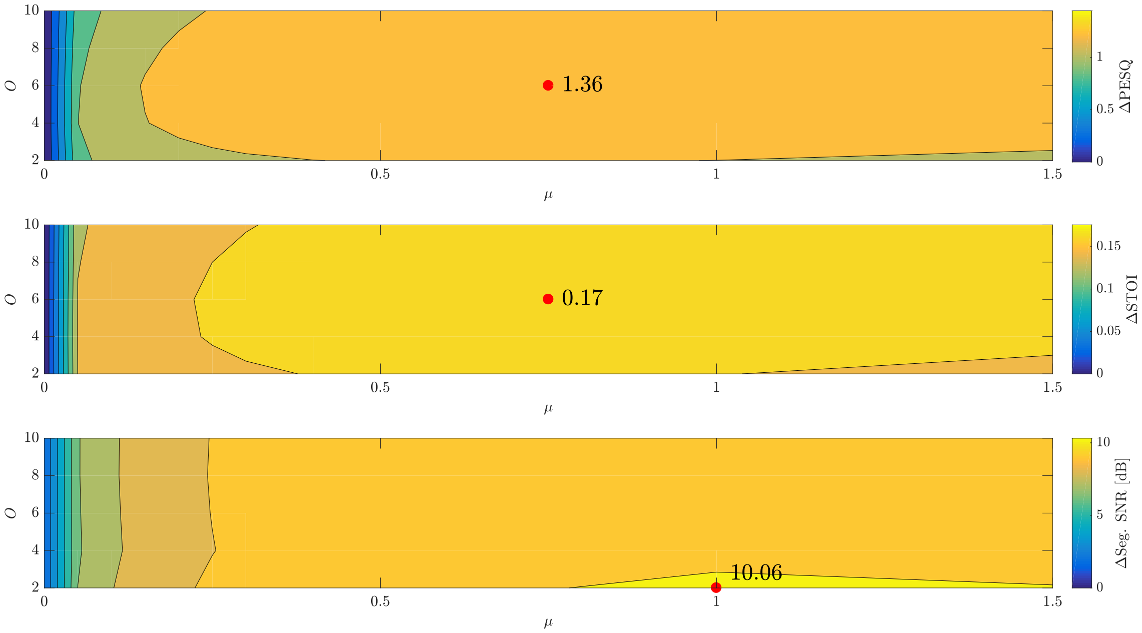

Fig. 4.30. SDW-IFWF-C tested for different values of O.

The seg. SNR improvement indicates that the most noise reduction occurs at $\mu = 1$ and $O = 2$. $\mu = 0.75$ and $O = 6$ produce the best PESQ and STOI improvements, therefore, it can be concluded that these are the best overall values for the SDW-IFWF-C under oracle conditions.

