This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2761699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

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

Kristian Timm Andersen, *Student Member, IEEE,* and Marc Moonen, *Fellow, IEEE,*

***Abstract*—In this paper, speech-distortion weighted (SDW) inter-frame Wiener filters (IFWFs) are investigated for single-channel noise reduction in a filter bank structure. The filters utilize a parameter μ that explicitly sets a trade-off between noise reduction and speech distortion and have traditionally been used in multi-channel applications under the term speech-distortion weighted multichannel Wiener filter (SDW-MWF). The application of these SDW-IFWFs relies on the estimation of inter-frame correlation (IFC) coefficients and it is shown that the IFC coefficients can be more robustly estimated using a secondary higher resolution filter bank (HRFB). It is then shown how real-valued scalar gains, which are optimal in the primary filter bank, can be applied directly in the HRFB instead of the inter-frame filtering in the primary filter bank, which leads to a more robust noise reduction performance for any value of μ. Computing these gains is also cheaper since matrix inversions are avoided and the primary filter bank is not needed in the actual implementation. Experimental results are given that support the claims, where the proposed methods are compared to relevant reference methods using measures such as the segmental SNR and the objective PESQ.**

I. Introduction

Noise reduction is an integral part of any real-time speech processing device. The speech signal is then usually processed through a filter bank such as a weighted overlap-add (WOLA) filter bank [1] where a real-valued scalar gain is applied to each subband signal. In communication devices, where real-time performance is important, the processing can be based on a filter bank with a suitably low delay. Low-delay filter banks are usually constrained in the frequency resolution that can be obtained. This is because the effective length of the analysis and synthesis window is limited by the delay. In such cases, instead of a scalar gain, a higher-order filter can be used in each subband to improve the noise reduction, e.g. a multiframe single-channel Wiener filter or minimum-variance distortionless-response (MVDR) filter [2], [3], [4].

One of the major limitations of noise reduction is that it introduces speech distortion, which can have a negative impact on the speech quality and intelligibility. In theory [2], speech distortion can be avoided in the single-channel case using a single-channel MVDR filter. In this paper we investigate higher-order subband filters in the framework of the speech-distortion weighted (SDW) inter-frame Wiener filter (IFWF) [5], [6], [7], [8]. The filters utilize a parameter μ that explicitly sets a trade-off between noise reduction and speech distortion and have traditionally been used in multi-channel applications under the term speech-distortion weighted multi-channel Wiener filter (SDW-MWF). We derive a number of different SDW-IFWFs from SDW-MWF theory [2], [3], [4] and show that they correspond to a type of rank-1 SDW-IFWF.

A crucial part in the performance of SDW-IFWFs is the estimation of the inter-frame correlation (IFC) coefficients. Usually, these IFC coefficients are estimated using first order recursions. In practice, however, SDW-IFWFs are very sensitive to estimation errors in the IFC coefficients and previous studies that have evaluated the filters using the true IFC coefficients have lead to unrealistically good performance. This has been addressed in [9], where a statistical analysis is carried out to estimate the noise and speech IFC coefficients using both a purely data-driven method and a *maximum-a-posteriori* (MAP) method. It has been shown that the MAP method, which relies on a model trained on offline data, improves the performance of the MVDR filter. The Wiener filter has also been considered, but it is commented that it is not as robust as the MVDR filter and consequently the evaluation has only focused on the MVDR filter.

In this paper we propose to estimate the IFC coefficients in a secondary higher resolution filter bank (HRFB) and investigate the robustness of the SDW-IFWFs using this approach. We use the term HRFB in the sense that the subbands in this filter bank have a smaller bandwidth than the subbands in the primary filter bank in which the noise reduction is defined. First it is shown that the IFC estimation in the HRFB is related to the IFC in the primary filter bank and can therefore be converted to SDW-IFWFs in the primary filter bank. Then, it is shown how real-valued scalar gains can be applied directly in the HRFB instead of the inter-frame filtering in the primary filter bank, which leads to a more robust noise reduction performance for any value of μ. It also means that the primary filter bank is not needed in the actual implementation. This approach has a number of advantages compared to the original SDW-IFWFs. First, the SDW-IFWFs are now more robust since a scalar gain can easily be limited below unity. Second, it is shown to improve the noise reduction in terms of both SNR and objective speech quality. Finally, it also leads to a lower computational complexity with a similar delay since it avoids the inversion of IFC matrices. An experimental evaluation is performed where results are given to support the claim that the proposed methods lead to more robust SDW-IFWFs compared to those obtained with the previously proposed methods[2], [9]. We investigate the noise reduction performance for a wide range of μ values, which includes both the MVDR and Wiener

This work was supported in part by The Danish Agency for Science, Technology, and Innovation and conducted in collaboration between KU Leuven and Widex A/S through the industrial Ph.D. program (case number 13-135472).

The authors are with the Department of Electrical Engineering (ESAT - STADIUS), KU Leuven, Leuven, Belgium (e-mail: kristian@esat.kuleuven.be; Marc.Moonen@esat.kuleuven.be). K.T. Andersen is also with Widex A/S, Lynge, Denmark.

2329-9290 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2761699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

filter. The performance is evaluated using the segmental signal-to-noise ratio (SNR), the objective PESQ measure and the speech-distortion index.

The main contribution of the paper is the proposed scalar SDW-IFWF gains in a HRFB, matching a criterion in the primary lower-resolution filter bank. The idea of estimating the IFC in a HRFB has also not been investigated previously. The resulting filter consists of a Wiener gain cascaded with a scaling factor to control the speech distortion. In this way it bears similarities with cascaded filter bank structures, e.g. [10] which have previously been proposed for single channel noise reduction. The cascaded filter bank allows for a flexible filter design, but the proposed methods in this paper effectively work in a single filter bank and allows for a very simple noise reduction and filter bank design.

The paper is structured as follows. In Section II, the noise reduction problem is formulated using a WOLA [1] filter bank. In Section III, the SDW-IFWFs are derived from SDW-MWF theory. The HRFB-based methods are presented in Section IV. An experimental evaluation is performed in Section V and a conclusion is given in Section VI.

II. Background and Problem Statement

The noisy speech signal $x[n]$ at time n is assumed to consist of a speech signal $s[n]$ degraded by additive noise $v[n]$:

$$x[n] = s[n] + v[n] \tag{1}$$

The signal $x[n]$ is given as input to a primary filter bank that splits the signal into K subband signals:

$$X[k,i] = S[k,i] + V[k,i], \quad k = -\frac{K}{2}+1, ..., \frac{K}{2} \tag{2}$$

where i is the downsampled time index (i.e. frame index). We assume that the filter bank is a WOLA [1] filter bank, so that each analysis filter is constructed from the same FIR lowpass analysis prototype filter $h_K[n]$ with length K :

$$X[k,i] = \sum_{n=-\infty}^{\infty} h_K[-n]x[n+iR]e^{-2\pi j(n+iR)k/K} \tag{3}$$

where R is the downsampling factor. This can be efficiently implemented using a fast Fourier transform (FFT) of size K. For each subband k, assuming $h_K[n]$ is an ideal lowpass filter, we can approximate the z-transform of $X[k,i]$ as [11]:

$$X_k(z) \approx \frac{1}{R} H_K\big(z^{1/R}\big) X\big(z^{1/R} e^{2\pi jk/K}\big) \tag{4}$$

where $X_k(z)$, $H_K(z)$ and $X(z)$ is the z-transform of $X[k,i]$, $h_K[n]$ and $x[n]$ respectively.

The signal $X[k,i]$ for frames $i, ..., i-M+1$ can be put into a length M noisy speech vector:

$$\boldsymbol{x}[k,i] = [X[k,i] \dots X[k,i-M+1]]^T \tag{5}$$

where ^{T} denotes the transpose operator. The speech vector $\boldsymbol{s}[k,i]$ and noise vector $\boldsymbol{v}[k,i]$ are similarly defined.

The noise reduction calculates an estimate of the speech $\hat{S}[k,i]$ from $\boldsymbol{x}[k,i]$:

$$\hat{S}[k,i] = \boldsymbol{w}^H[k,i]\boldsymbol{x}[k,i] \tag{6}$$

where ^{H} is the Hermitian transpose operator and $\boldsymbol{w}[k,i]$ is the time-varying noise reduction filter. The use of a filter $\boldsymbol{w}[k,i]$ of order $M-1>0$ makes it possible to improve the noise reduction compared to just using a scalar gain in each subband. In particular, it has been shown [2], [3] that single-channel MVDR filters can be derived that in theory provide noise reduction with no speech distortion. The derivation of SDW-IFWFs, which contain MVDR filters as a special case and provide an explicit trade-off between noise reduction and speech distortion, will be given in Section III. Each subband k is treated independently of other subbands and the speech-distortion weighting is therefore also independent of other subbands.