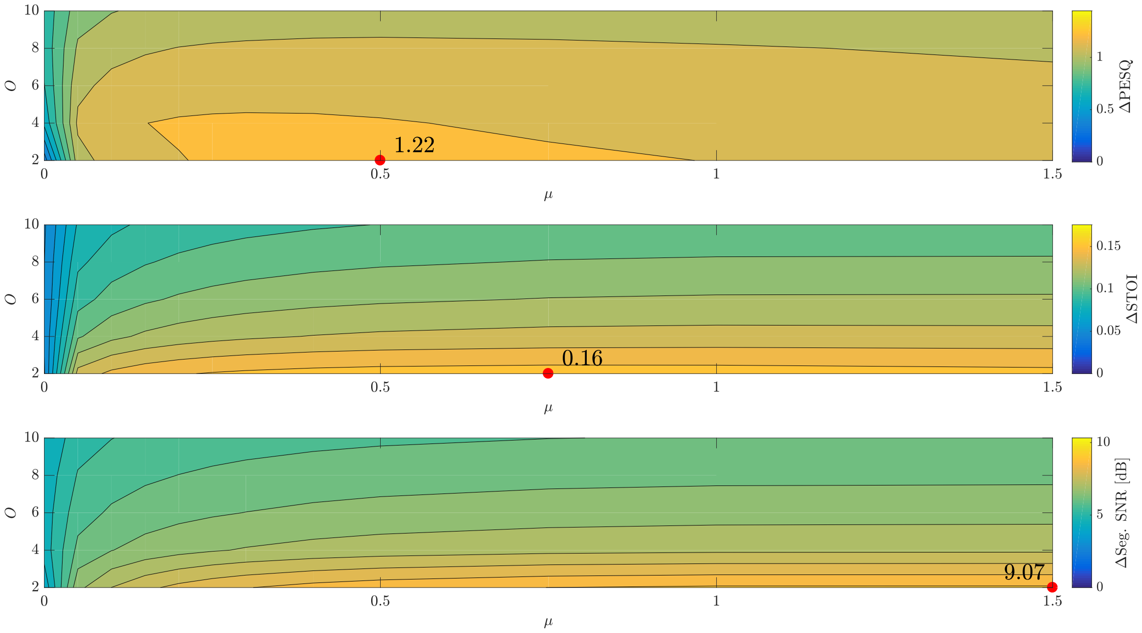

Fig. 4.31. SDW-IFWF-C1 tested for different values of O.

The SDW-IFWF-C1 performs worse than the SDW-IFWF-C in terms of PESQ, STOI, and seg. SNR improvement, however, its maximal performance at $O = 2$ and $\mu = 0.75$ is only slightly lower than the maximal performance of the SDW-IFWF-C. The higher the oversampling value O, the worse the overall performance is in all measures.

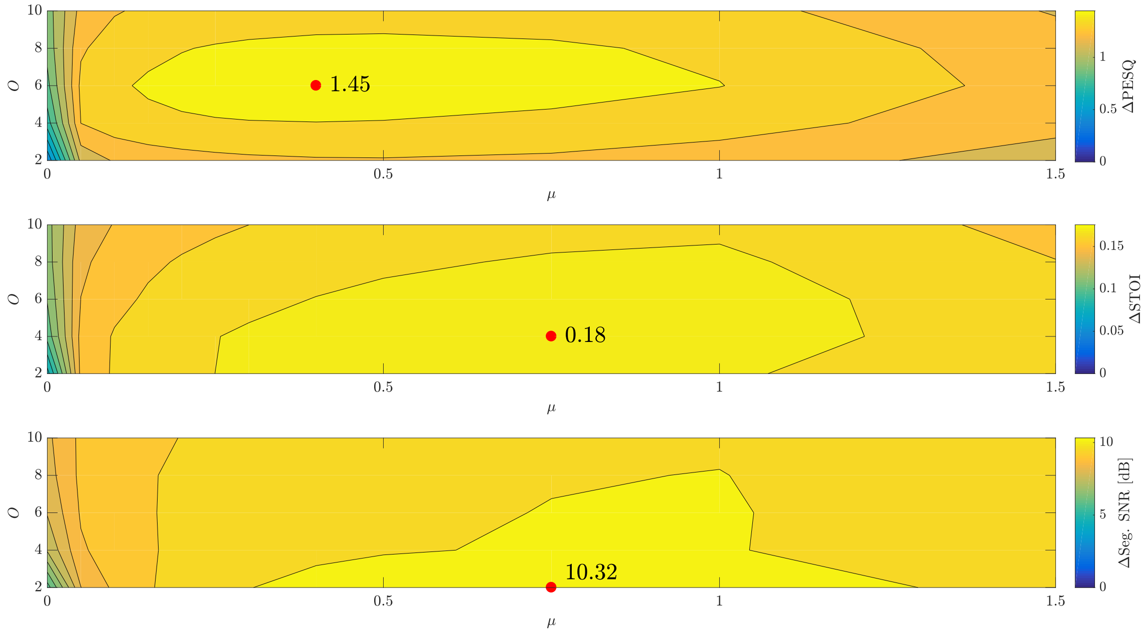

Fig. 4.32. SDW-IFWF-CX tested for different values of O.

For the SDW-IFWF-CX, $O = 6$ and $\mu = 0.4$ produce the highest PESQ improvement, however, $O = 4$ and $O = 2$ produce the best STOI and seg. SNR improvements, respectively, with $\mu = 0.75$. The best overall performance is, thus, obtained with $O = 6$ and $\mu = 0.75$ when taking into account all tests. At $\mu = 0$ (equivalent to IFMPDR-C), an increase in O results in a performance increase in all measures, where $O = 10$ produces the best performance.