In Section IV-A we show that the filters $\boldsymbol{w}[k,i]$ can be more robustly estimated in a secondary HRFB with subband signals:

$$X[l,i] = S[l,i] + V[l,i], \quad l = -\frac{L}{2}+1, ..., \frac{L}{2} \tag{7}$$

where i is again the frame index corresponding to the same downsampling factor R. As before, we assume that the filter bank is a WOLA filter bank so that each analysis filter is constructed from the same FIR lowpass analysis prototype filter $h_L[n]$ with length L. For each subband l, the z-transform of $X[l,i]$ can then be written in a similar way:

$$X_l(z) \approx \frac{1}{R} H_L\big(z^{1/R}\big) X\big(z^{1/R} e^{2\pi jl/L}\big) \tag{8}$$

where $H_L(z)$ is the z-transform of $h_L[n]$. Evaluating the z-transforms on the unit circle, the higher frequency resolution in $X_l(z)$ implies that $h_L[n]$ has a more narrow lowpass frequency response than $h_K[n]$ and that $L > K$. It has been shown that WOLA filter banks with longer prototype filters, such as $h_L[n]$, can still be synthesized with the same low delay as $h_K[n]$, see for instance [12], [13], [14].

The HRFB gives an improved frequency resolution representation of $x[n]$ under the assumption that the signal is stationary over L samples. We use the term *wide-sense stationary* to indicate that the second order statistics of $x[n]$ are constant over L samples. By considering the higher frequency resolution in (8) compared to (4) we propose that the filters $\boldsymbol{w}[k,i]$ in (6), which are derived from statistical parameters in the primary filter bank defining $X[k,i]$, can be converted into high resolution scalar gains $g[l,i]$ in the secondary HRFB that are then applied to $X[l,i]$:

$$\hat{S}[l,i] = g[l,i]X[l,i] \tag{9}$$

In Section IV-B, we show that under certain conditions this is indeed the case. It is then experimentally shown that these filters are more robust and lead to improved performance under realistic test conditions compared to previously proposed filters. Since the resulting gains are calculated in the HRFB it also turns out that the primary filter bank (3) is not needed in the actual implementation and is only used to define the noise reduction problem. This, together with the scalar gains, results in a much cheaper solution. In the rest of the paper, we mostly omit the indices k and i for matrices and vectors, but it should be remembered that each subband k in $X[k,i]$ is treated independently.

2329-9290 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2761699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

III. Inter-frame Wiener filters

In this section, filters that take the IFC in $X[k,i]$ into account are derived. The idea of applying an IFWF in each subband of a filter bank has been proposed in [3] and the resulting filters are already found in the given references, but the derivation in this section follows the theory from SDW-MWFs. This is necessary to express the filters as a function of IFC matrices, which will be used in the rest of the paper.

The IFWF is found by minimizing the mean-square error (MSE):

$$J_{\text{IFWF}} = \mathbf{E}\{|\boldsymbol{w}^H\boldsymbol{x} - s_1|^2\} \tag{10}$$

where $\mathbf{E}\{\}$ is the expectation operator, $s_1 = \boldsymbol{e}_1^H\boldsymbol{s}$ is the desired speech signal and $\boldsymbol{e}_1$ is an all-zero vector except for a one in the first position.

The solution to (10) is given as:

$$\boldsymbol{w}_{\text{IFWF}} = \boldsymbol{R}_x^{-1}\boldsymbol{R}_s\boldsymbol{e}_1 = (\boldsymbol{R}_s + \boldsymbol{R}_v)^{-1}\boldsymbol{R}_s\boldsymbol{e}_1 \tag{11}$$

where

$$\boldsymbol{R}_x = \mathbf{E}\{\boldsymbol{x}\boldsymbol{x}^H\} \tag{12}$$

$$\boldsymbol{R}_s = \mathbf{E}\{\boldsymbol{s}\boldsymbol{s}^H\} \tag{13}$$

$$\boldsymbol{R}_v = \mathbf{E}\{\boldsymbol{v}\boldsymbol{v}^H\} \tag{14}$$

are the noisy speech, speech and noise IFC matrix respectively. The IFC matrices $\boldsymbol{R}_x$ and $\boldsymbol{R}_v$ can be estimated during speech-plus-noise and noise-only periods respectively, which requires a voice-activity detector (VAD), while $\boldsymbol{R}_s$ can be found as:

$$\boldsymbol{R}_s = \boldsymbol{R}_x - \boldsymbol{R}_v \tag{15}$$

• **SDW-IFWF**

The SDW-MWF has been proposed as a way to provide a trade-off between noise reduction and speech distortion [5], [6], [7], [8] and used as an IFWF it corresponds to minimizing the following MSE:

$$J_{\text{SDW-IFWF}} = \mathbf{E}\{|\boldsymbol{w}^H\boldsymbol{s} - s_1|^2\} + \mu\mathbf{E}\{|\boldsymbol{w}^H\boldsymbol{v}|^2\} \tag{16}$$

where $\mu \geq 0$ is the trade-off parameter. The solution is given as:

$$\boldsymbol{w}_{\text{SDW-IFWF}} = (\boldsymbol{R}_s + \mu\boldsymbol{R}_v)^{-1}\boldsymbol{R}_s\boldsymbol{e}_1 \tag{17}$$

When $\boldsymbol{R}_s$ is a full rank matrix, $\mu = 0$ results in no filtering:

$$\boldsymbol{w}_{\text{SDW-IFWF}\,(\mu=0)} = \boldsymbol{e}_1 \tag{18}$$

• **SDW-IFWF-1**

Since the desired speech signal in (16) is s_1 rather than the full speech vector $\boldsymbol{s}$, the speech IFC is fully determined by a speech IFC vector[1]:

$$\boldsymbol{\gamma}_s = \frac{\mathbf{E}\{\boldsymbol{s}s_1^H\}}{\mathbf{E}\{s_1s_1^H\}} = \frac{\boldsymbol{R}_s\boldsymbol{e}_1}{\sigma_s} \tag{19}$$

where σ_s is the first element in the speech IFC matrix. The noisy speech and noise IFC vectors are similarly defined.

As in [2], [3], [4], the signal model is now written as a linear function of the desired speech signal:

$$\boldsymbol{x} = \boldsymbol{\gamma}_s s_1 + \boldsymbol{v} \tag{20}$$

[1]This is a slight deviation from previous papers where $\boldsymbol{\gamma}_s$ was defined from the first row rather than column.

From this, it is seen that the speech IFC matrix is a rank-1 matrix:

$$\boldsymbol{R}_{s1} = \mathbf{E}\{\boldsymbol{\gamma}_s s_1(\boldsymbol{\gamma}_s s_1)^H\} = \boldsymbol{\gamma}_s\boldsymbol{\gamma}_s^H\sigma_s \tag{21}$$

Using this signal model in (16) corresponds to replacing $\boldsymbol{R}_s$ with $\boldsymbol{R}_{s1}$ in (17) and applying the matrix inversion lemma gives [15], [16]:

$$\begin{aligned}\boldsymbol{w}_{\text{SDW-IFWF-1}} &= \frac{\boldsymbol{R}_v^{-1}\boldsymbol{R}_s\boldsymbol{e}_1\boldsymbol{e}_1^H\boldsymbol{R}_s\boldsymbol{e}_1}{\mu\boldsymbol{e}_1^H\boldsymbol{R}_s\boldsymbol{e}_1 + \text{Tr}\{\boldsymbol{R}_v^{-1}\boldsymbol{R}_s\boldsymbol{e}_1\boldsymbol{e}_1^H\boldsymbol{R}_s\}} \\ &= \frac{\boldsymbol{R}_v^{-1}\boldsymbol{\gamma}_s\sigma_s}{\mu + \boldsymbol{\gamma}_s^H\boldsymbol{R}_v^{-1}\boldsymbol{\gamma}_s\sigma_s}\end{aligned} \tag{22}$$

where $\text{Tr}\{\}$ is the trace operator. It is seen that setting $\mu = 0$ gives an MVDR filter:

$$\begin{aligned}\boldsymbol{w}_{\text{SDW-IFWF-1}\,(\mu=0)} &= \boldsymbol{R}_v^{-1}\boldsymbol{R}_s\boldsymbol{e}_1\frac{\boldsymbol{e}_1^H\boldsymbol{R}_s\boldsymbol{e}_1}{\text{Tr}\{\boldsymbol{R}_v^{-1}\boldsymbol{R}_s\boldsymbol{e}_1\boldsymbol{e}_1^H\boldsymbol{R}_s\}} \\ &= \frac{\boldsymbol{R}_v^{-1}\boldsymbol{\gamma}_s}{\boldsymbol{\gamma}_s^H\boldsymbol{R}_v^{-1}\boldsymbol{\gamma}_s} = \boldsymbol{w}_{\text{MVDR-1}}\end{aligned} \tag{23}$$

• **SDW-IFWF-X**

The filter $\boldsymbol{w}_{\text{SDW-IFWF-1}}$ uses the noise IFC matrix $\boldsymbol{R}_v$ and a rank-1 speech IFC matrix $\boldsymbol{R}_{s1}$. If $\boldsymbol{R}_v$ is estimated during noise-only periods, this means that $\boldsymbol{w}_{\text{SDW-IFWF-1}}$ is unable to react to the part of $\boldsymbol{R}_s$ that is not correlated with the desired speech signal s_1. An alternative filter that takes this into account can be found where $\boldsymbol{R}_v$ is re-estimated by subtracting the rank-1 matrix $\boldsymbol{R}_{s1}$ from $\boldsymbol{R}_x$:

$$\begin{aligned}\boldsymbol{w}_{\text{SDW-IFWF-X}} &= (\boldsymbol{R}_{s1} + \mu(\boldsymbol{R}_x - \boldsymbol{R}_{s1}))^{-1}\boldsymbol{R}_{s1}\boldsymbol{e}_1 \\ &= ((1-\mu)\boldsymbol{R}_{s1} + \mu\boldsymbol{R}_x)^{-1}\boldsymbol{R}_{s1}\boldsymbol{e}_1\end{aligned} \tag{24}$$

Using the matrix inversion lemma in (24) gives:

$$\begin{aligned}\boldsymbol{w}_{\text{SDW-IFWF-X}} &= \frac{\boldsymbol{R}_x^{-1}\boldsymbol{R}_s\boldsymbol{e}_1\boldsymbol{e}_1^H\boldsymbol{R}_s\boldsymbol{e}_1}{\mu\boldsymbol{e}_1^H\boldsymbol{R}_s\boldsymbol{e}_1 + (1-\mu)\text{Tr}\{\boldsymbol{R}_x^{-1}\boldsymbol{R}_s\boldsymbol{e}_1\boldsymbol{e}_1^H\boldsymbol{R}_s\}} \\ &= \frac{\boldsymbol{R}_x^{-1}\boldsymbol{\gamma}_s\sigma_s}{\mu + (1-\mu)\boldsymbol{\gamma}_s^H\boldsymbol{R}_x^{-1}\boldsymbol{\gamma}_s\sigma_s}\end{aligned} \tag{25}$$

An equivalent filter to $\boldsymbol{w}_{\text{SDW-IFWF-X}}$, called the "Tradeoff filter", is derived in [3] [4] using a Lagrange multiplier procedure in a constrained optimization problem. It is seen that setting $\mu = 1$ in (25) gives the IFWF in (11):

$$\boldsymbol{w}_{\text{SDW-IFWF-X}\,(\mu=1)} = \boldsymbol{R}_x^{-1}\boldsymbol{\gamma}_s\sigma_s = \boldsymbol{R}_x^{-1}\boldsymbol{R}_s\boldsymbol{e}_1 = \boldsymbol{w}_{\text{IFWF}} \tag{26}$$

and that setting $\mu = 0$ gives an MVDR filter:

$$\begin{aligned}\boldsymbol{w}_{\text{SDW-IFWF-X}\,(\mu=0)} &= \boldsymbol{R}_x^{-1}\boldsymbol{R}_s\boldsymbol{e}_1\frac{\boldsymbol{e}_1^H\boldsymbol{R}_s\boldsymbol{e}_1}{\text{Tr}\{\boldsymbol{R}_x^{-1}\boldsymbol{R}_s\boldsymbol{e}_1\boldsymbol{e}_1^H\boldsymbol{R}_s\}} \\ &= \frac{\boldsymbol{R}_x^{-1}\boldsymbol{\gamma}_s}{\boldsymbol{\gamma}_s^H\boldsymbol{R}_x^{-1}\boldsymbol{\gamma}_s} = \boldsymbol{w}_{\text{MVDR-X}}\end{aligned} \tag{27}$$

It is noted that the difference between $\boldsymbol{w}_{\text{SDW-IFWF}}$, $\boldsymbol{w}_{\text{SDW-IFWF-1}}$ and $\boldsymbol{w}_{\text{SDW-IFWF-X}}$ is how $\boldsymbol{R}_s$ is treated when it is not a rank-1 matrix [17], i.e.:

$$\boldsymbol{R}_s = \boldsymbol{R}_{s1} + \boldsymbol{R}_i \tag{28}$$

where $\boldsymbol{R}_i$ corresponds to the part of $\boldsymbol{R}_s$ that is not correlated with s_1 (the first row and column are equal to zero). $\boldsymbol{R}_i$ is

2329-9290 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2761699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

treated as noise in $\boldsymbol{w}_{\text{SDW-IFWF-X}}$, which is seen by rewriting (24) as:

$$\boldsymbol{w}_{\text{SDW-IFWF-X}} = (\boldsymbol{R}_{s1} + \mu(\boldsymbol{R}_v + \boldsymbol{R}_i))^{-1}\boldsymbol{R}_{s1}\boldsymbol{e}_1 \quad (29)$$

The MSE problem that has (29) as a minimum can be found by writing out the MSE in (10), using (15) and (28) and then applying the trade-off parameter μ on $\boldsymbol{w}^H(\boldsymbol{R}_v + \boldsymbol{R}_i)\boldsymbol{w}$:

$$J_{\text{SDW-IFWF-X}} = \boldsymbol{w}^H\boldsymbol{R}_{s1}\boldsymbol{w} + \boldsymbol{e}_1^H\boldsymbol{R}_{s1}\boldsymbol{e}_1 - \boldsymbol{w}^H\boldsymbol{R}_{s1}\boldsymbol{e}_1 - \boldsymbol{e}_1^H\boldsymbol{R}_{s1}\boldsymbol{w} + \mu\boldsymbol{w}^H(\boldsymbol{R}_v + \boldsymbol{R}_i)\boldsymbol{w} \quad (30)$$

Conversely, rewriting (17) as:

$$\boldsymbol{w}_{\text{SDW-IFWF}} = (\boldsymbol{R}_{s1} + \mu(\boldsymbol{R}_v + \frac{1}{\mu}\boldsymbol{R}_i])^{-1}\boldsymbol{R}_{s1}\boldsymbol{e}_1 \quad (31)$$

it is seen that $\boldsymbol{w}_{\text{SDW-IFWF}}$ treats $\frac{1}{\mu}\boldsymbol{R}_i$ as noise. Consequently, $\boldsymbol{w}_{\text{SDW-IFWF}}$ and $\boldsymbol{w}_{\text{SDW-IFWF-X}}$ will have similar performance when $\mu \approx 1$, while $\boldsymbol{w}_{\text{SDW-IFWF-X}}$ will provide more noise reduction reduction when $\mu \to 0$, since then $\boldsymbol{w}_{\text{SDW-IFWF}}$ becomes $\boldsymbol{e}_1$ according to (18). It is seen that $\boldsymbol{R}_i$ is ignored in $\boldsymbol{w}_{\text{SDW-IFWF-1}}$ since it does not use the full speech matrix $\boldsymbol{R}_s$. For this reason we prefer $\boldsymbol{w}_{\text{SDW-IFWF-X}}$ over $\boldsymbol{w}_{\text{SDW-IFWF-1}}$, which will not be used in the rest of the paper.

IV. HRFB-based inter-frame Wiener filters

In this section, we consider the estimation of the IFC matrices used in the SDW-IFWFs. First we present the traditional way of estimating the IFC matrices. Then, in Section IV-A we show how the IFC coefficients can be estimated using a HRFB. In Section IV-B, we show how real-valued scalar gains can be applied directly in the HRFB instead of the SDW-IFWFs in the primary filter bank. In Section IV-C, a MAP method proposed in [9] for estimating the normalized IFC vector γ_s is described.

The IFC matrix $\boldsymbol{R}_x$ is usually estimated during speech+noise periods [2], [3], [9] using a first order recursion:

$$\hat{\boldsymbol{R}}_x[i] = \lambda\hat{\boldsymbol{R}}_x[i-1] + (1-\lambda)\boldsymbol{x}\boldsymbol{x}^H \quad (32)$$

where λ is an exponential forgetting factor that determines the effective time window, wherein the IFC matrix is estimated. Since $\boldsymbol{x}$ contains $X[k,i]$ for successive frames, a value of λ close to 1 means that $\hat{\boldsymbol{R}}_x$ becomes approximately a Hermitian Toeplitz matrix. However, the temporal inter-frame stationarity for a speech-like signal is usually around $10-50$ ms, which implies that λ must be set much smaller than 1.

Alternatively, assuming $X[k,i]$ is wide-sense stationary over a frame of L/R samples as defined in Section II, the IFC coefficients can be considered to only depend on the time lag:

$$r_k^x[m,i] = \mathbf{E}\{X[k,i]X^*[k,i-m]\} = \mathbf{E}\{X[k,i+t]X^*[k,i-m+t]\} = r_k^x[m] \quad (33)$$

where $t \in \mathbf{Z}$ and covers the range of stationarity. Therefore, the elements of the Hermitian Toeplitz matrix can be estimated as:

$$\hat{r}_k^x[m] = \frac{R}{L}\sum_{m'=0}^{L/R-m} X[k,i-m']X^*[k,i-m-m'] \quad (34)$$

and put into the Hermitian Toeplitz matrix $\hat{\boldsymbol{R}}_x$. This is similar to the auto-correlation method [18] in linear prediction that is known to result in stable filters.