After weighing up the PESQ, STOI, and seg. SNR improvements, $O = 6$ and $\mu = 0.75$ produced the best overall performance in the SDW-IFWF-C and SDW-IFWF-CX. Out of the real-valued multi-frame filters, the SDW-IFWF-CX produced both the best maximal and overall performance in all tests.

4.3 Blind Implementations of Multi-Frame Filters

In this part of the evaluation, the best performing implementable filters from the oracle evaluations in Sections 4.1 and 4.2 were tested under blind conditions (with only access to the noisy speech signal). The noise PSD was estimated directly using the noise power estimator from [13] and the speech PSD was estimated using (2.19) and (2.18). In the case of the real-valued multi-frame filters, the old speech estimate used in the a-priori SNR was only updated every 4 frames since this was found to be the optimal update rate which produced the least musical noise. In the case of the complex-valued filters in the K-filterbank, the old speech estimate was updated every frame.

The best performing IFC matrix estimators and filters from Section 4.1 were tested, and are shown in Table 7. Even though the SDW-IFWF and SDW-IFWF-1 performed well, they were not tested in the blind implementation since they required estimates of $\boldsymbol{R}^s_{k,l}$ and/or $\boldsymbol{R}^v_{k,l}$.

Table 7: Blind implementations with the corresponding IFC matrix or PSD coefficient estimation methods

Filter Method	**Quantities**		**IFC or PSD Diagonal Matrix Estimation Method**		
SDW-IFWF-X	$\boldsymbol{R}^x_{k,l}$	$\gamma^s_{k,l}, \phi^s_{k,l}$	FORS	ACS	MACM
SDW-IFWF-C	$\boldsymbol{\Phi}^s_{k,l}$	$\boldsymbol{\Phi}^v_{k,l}$	Periodogram		
SDW-IFWF-X-C	$\boldsymbol{\Phi}^s_{k,l}$	$\boldsymbol{\Phi}^x_{k,l}$			

Figs. 4.33, 4.34, 4.35, show the results of the SDW-IFWF-X implementations using the FORS, ACS, and MACM IFC estimates and the results of the SDW-IFWF-C and SDW-IFWF-CX using the periodogram can be seen in Figs. 4.36 and 4.37, respectively.