The IFC matrix $\boldsymbol{R}_v$ can be estimated during noise-only periods in the same way as (32) [19] [20]. The IFC matrix $\hat{\boldsymbol{R}}_s$ can then be estimated as:

$$\hat{\boldsymbol{R}}_s = \hat{\boldsymbol{R}}_x - \hat{\boldsymbol{R}}_v \quad (35)$$

From (35), $\boldsymbol{\gamma}_s$ can be calculated as the first column of $\hat{\boldsymbol{R}}_s$ divided by the first element. Note that when using estimates of the IFC matrices, then (35) does not guarantee that $\hat{\boldsymbol{R}}_s$ is a positive semidefinite matrix.

A. HRFB-based estimation of IFC coefficients

The power spectral density (PSD) of subband signal $X[k,i]$ is given as the DTFT of the IFC:

$$P_k^x(f) = \sum_{m=-\infty}^{\infty} r_k^x[m]e^{-2\pi jfm}, \quad -\frac{1}{2} < f \le \frac{1}{2} \quad (36)$$

This will be the basic formula for the estimation of the IFC coefficients, rather than (33). Each subband signal $X[k,i]$ in the WOLA filter bank [1] is a filtered, modulated and R-fold downsampled version of $x[n]$, hence, assuming $h_K[n]$ is an ideal low-pass filter, we can write [11]:

$$P_k^x(f) \approx \frac{1}{R}\Big|H_K\Big(\frac{f}{R}\Big)\Big|^2 P^x\Big(\frac{f}{R} + f_k\Big), \quad -\frac{1}{2} < f \le \frac{1}{2} \quad (37)$$

where $H_K(f)$ is the DTFT of $h_K[n]$, $P^x(f)$ is the PSD of $x[n]$ and f_k is the normalized center frequency of subband k. From (5) and (6) it is seen that $r_k^x[m]$ is needed for $m = 0, ..., M-1$ to compute $\hat{\boldsymbol{R}}_x$. To obtain a discrete sampling of $P_k^x(f)$ that can be used to estimate these IFC coefficients, we aproximate the infinite sum in (36) by a finite sum $m = -M+1, ..., M$, i.e. all other IFC coefficients are assumed to be zero. These $2M$ IFC coefficients can then be computed by means of a $2M$ inverse DFT (IDFT):

$$r_k^x[m] = \frac{1}{2M}\sum_{m'=-M+1}^{M} P_k^x\Big(\frac{m'}{2M}\Big)e^{2\pi jmm'/(2M)}, \quad m = -M+1, ..., M \quad (38)$$

where $P_k^x\Big(\frac{m'}{2M}\Big)$ are $2M$ discrete samples of (37):

$$P_k^x\Big(\frac{m'}{2M}\Big) = \frac{1}{R}\Big|H_K\Big(\frac{m'}{2MR}\Big)\Big|^2 P^x\Big(\frac{m'}{2MR} + \frac{k}{K}\Big), \quad m' = -M+1, ..., M \quad (39)$$

Assuming an oversampling factor $O = K/R$ in $X[k,i]$, this can be written:

$$P_k^x\Big(\frac{m'}{2M}\Big) = \frac{O}{K}\Big|H_K\Big(\frac{Om'}{2MK}\Big)\Big|^2 P^x\Big(\frac{Om'}{2MK} + \frac{k}{K}\Big), \quad m' = -M+1, ..., M \quad (40)$$

This sampling of $P^x(f)$ effectively corresponds to a frequency resolution that is $\frac{2M}{O}$ times higher than the frequency resolution K of the WOLA filter bank defining $X[k,i]$. Hence, to calculate $P^x\Big(\frac{Om'}{2MK} + \frac{k}{K}\Big)$, a HRFB is needed with a $\frac{2M}{O}$ times

2329-9290 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2761699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE I
ALGORITHM FOR CALCULATING $\hat{\boldsymbol{R}}_x$.

Algorithm steps
For $l = -\frac{L}{2}+1, ..., \frac{L}{2}$ (with $L = \frac{2M}{O}K$) - Calculate $\hat{P}^x[l,i]$ using (41)
For $k = -\frac{K}{2}+1, ..., \frac{K}{2}$ - Calculate $P_k^x(\frac{m'}{2M})$ for $m' = -M+1, ..., M$ using (40) and $\hat{P}^x[l,i]$ - Calculate $r_k^x[m]$ for $m = -M+1, ..., M$ using (38) - Construct Hermitian Toeplitz matrix $\hat{\boldsymbol{R}}_x$ with first row equal to $r_k^x[m]$ for $m = 0, \ldots, M-1$

higher frequency resolution. This will be the HRFB defining $X[l,i]$ with $L = \frac{2M}{O}K$:

$$\hat{P}^x[l,i] = \frac{1}{L}|X[l,i]|^2 \approx P^x[l,i] = P^x\Big(\frac{l}{L}\Big), \quad l = -\frac{L}{2}+1, ..., \frac{L}{2} \tag{41}$$

With this, we have established a relation between the IFC of $X[k,i]$ and the HRFB $X[l,i]$, which can be used to calculate the IFC coefficients $r_k^x[m]$ by inserting the approximation (41) in (40) and then using (38). It is noted that when $M \leq \frac{O}{2}$, then no HRFB is needed.

From (33) it follows that $\boldsymbol{R}_x$ is a Hermitian Toeplitz matrix with elements $r_k^x[m]$. For each new time frame i, the IFC coefficients for $X[k,i]$ are estimated using the PSD obtained from $X[l,i]$ and then put into $\hat{\boldsymbol{R}}_x$. From its definition in (12), it is seen that $\boldsymbol{R}_x$ contains the IFC for lag $m = 0, ..., M-1$. Therefore only half of the IFC coefficients from the IDFT in (38) is needed. The overall algorithm for calculating $\hat{\boldsymbol{R}}_x$ is summarized in Table I.

The noise IFC matrix $\boldsymbol{R}_v$ can be estimated in a similar way. A noise estimation method, for instance based on minimum statistics [21] [22] or a VAD [19] [20], can be applied to $\hat{P}^x[l,i]$ to find a noise estimate $\hat{P}^v[l,i]$. The part for $k = -\frac{K}{2}+1, ..., \frac{K}{2}$ in Table I can then be repeated, replacing $\hat{P}^x[l,i]$ with $\hat{P}^v[l,i]$ to find $\hat{\boldsymbol{R}}_v$.

The speech IFC matrix $\boldsymbol{R}_s$ can be estimated as follows. The *a-priori* signal-to-noise ratio (SNR) can be estimated using the well-known *decision-directed* approach [23]:

$$\xi[l,i] = a\frac{|\hat{S}[l,i-1]|^2}{\hat{P}^v[l,i-1]} + (1-a)\max\Big(\frac{\hat{P}^x[l,i]}{\hat{P}^v[l,i]} - 1, 0\Big) \tag{42}$$

where $\xi[l,i]$ is the *a-priori* SNR, a is the smoothing parameter and $\hat{S}[l,i-1]$ is the speech estimate from the previous frame. The *a-priori* SNR $\xi[l,i]$ is then used in a power subtraction [24], [25]:

$$\hat{P}^s[l,i] = \frac{\xi[l,i]}{1+\xi[l,i]}\hat{P}^x[l,i] \tag{43}$$

The part for $k = -\frac{K}{2}+1, ..., \frac{K}{2}$ in Table I can then be repeated, replacing $\hat{P}^x[l,i]$ with $\hat{P}^s[l,i]$ to find $\hat{\boldsymbol{R}}_s$.

Comparing the proposed method to the first-order recursion (32), the IFC matrices are explicitly modelled as Hermitian Toeplitz matrices, which means they can be efficiently inverted in the filter equations (17) and (25), for instance using the Levinson recursion [26]. The robustness of these filters will be demonstrated in Section V.

B. HRFB-based realisation of SDW-IFWFs

In this section, we use SDW-IFWFs that are symmetric around $m = 0$ by redefining (5):

$$\boldsymbol{x}[k,i] = [X[k,i+M], ..., X[k,i], ..., X[k,i-M+1]]^T \tag{44}$$

As will be shown, this symmetry means that the SDW-IFWFs can be converted into real-valued scalar gains $g[l,i]$ and hence realized in the HRFB $X[l,i]$. Using the vector (44) in the primary filter bank would make the resulting filter non-causal. However, since the filter can be converted to real-valued scalar gains in the HRFB, the only delay in the resulting system is the one from the HRFB. Thus, the design of the HRFB determines the delay of the system which will be explained in Section V-A. It will be shown that the resulting $g[l,i]$ have a dependency on the standard Wiener gains in neighboring subbands of $X[l,i]$, which makes it possible, by means of the MVDR filters obtained with $\mu = 0$, to apply real-valued noise reduction gains in $X[l,i]$ while preserving the speech PSD in the frequency range covered by a subband of $X[k,i]$.

To realise the SDW-IFWFs in the HRFB $X[l,i]$ with a frequency resolution equal to L, a $2M$ point DFT is used over the range $m = -M+1, ..., M$, which can be written in matrix notation as:

$$\boldsymbol{W}[k,i] = \boldsymbol{D}\boldsymbol{w}[k,i] \tag{45}$$

where $\boldsymbol{D}$ is a permuted $2M \times 2M$ DFT matrix:

$$\boldsymbol{D} = \begin{bmatrix} e^{-2\pi j(-M+1)(-M+1)/(2M)} & \cdots & e^{-2\pi jM(-M+1)/(2M)} \\ \vdots & \ddots & \vdots \\ e^{-2\pi j(-M+1)M/(2M)} & \cdots & e^{-2\pi jMM/(2M)} \end{bmatrix} \tag{46}$$

and

$$\boldsymbol{W}[k,i] = [W_k[-M+1,i], ..., W_k[M,i]]^T \tag{47}$$

With $\boldsymbol{x}[k,i]$ redefined as in (44), the IFC matrix for the noisy speech can be approximated, using matrix notation, as the $2M \times 2M$ circulant matrix $\boldsymbol{C}_x$ [27]:

$$\boldsymbol{C}_x = \frac{1}{2M}\boldsymbol{D}^H\boldsymbol{P}_x\boldsymbol{D} \tag{48}$$

where $\boldsymbol{P}_x$ is a diagonal matrix containing $P_k^x\Big(\frac{m'}{2M}\Big), m' = -M+1, ..., M$ from (40) in the diagonal and similarly for $\boldsymbol{R}_s$ and $\boldsymbol{R}_v$, which are calculated from diagonal matrices $\boldsymbol{P}_s$ and $\boldsymbol{P}_v$. Then, $\boldsymbol{C}_x$ can replace $\boldsymbol{R}_x$ in the SDW-IFWFs in Section III and similar approximations can be made for $\boldsymbol{C}_s$ and $\boldsymbol{C}_v$.

The IFWF MSE (10) can be rewritten by using the circulant matrices $\boldsymbol{C}_x$ and $\boldsymbol{C}_s$ and the corresponding versions of (48):

$$J_{\text{IFWF}} = \mathbf{E}\{|\boldsymbol{w}^H\boldsymbol{x} - \boldsymbol{e}_{M+1}^H\boldsymbol{s}|^2\} = \boldsymbol{w}^H\boldsymbol{C}_x\boldsymbol{w} + \boldsymbol{e}_{M+1}^H\boldsymbol{C}_s\boldsymbol{e}_{M+1} - \boldsymbol{w}^H\boldsymbol{C}_s\boldsymbol{e}_{M+1} - \boldsymbol{e}_{M+1}^H\boldsymbol{C}_s\boldsymbol{w} \propto \boldsymbol{W}^H\boldsymbol{P}_x\boldsymbol{W} + \mathbf{1}^H\boldsymbol{P}_s\mathbf{1} - \boldsymbol{W}^H\boldsymbol{P}_s\mathbf{1} - \mathbf{1}^H\boldsymbol{P}_s\boldsymbol{W} \tag{49}$$

2329-9290 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2761699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

where $\mathbf{1}$ is a vector of ones of suitable size. The solution to (49) is given by:

$$\boldsymbol{W}_{\text{IFWF}} = \boldsymbol{P}_x^{-1}\boldsymbol{P}_s\mathbf{1} \tag{50}$$

which is the standard Wiener gain. Since $\boldsymbol{P}_x$ and $\boldsymbol{P}_s$ are diagonal matrices, the standard Wiener gain does not depend on neighbouring subbands.

The MSE for the rank-1 SDW-IFWF-X (30) can be rewritten in a similar way:

$$\begin{aligned} J_{\text{SDW-IFWF-X}} \propto \mu\boldsymbol{W}^H\boldsymbol{P}_x\boldsymbol{W} + (1-\mu)\boldsymbol{W}^H\frac{\boldsymbol{P}_s\mathbf{1}\mathbf{1}^H\boldsymbol{P}_s}{\mathbf{1}^H\boldsymbol{P}_s\mathbf{1}}\boldsymbol{W} \\ + \mathbf{1}^H\boldsymbol{P}_s\mathbf{1} - \boldsymbol{W}^H\boldsymbol{P}_s\mathbf{1} - \mathbf{1}^H\boldsymbol{P}_s\boldsymbol{W} \end{aligned} \tag{51}$$

For $\mu = 0$, this can be written as:

$$J_{\text{SDW-IFWF-X}(\mu=0)} \propto \frac{|\boldsymbol{W}^H\boldsymbol{P}_s\mathbf{1} - \mathbf{1}^H\boldsymbol{P}_s\mathbf{1}|^2}{\mathbf{1}^H\boldsymbol{P}_s\mathbf{1}} \tag{52}$$

Using the matrix inversion lemma, the solution to (51) is obtained as:

$$\boldsymbol{W}_{\text{SDW-IFWF-X}} = \frac{\boldsymbol{P}_x^{-1}\boldsymbol{P}_s\mathbf{1}\mathbf{1}^H\boldsymbol{P}_s\mathbf{1}}{\mu\mathbf{1}^H\boldsymbol{P}_s\mathbf{1} + (1-\mu)\mathbf{1}^H\boldsymbol{P}_s\boldsymbol{P}_x^{-1}\boldsymbol{P}_s\mathbf{1}} \tag{53}$$

This can be written in terms of the standard Wiener gain (50):

$$\boldsymbol{W}_{\text{SDW-IFWF-X}} = \boldsymbol{W}_{\text{IFWF}}\frac{1}{\mu + (1-\mu)A} \tag{54}$$

where

$$A = \frac{\mathbf{1}^H\boldsymbol{P}_s\boldsymbol{W}_{\text{IFWF}}}{\mathbf{1}^H\boldsymbol{P}_s\mathbf{1}} \tag{55}$$

is a weighted average of $\boldsymbol{W}_{\text{IFWF}}$. The average is weighted by the speech PSD and leads to an intuitive interpretation of the single-channel MVDR filter:

$$\boldsymbol{W}_{\text{SDW-IFWF-X}(\mu=0)} = \boldsymbol{W}_{\text{IFWF}}\frac{1}{A} \tag{56}$$

It is seen that the MVDR gain is the standard Wiener gain scaled by the inverse of the weighted average of the standard Wiener gains, weighted by how much speech power is in each respective subband. Thus, the single-channel MVDR gain effectively scales the entire subband signal $x[k,i]$ according to how much gain reduction is applied to the individual subband speech signals $\boldsymbol{s}[k,i]$ by the Wiener gains.

Finally, the vectors $\boldsymbol{W}[k,i]$ can be put into the gain vector $g[l,i]$ that is applied to the noisy speech in the WOLA filter bank defining $X[l,i]$, to replace (6) by:

$$\hat{S}[l,i] = g[l,i]X[l,i] \tag{57}$$

This is done using an "overlap" procedure, where $g[l,i]$ is a weighted combination of the vectors $\boldsymbol{W}[k,i]$ that use the same subbands of the WOLA filter bank defining $X[l,i]$. There are O overlapping $\boldsymbol{W}[k,i]$'s with a spacing of $\frac{L}{K}$ subbands and so the "overlap" can be written as:

$$g[l,i] = \sum_{o=0}^{O-1} H_K[l' + o\frac{L}{K}]W_{\frac{K}{L}(l-l')-o}[(l' + o\frac{L}{K})R, i] \tag{58}$$

$$l' = \mod(l + M - 1, \frac{L}{K}) - M + 1 \tag{59}$$

The weighting $H_K[m']$ corresponds to the analysis prototype filter of the filter bank defining $X[k,i]$. A more formal proof of this "overlap" procedure is given in the Appendix. The estimated speech $\hat{S}[l,i]$ is then synthesized as usual now in the WOLA filter bank defining $X[l,i]$.

Since the SDW-IFWF-X in this section is realized by means of scalar gains, it is inherently robust. It is also significantly cheaper than the SDW-IFWFs in previous sections, since it does not need any matrix inversions. The performance of the SDW-IFWF-X will be demonstrated in Section V.

C. MAP estimation of the IFC coefficient vector

It has previously been shown [9] that the MVDR filter can be robustly computed using a statistical estimation of $\boldsymbol{\gamma}_s$ in a maximum likelihood (ML) approach and further improved in a MAP approach. In the following we briefly describe the method and show that a simpler MAP estimation can be done in the HRFB $X[l,i]$.