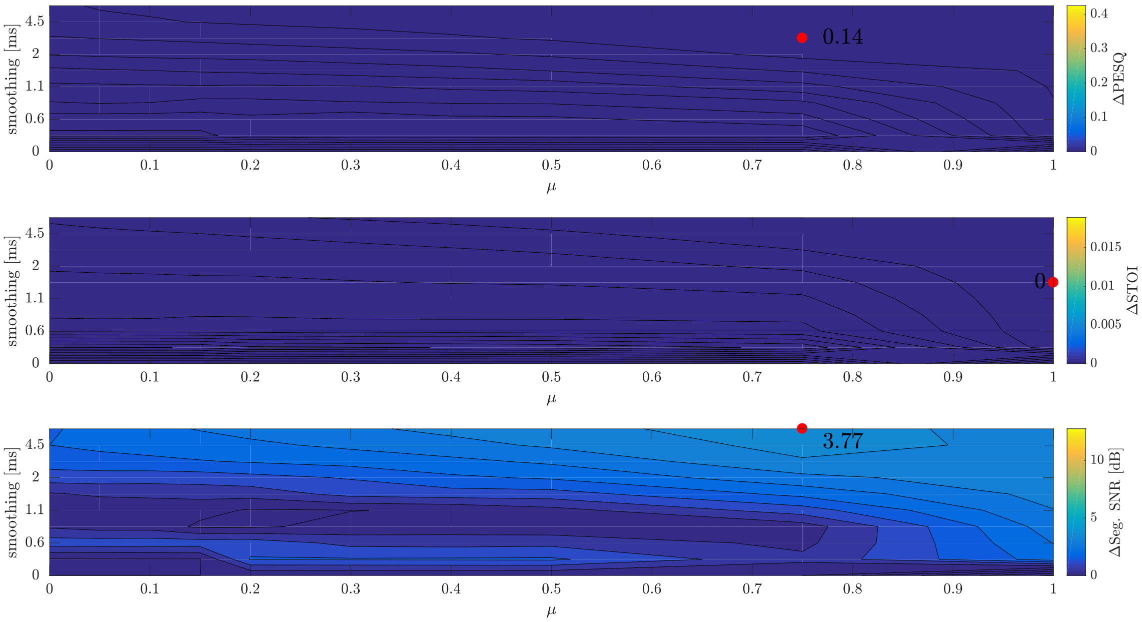

Fig. 4.33. Objective scores of the blind SDW-IFWF-X implementation with FORS IFC matrix estimation.

The blind SDW-IFWF-X using the FORS IFC estimate produced unstable results for all parameters across all tests, producing very low PESQ and seg. SNR improvements as well as negative STOI improvements, many artifacts could be heard in informal listening tests.

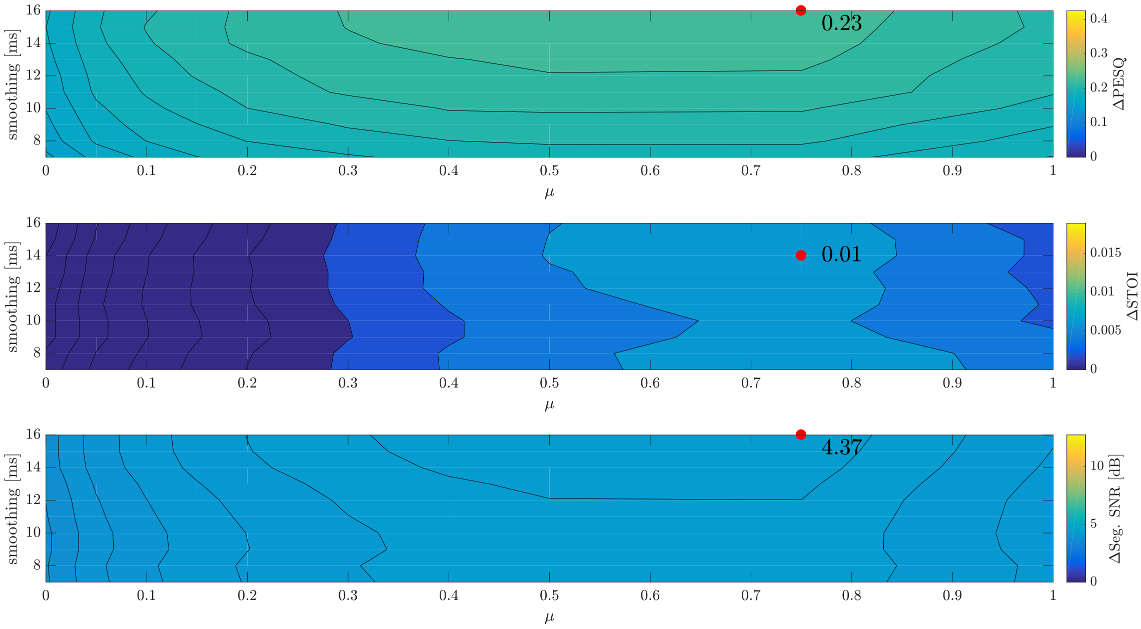

Fig. 4.34. Objective scores of the blind SDW-IFWF-X implementation with ACS IFC matrix estimation.

Taking into account the PESQ, STOI, and seg. SNR improvements, the SDW-IFWF-X using the ACS IFC estimate produced the best results for $\rho = 0.16$ (16 ms smoothing) and $\mu = 0.75$. A fraction of the noise reduction is apploed compared to the oracle oracle conditions, however, out of the blind SDW-IFWF-X, the ACS IFC estimate produced the best results, which was also reflected in informal listening tests.