Assuming normally-distributed, multivariate random variables and that the noise is white, $\boldsymbol{\gamma}_s$ can be estimated as:

$$\hat{\boldsymbol{\gamma}}_s = \frac{\sigma_x}{\sigma_s}\hat{\boldsymbol{\gamma}}_x - \frac{\sigma_v}{\sigma_s}\boldsymbol{\mu}_v \tag{60}$$

where $\boldsymbol{\mu}_v$ is the long-term IFC coefficient vector of the noise, of which the ith element can be computed from the prototype filter $h_K[n]$:

$$\boldsymbol{\mu}_v[m] = \frac{\sum_n h_K[n]h_K[n+mR]}{\sum_n h_K^2[n]} \tag{61}$$

Using multivariate normal distributions for the likelihood and prior probabilities, the MAP estimate is then given as:

$$\begin{aligned} \hat{\boldsymbol{\gamma}}_{s,MAP} = \boldsymbol{\Phi}_{\boldsymbol{\gamma}_v}\left[\frac{\sigma_v^2}{\sigma_s^2}\boldsymbol{\Phi}_{\boldsymbol{\gamma}_v} + \boldsymbol{\Phi}_{\boldsymbol{\gamma}_s}\right]^{-1}\boldsymbol{\Phi}_{\boldsymbol{\gamma}_s}\boldsymbol{\Phi}_{\boldsymbol{\gamma}_v}^{-1}\hat{\boldsymbol{\gamma}}_s \\ + \frac{\sigma_v^2}{\sigma_s^2}\boldsymbol{\Phi}_{\boldsymbol{\gamma}_v}\left[\frac{\sigma_v^2}{\sigma_s^2}\boldsymbol{\Phi}_{\boldsymbol{\gamma}_v} + \boldsymbol{\Phi}_{\boldsymbol{\gamma}_s}\right]^{-1}\boldsymbol{\mu}_{\boldsymbol{\gamma}_s} \end{aligned} \tag{62}$$

where $\boldsymbol{\mu}_{\boldsymbol{\gamma}_s} = \mathbf{E}\{\boldsymbol{\gamma}_s\}$, $\boldsymbol{\Phi}_{\boldsymbol{\gamma}_s} = \mathbf{E}\{(\boldsymbol{\gamma}_s - \boldsymbol{\mu}_{\boldsymbol{\gamma}_s})(\boldsymbol{\gamma}_s - \boldsymbol{\mu}_{\boldsymbol{\gamma}_s})^H\}$ and similarly for the noise. The speech IFC mean vector and the noise and speech IFC covariance matrices should be calculated offline on a training data set. For low SNR $\frac{\sigma_v}{\sigma_s} \to \infty$ it is seen that $\hat{\boldsymbol{\gamma}}_{s,MAP} \to \boldsymbol{\mu}_{\boldsymbol{\gamma}_s}$, while for high SNR $\frac{\sigma_v}{\sigma_s} \to 0$ it is seen that $\hat{\boldsymbol{\gamma}}_{s,MAP} \to \hat{\boldsymbol{\gamma}}_s$.

By following the method in Section IV-B we can define a speech IFC coefficient vector $\boldsymbol{\Gamma}_s$ in the HRFB. A MAP estimate of $\boldsymbol{\Gamma}_s$ can then be found using an expression similar to (62), where $\boldsymbol{\gamma}_s$ is replaced with $\boldsymbol{\Gamma}_s$ and similarly for the noise and noisy speech. We can define the speech IFC coefficient vector in the HRFB as:

$$\boldsymbol{\Gamma}_s = \frac{\boldsymbol{P}_s\mathbf{1}}{\mathbf{1}^H\boldsymbol{P}_s\mathbf{1}} \tag{63}$$

However, a much simpler MAP estimation can be found by considering that the covariance in the IFC covariance matrix is much smaller in the HRFB. Using 30 minutes of speech data, the statistical parameters in the HRFB have been calculated and the covariance matrix $\boldsymbol{\Phi}_{\boldsymbol{\Gamma}_s}$ of Γ_s for k corresponding to the center frequencies $0.5, 1.5, 3$ and 6 kHz can be seen in Figure 1. It is seen that to a reasonable approximation we

2329-9290 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2761699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

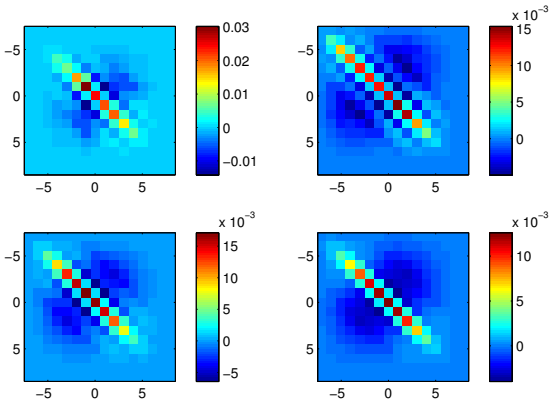


Fig. 1. Covariance matrix of $\mathbf{\Gamma}_s$ for k corresponding to a center frequency of (Top left) 500 Hz, (Top right) 1.5 kHz, (Bottom left) 3 kHz, (Bottom right) 6 kHz.

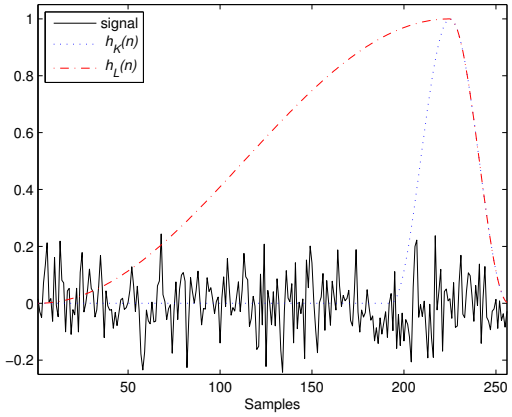


Fig. 2. Analysis prototype windows used in experimental evaluation.

can model the elements in $\mathbf{\Gamma}_s$ as independent, i.e. $\mathbf{\Phi}_{\mathbf{\Gamma}_s}$ is a diagonal matrix, which greatly simplifies the MAP estimation. Similar observations are made for other values of k and writing the MAP estimation as a 1-dimensional variable we get:

$$\hat{\Gamma}^s_{k,MAP}[m'] = \frac{\phi_{\Gamma_s,k}[m']}{\frac{\sigma^2_{v,k}[m']}{\sigma^2_{s,k}[m']}\phi_{\Gamma_v,k}[m'] + \phi_{\Gamma_s,k}[m']}\hat{\Gamma}^s_k[m'] + \frac{\frac{\sigma^2_{v,k}[m']}{\sigma^2_{s,k}[m']}\phi_{\Gamma_v,k}[m']}{\frac{\sigma^2_v[l,i]}{\sigma^2_s[l,i]}\phi_{\Gamma_v,k}[m'] + \phi_{\Gamma_s,k}[m']}\mu_{\Gamma_s,k}[m'] \quad (64)$$

where $\phi_{\Gamma_s,k}[m']$ is the diagonal elements in the matrix shown in Figure 1 and similar definitions exist for $\phi_{\Gamma_v,k}[m']$ and $\mu_{\Gamma_x,k}[m']$. In practice we then need to calculate $\hat{\Gamma}^s_k[m']$ from (63) using a speech power estimate $\hat{P}^s[l,i]$ and $\frac{\sigma^2_{v,k}[m']}{\sigma^2_{s,k}[m']}$ which is the inverse squared of the SNR. Both of these can be estimated using the *decision-directed* SNR estimator (42) and (43). All other parameters are calculated offline from speech and noise data.

The MAP speech power is then calculated using the inverse of (63):

$$\hat{P}^s_{k,MAP}[m'] = \hat{\Gamma}^s_{k,MAP}[m'] \sum_p \hat{P}^s_k[m'] \quad (65)$$

The power estimate $\hat{P}_{s,MAP}[l,i]$ is then used in the SDW-IFWF gain in Section IV-B.

V. Experimental results

In this section an experimental evaluation of the proposed methods is performed and comparisons are made with existing reference methods. First the setup and test data are described along with implementation details and reference methods used in the evaluation. Then the SDW-IFWF methods of Section III are compared to the HRFB gain based methods and the MAP estimation methods.