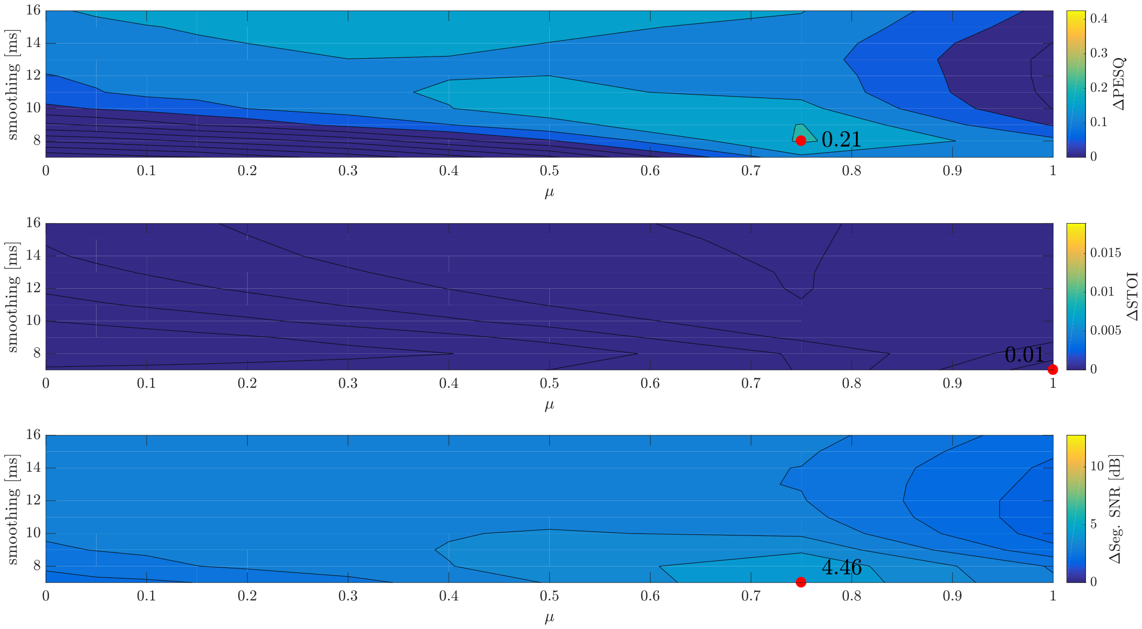

Fig. 4.35. Objective scores of the blind SDW-IFWF-X implementation with MACM IFC matrix estimation.

The MACM estimate produced robust results, however, in contrast to the ACS which favoured longer smoothing windows, the MACF also performed the most noise reduction at $\rho = 7$ (7 ms smoothing). Taking into account PESQ and seg. SNR improvements, the optimal combination is $\rho = 7$ (7 ms smoothing) and $\mu = 0.75$.

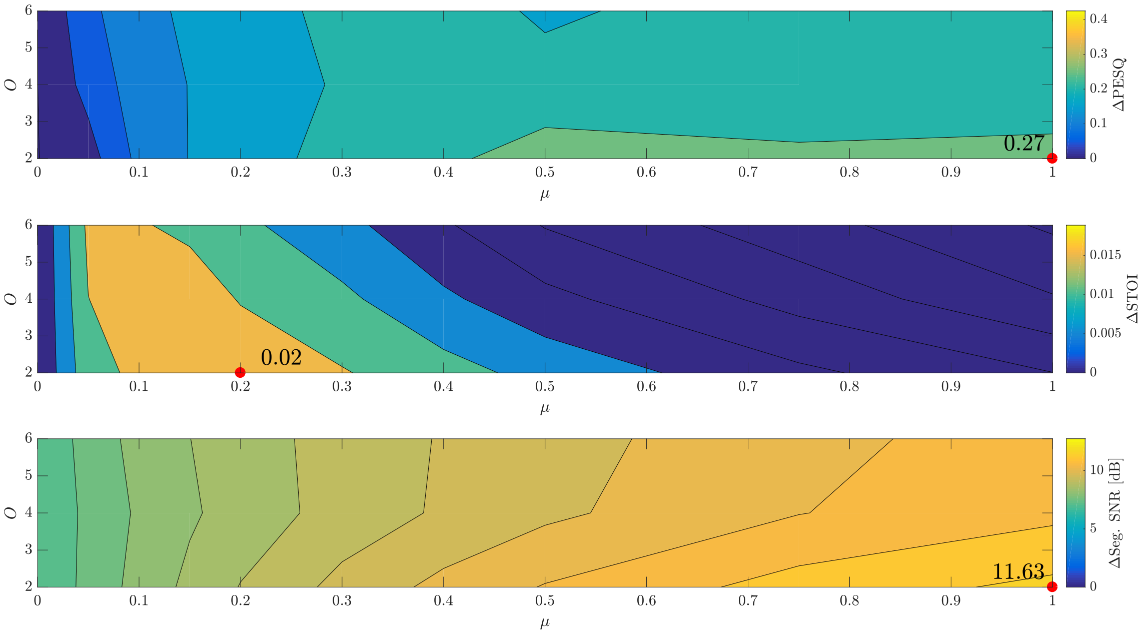

Fig. 4.36. Objective scores of the blind SDW-IFWF-C implementation using the periodogram.

The PESQ and seg. SNR improvements indicate the best performance of the SDW-IFWF-C when the tradeoff parameter $\mu = 1$ and the oversampling factor $O = 2$. The best STOI improvement contradicts the PESQ and seg. SNR scores in terms of which μ value produced the best performance and indicates the best performance at $\mu \approx 0.15$ for all values of the oversampling factor O. Taking into account all measures, the best performance is seen at $\mu = 1$ and $O = 2$.

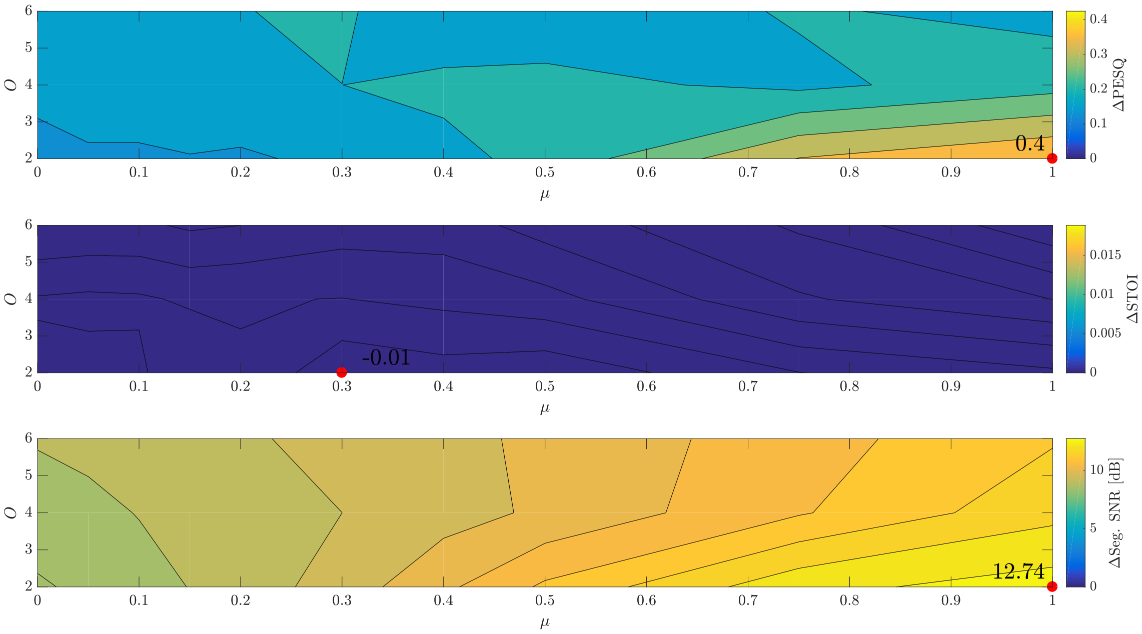

Fig. 4.37. Objective scores of the blind SDW-IFWF-CX implementation using the periodogram.

For the SDW-IFWF-CX, the best PESQ and seg. SNR improvements are produced at $\mu = 1$ and $O = 2$ as with the SDW-IFWF-C. Similar to the oracle implementation in Section 4.2, at $\mu = 0$ (equivalent to the IFMPDR-C) the SDW-IFWF-CX produces higher PESQ and seg. SNR improvement with increasing O. The SDW-IFWF-X implementations using the ACS and MACM IFC estimates as well as the SDW-IFWF-C and SDW-IFWF-CX implementations produced stable results throughout all tests of the blind implementations.