A. Setup and test data

A WOLA filter bank defining $X[k,i]$ was implemented with $K = 64$, $O = 4$ and $R = 64/4 = 16$. A periodic Hann window was used as the analysis prototype filter:

$$h_K[n] = \frac{1}{2}\left(1 - \cos\left(\frac{2\pi n}{K}\right)\right), \quad 0 \le n < K \quad (66)$$

Different choices of the HRFB defining $X[l,i]$ were compared based on performance and computational complexity and a value of $L = 256$ was selected for further testing. This corresponds to applying an $M = OL/(2K) = 8$ SDW-IFWF in each subband. Larger values of M were found to lead to a very computationally demanding procedure due to the required matrix inversions. The gain based methods in Section IV-B do not have this problem, but to be able to compare the different methods they also used a value of $M = 8$. The analysis prototype filter $h_L[n]$ of the HRFB selects the frame of $x[n]$ that is used for estimating the SDW-IFWFs. This frame can not exceed the last sample of $h_K[n]$ to maintain the delay of $X[k,i]$. An asymmetric $h_L[n]$ was chosen that keeps most of the energy concentrated at the most recent part of the frame and was created by concatenating two "half Hann" windows of different lengths to create a window of total length L. The prototype windows $h_K[n]$ and $h_L[n]$ are shown in Figure 2. The used synthesis prototype filter was $h_K[n]$. Thus, the resulting HRFB has an analysis window of length $L = 256$ to obtain a high frequency resolution, but a short synthesis window of length $K = 64$ to keep the total delay of the HRFB equal to the primary filter bank [12], [13], [14]. The derived scalar gain values in the HRFB are real-valued and therefore does not impose any delay. All compared methods have the same delay and since the gain-based methods are much cheaper, they are considered superior in terms of real-time performance.

To avoid boosting any noise components and make the gain more robust against estimation errors, it was found advantageous to limit the gain (54) in the HRFB to a maximum value $\delta_{\max}$:

$$\boldsymbol{W}_{\text{SDW-IFWF-X}} = \min\left(\boldsymbol{W}_{\text{IFWF}}\frac{1}{\mu + (1-\mu)A}, \delta_{\max}\right) \quad (67)$$

2329-9290 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2761699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

where $\delta_{\max} = 1$ limits the gain to only attenuate the signal. In theory this limit can increase the speech distortion when the true IFC matrices are used. In practice, however, it was found that the limit only had a minor negative influence on the speech distortion when IFC estimates are used and a positive impact on the SNR. Since this paper investigates robust SDW-IFWFs we consider this simple limit a further advantage of the proposed gain-based method when dealing with estimated IFC matrices.

All methods considered in the experimental comparison use estimates of the speech and noise IFC matrices and the results can be considered as real-life performance when implemented in a speech enhancement system. This is important since we are proposing methods that are robust when dealing with estimates of the true values and shows the improvement of using the MAP estimates. The method using the first order recursion (32) together with the SDW-IFWF (25) is denoted as **FORWX**. The forgetting factor λ was set to 0.85, which seemed to give the best performance. The IFC matrices were also estimated as described in Section IV and the noise was estimated using the method described in [20]. The speech IFC vector $\boldsymbol{\gamma}_s$ was estimated using the ML method (60). The method similar to **FORWX**, but using the auto-correlation estimate (34) instead is denoted **ACWX**. The method using the HRFB to derive a filter in $X[k,i]$ described in Section IV-A together with the SDW-IFWFs (17) and (25) are denoted as **HRW** and **HRWX** respectively. The HRFB method in Section IV-B with the gain (67) is denoted as **HRGX**. The SDW Wiener gain:

$$\boldsymbol{w}_{\text{SDW-W}}[k,i] = \frac{\xi[k,i]}{\mu + \xi[k,i]} \tag{68}$$

where $\xi[k,i]$ is the *a-priori* SNR smoother (42) was also implemented and is denoted as **SDWG**. The smoothing parameter a was set to 0.97. The old speech estimate in the *a-priori* SNR smoother was updated every 64 samples as a more often update was found to produce excessive musical noise. This was also done in the methods using the HRFB $X[l,i]$. It is noted that according to (50) the method **HRGX** reduces to the standard Wiener gain in the HRFB when $\mu = 1$. This means that for $\mu = 1$, the improvement in performance of **HRGX** compared to **SDWG** can be seen as the difference in performance of the Wiener gain when using a HRFB instead of the primary filter bank. The MAP method for estimating $\boldsymbol{\gamma}_s$ in Section IV-C was also tested and the corresponding method is indicated by a **-MAP** suffix in the method name.

The test data consisted of 8 female and 8 male speech signals taken from the TIMIT database and sampled at 16kHz. Each speech signal consisted of 5 sentences and had a length of $12 - 13$ seconds. The used noise signals consisted of a car engine signal that had a significant low frequency characteristic, an outdoor traffic signal that had a high dynamic range and a drilling noise that had a wide-band frequency characteristic with some peaks in the $200-500$Hz range. Each speech signal was mixed at some SNR with each noise signal and the evolution was then averaged over all the data. Each averaged result corresponded to approximately 10 minutes of data.

The time-domain output signal $\hat{s}[n]$ was synthesized in the WOLA filter bank from either $\hat{S}[k,i]$ or $\hat{S}[l,i]$ and can be decomposed as:

$$\hat{s}[n] = \hat{s}_s[n] + \hat{s}_v[n] \tag{69}$$

where $\hat{s}_s[n]$ is the filtered speech component and $\hat{s}_v[n]$ is the filtered noise component.

The methods were evaluated using the segmental SNR [28]:

$$\text{seg. SNR} = \frac{1}{|\mathbf{M}|}\sum_{m\in\mathbf{M}} 10\log_{10}\frac{\sum_{n=0}^{N-1} s^2[n+mR]}{\sum_{n=0}^{N-1}(s[n+mR]-\hat{s}[n+mR])^2} \tag{70}$$

where N is the number of samples in a 30ms frame and $\mathbf{M}$ is the set of frames with significant speech energy. We also used the objective PESQ measure, which has a good correlation with the perceptual speech quality [29]. We also used the segmental speech-distortion index [30]:

$$\text{seg. SD} = \frac{1}{|\mathbf{M}|}\sum_{m\in\mathbf{M}} 10\log_{10}\frac{\sum_{n=0}^{N-1}(s[n+mR]-\hat{s}_s[n+mR])^2}{\sum_{n=0}^{N-1} s^2[n+mR]} \tag{71}$$

Unlike the seg. SNR, the seg. SD only measures the distortion of the speech component.

B. Comparison of SDW-IFWFs and reference methods

In this section, the methods **HRGX**, **FORWX**, **HRW**, **HRWX**, **SDWG**, **ACWX** described in the previous section are compared. As mentioned in Section III, the rank-1 assumption is essential for allowing noise reduction for $\mu = 0$. This is shown in Section IV-B to correspond to a constraint on the Wiener gain in a HRFB. It will be evaluated how this constraint affects the trade-off between noise reduction and speech distortion by varying the parameter μ in the range from 0 to 2. The results for an input SNR of 0dB and 10dB are given in Figure 3 and Figure 4 respectively. It is seen that the seg. SNR is a monotonic function of μ for all the investigated filters except **FORWX** and **ACWX**. This is expected as a higher value of μ puts more emphasis on the noise in the cost function (16). The improvement saturates for large values of μ. For small values of μ, **HRW** and **SDWG** perform significantly worse than the other methods. This is because they do not apply any gain reduction for $\mu = 0$ as seen in the analysis in Section III. It is also seen that **HRW** and **HRWX** have similar, but not identical, performance for $\mu = 1$ as would otherwise be expected from the analysis of the filters. This is because **HRW** uses $\boldsymbol{R}_x$ directly while **HRWX** uses $\boldsymbol{R}_s + \boldsymbol{R}_v$ where $\boldsymbol{R}_s$ has been estimated using the *a-priori* SNR smoother, so in general $\boldsymbol{R}_x \neq \boldsymbol{R}_s + \boldsymbol{R}_v$. The difference between **HRW** and **HRWX** is largest for small values of μ, which is due to the different weighting of $\boldsymbol{R}_i$ as explained in the analysis of the filters. It is also seen that the use of the Hermitian Toeplitz assumption in the calculation of the IFC matrices in **ACWX** improves the performance compared to **FORWX**. However, **ACWX** still performs worse than the other investigated methods in

2329-9290 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2761699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

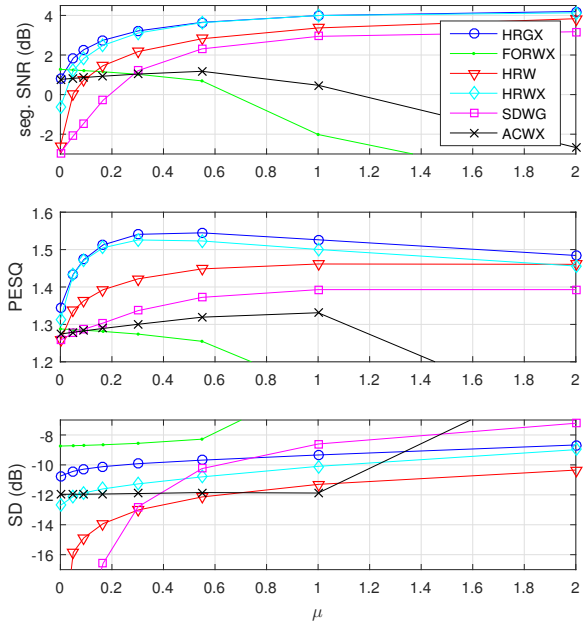


Fig. 3. Comparison of methods for an input SNR of 0dB.