Overall, the STOI improvement measure seemed very sensitive to estimation errors in the blind implementations. Considering all blind implementations, the SDW-IFWF-X with the FORS IFC matrix estimation performed poorly due to introducing high levels of artifacts and distortion. The SDW-IFWF-X with the ACS IFC estimation produced the best results for $\rho = 16$ (16 ms smoothing) and $\mu = 0.75$. Using the MACM IFC estimate, the speech quality improvement and noise reduction was only slightly worse than the ACS, reflected by the PESQ and seg. SNR improvements, and produced the best results for $\mu = 0.75$ and $\rho = 7$ (7 ms smoothing). Informal listening tests also indicated that the ACS performed slightly better than the MACM.

The SDW-IFWF-CX produced the best results out of all of the blind implementations in terms of PESQ and seg. SNR improvement. The SDW-IFWF-C and SDW-IFWF-CX both provided similar noise reduction, with the SDW-IFWF-CX having a slight edge. Higher Δseg. SNR scores favour higher values of μ in the gain filters and increasing the oversampling factor O produces worse scores in terms of PESQ, STOI, and seg. SNR improvement except for when $\mu \leq 0.2$. Although the SDW-IFWF-Cs produced better PESQ and segmental SNR improvement scores than the SDW-IFWF-Xs, informal listening tests confirmed the SDW-IFWF-Cs introduced more musical noise than the SDW-IFWF-X.

5 Conclusions

In this thesis, speech quality and intelligibility improvement was investigated for a wide range of single-channel filters which can influence a trade-off between noise reduction and speech distortion. The filters can be divided into two groups, those within the single-frame model which assume speech and noise to be uncorrelated across consecutive time frames and those within the multi-frame model where the IFC is exploited to derive the optimal filter coefficients for noise reduction. Overall, the complex-valued multi-frame filters performed the best under oracle conditions (knowledge of all required quantities) and the real-valued multi-frame filters were the most effective under blind conditions (only the noisy speech signal available).

The influence of the frequency resolution of real-valued multi-frame filters was evaluated with the aim of improving their performance. When the periodogram was used to estimate the PSD with the oversampling factor $O = 6$ (which corresponds to a bandwidth of 42 Hz per frequency band), the real-valued multi-frame filters applied the most noise reduction under oracle conditions. Under blind conditions, a lower frequency resolution corresponding to $O = 2$ (equivalent to a bandwidth of 125 Hz per frequency band) yielded the best performance. A drawback of the real-valued multi-frame filters is that in most cases they introduced more musical noise than the robust implementations of complex-valued multi-frame filters.

The real-valued multi-frame filter gains were derived by assuming that the IFC matrices are Hermitian circulant structured, which matched or improved upon the performance of real-valued WGs within the single-frame model, especially at $\mu = 0$, where the WGs do not provide any noise reduction. It was shown that when $\mu = 0$, a real-valued rank 1 SDW-IFWF is equivalent to a real-valued MVDR gain and that the real-valued multi-frame IFWF can be decomposed into an MPDR gain multiplied by a WG.

The MACM was proposed as a method to estimate the IFC matrix and produced promising results in the complex-valued SDW-IFWF filter under oracle conditions. The objective speech quality and intelligibility improvement matched that of the commonly used FORS method, in addition to producing more natural sounding speech estimates in informal listening tests. The complex-valued SDW-IFWF-1 using an instantaneous IFC matrix estimate (with no smoothing applied) performed the best out of all implementations under oracle conditions. The ACS and the proposed MACM method to estimate the IFC matrices produced robust results in the complex-valued multi-frame filters under blind conditions, unlike the FORS which produced good results only under oracle conditions.

Future research could explore the performance of different IFC matrix estimators/-parameters within one filter implementation. Since speech and noise are different in terms of stationarity, the performance of the oracle implementations can surely be improved by choosing different smoothing parameters for each IFC matrix estimate of speech and noise or noisy speech.

A References

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

[2] Y. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1256–1269, May 2012.

[3] J. Benesty and Y. Huang, "A single-channel noise reduction MVDR filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011, pp. 273–276.

[4] D. Fischer and T. Gerkmann, "Single-microphone speech enhancement using MVDR filtering and Wiener post-filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 201–205.

[5] A. Schasse and R. Martin, "Estimation of subband speech correlations for noise reduction via MVDR processing," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 9, pp. 1355–1365, Sep. 2014.

[6] K. T. Andersen and M. Moonen, "Robust speech-distortion weighted interframe Wiener filters for single-channel noise reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 26, no. 1, pp. 97–107, Jan. 2018.

[7] B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1368–1381, July 2011.

[8] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367 – 2387, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168404002002

[9] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction," *Speech Communication*, vol. 49, no. 7, pp. 636 – 656, 2007, speech Enhancement. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639307000313

[10] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen, “Incorporating the conditional speech presence probability in multi-channel wiener filter based noise reduction in hearing aids,” *EURASIP J. Adv. Signal Process*, vol. 2009, pp. 7:1–7:11, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1155/2009/930625

[11] T. Gerkmann and R. C. Hendriks, “Unbiased mmse-based noise power estimation with low complexity and low tracking delay,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

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

[13] T. Gerkmann and R. C. Hendriks, “Noise power estimation based on the probability of speech presence,” in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2011, pp. 145–148.

[14] D. Mauler and R. Martin, “A low delay, variable resolution, perfect reconstruction spectral analysis-synthesis system for speech enhancement,” in *2007 15th European Signal Processing Conference*, Sept 2007, pp. 222–226.

[15] P. Welch, “The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms,” *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, June 1967.

[16] D. J. Thomson, “Spectrum estimation and harmonic analysis,” *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, Sept 1982.

[17] D. B. Percival and A. T. Walden, *Spectral analysis for physical applications : multitaper and conventional univariate techniques / Donald B. Percival and Andrew T. Walden.* Cambridge University Press Cambridge ; New York, N.Y., U.S.A, 1993. [Online]. Available: http://www.loc.gov/catdir/toc/cam021/92045862.html

[18] 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.

[19] J. S. Lim and A. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” vol. 67, pp. 1586 – 1604, 01 1980.

[20] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, April 1979.

[21] R. M. Gray, “Toeplitz and circulant matrices: A review,” *Foundations and Trends in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006. [Online]. Available: http://dx.doi.org/10.1561/0100000006

[22] S. L. Marple, *Digital Spectral Analysis: With Applications.* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1986.

[23] “ITU-T recommendation P.862. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” Feb. 2001.

[24] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 4214–4217.

[25] J. S. Garofolo, “DARPA TIMIT acoustic-phonetic speech database,” in *National Institute of Standards and Technology (NIST)*, 1988.

[26] A. Varga and H. J. M. Steeneken, “Assessment for automatic speech recognition ii: Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993. [Online]. Available: http://dx.doi.org/10.1016/0167-6393(93)90095-3

[27] D. Fischer and S. Doclo, “Sensitivity analysis of the multi-frame mvdr filter for single-microphone speech enhancement,” in *2017 25th European Signal Processing Conference (EUSIPCO)*, Aug 2017, pp. 603–607.

Acknowledgements

Many thanks to all who helped and supported me with my Bachelor thesis.

I would also like to express my gratitude to Simon Doclo for offering me this opportunity to help me develop my research skills and providing me with a lovely office and equipment.

A very special thanks go to Dörte Fischer, who as well as offering a lot of support, provided very useful suggestions and constructive criticism to my work. This kind of supervision was very much appreciated and is not easy to find as an undergraduate student!

Selbstständigkeitserklärung

Name: Brümann

Vorname: Klaus

Hiermit versichere ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Außerdem versichere ich, dass ich die allgemeinen Prinzipien wissenschaftlicher Arbeit und Veröffentlichung, wie sie in den Leitlinien guter wissenschaftlicher Praxis der Carl von Ossietzky Universität Oldenburg festgelegt sind, befolgt habe

Oldenburg, 19.11.2018 Unterschrift Studierende/r: Kl Brümann