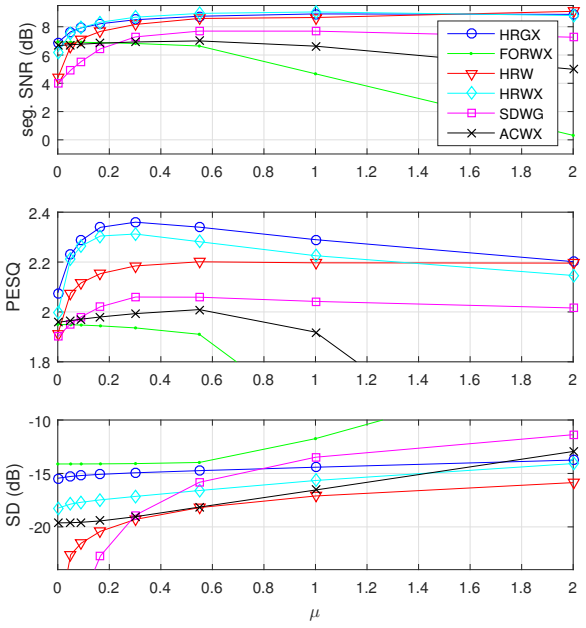


Fig. 4. Comparison of methods for an input SNR of 10dB.

terms of SNR and PESQ. Looking at the PESQ values it is seen that the rank-1 filters **HRGX** and **HRWX** achieve higher values than the full-rank **HRW** and the Wiener gain **SDWG**, especially for smaller values of μ around $0.1 - 0.5$. Small values of μ leads to less speech distortion, which is not necessarily the same as the speech quality measured by PESQ. Indeed, it is seen that PESQ drops as $\mu \to 0$. It is interesting that the best perceptual quality for the rank-1 methods seems to be obtained for a speech distortion weighting somewhere between the MVDR value ($\mu = 0$) and standard Wiener value ($\mu = 1$). Informal listening tests have confirmed that the subjective speech quality is in accordance with the PESQ measure. Looking at the seg. SD values it is seen that the seg. SD of all methods increases monotonically with μ. The method **FORWX** has the largest seg. SD while the seg. SD of **HRW** and **SDWG** become zero for $\mu = 0$. This is due to the fact that **HRW** and **SDWG** do not perform any noise reduction for $\mu = 0$.

It is seen that **FORWX** performs best for $\mu = 0$. For increasing values of μ both the seg. SNR and PESQ decrease monotonically. Investigating the output signal it was found that **FORWX** can produce unexpectedly large outputs for $\mu > 0$. This was not found in [2], [3] as the true noise and speech signals were used to calculate the ICF matrices. In [9] only experimental results are shown for the MVDR filter (i.e. $\mu = 0$). For $\mu = 0$, the seg. SNR is similar to **HRGX** while the PESQ values are worse. The main difference, however, seems to be that **HRGX** is robust for $\mu > 0$, where **FORWX** produces unexpected results. Looking at the IFWFs, it is seen that the inverse IFC matrix $\boldsymbol{R}_s^{-1}$ appears in both the numerator and denominator of the MVDR filter ($\mu = 0$). This power normalization makes it more robust against estimation errors.

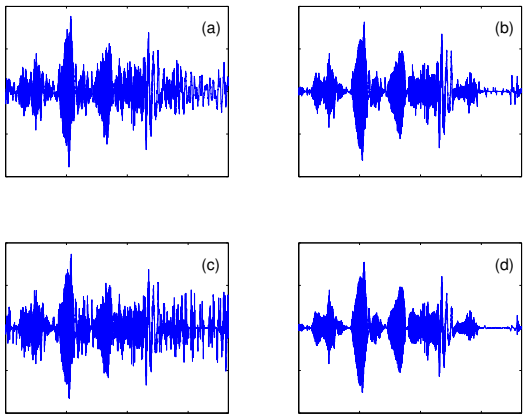


Fig. 5. Example of output signal. **FORWX** produces unexpected results for $\mu > 0$. (a) Noisy input signal. (b) Output signal for **FORWX** ($\mu = 0$). (c) Output signal for **FORWX** ($\mu = 1$). (d) Output signal for **HRGX** ($\mu = 1$).

An example of the output signal is given in Figure 5. Finally it is seen that **SDWG** always performs worse than methods using the HRFB.

C. Comparison of MAP estimation

Of the proposed methods in the previous section, **HRGX** has the best performance and in this section the MAP estimation, for the normalized IFC speech vector γ_s as described in Section IV-C, is evaluated against the MAP estimation using **FORWX** and $\mu = 0$, which is the method proposed in [9]. The results are shown in Figure 6, where the input SNR is varied from -5 to 15dB. It is seen that the MAP method

2329-9290 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2761699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

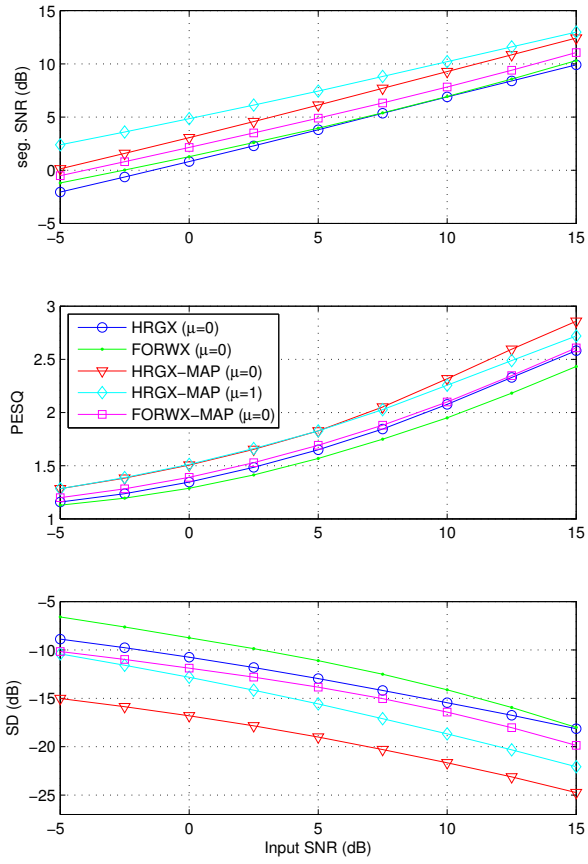


Fig. 6. Comparison of MAP estimation for varying input SNR.

improves both the seg. SNR and PESQ for both **HRGX** and **FORWX**. For $\mu = 1$, the seg. SNR is improved compared to $\mu = 0$ (**FORWX** is not shown for $\mu = 1$ since it is not robust as explained in the previous section). Looking at the SD, the opposite result is seen. The MVDR filter ($\mu = 0$) has a lower seg. SD compared to the corresponding standard Wiener filter ($\mu = 1$). This illustrates that while the standard Wiener filter provide a better reduction of speech plus noise distortion, the MVDR filter is superior if only speech distortion is considered. It is also seen that **HRGX-MAP** that was proposed in this paper has a superior performance compared to **FORWX-MAP**.

VI. Conclusion

SDW-IFWFs have been proposed in this paper for single channel noise reduction. These filters can realize a trade-off between speech distortion and noise reduction by means of a parameter μ. In particular, rank-1 SDW-IFWFs have been derived that for $\mu = 0$ reduce to an MVDR filter. It has been shown that the SDW-IFWFs can be realized by means of real-valued gains in a HRFB and further improved using a MAP estimation scheme. Noise reduction experiments have been performed to show the performance of these methods in realistic scenarios and that they result in robust noise reduction for a wide range of μ values. The proposed methods have been compared to methods based on the traditional first order recursion estimation of the IFC coefficients, which is experimentally shown to only be robust for $\mu = 0$. It has been shown that larger μ values provide better performance in terms of seg. SNR, while the MVDR filters perform better in terms of speech distortion. The MAP estimation method has been shown to improve the noise reduction in the HRFB and outperform a previously proposed MAP method. A further advantage of the proposed methods is that they can avoid the need for matrix inversions and are therefore computationally cheaper than previous methods.

Appendix

This Appendix contains a more formal proof for the "overlap" procedure used in Section IV-B. The synthesis of a frame i of the estimated speech (6) in the WOLA filter bank defining $X[k,i]$ is given as:

$$\hat{s}_i[n] = f[n - iR] \sum_{k=0}^{K-1} \boldsymbol{w}^H[k,i]\boldsymbol{x}[k,i]e^{2\pi jkn/K} \tag{72}$$

where $f[n]$ is the synthesis prototype filter. Using the HRFB $X[l,i]$ from Section IV-A, $\boldsymbol{w}^H[k,i]\boldsymbol{x}[k,i]$ is given as:

$$\boldsymbol{w}[k,i]^H\boldsymbol{x}[k,i] = \frac{1}{2M}\sum_{m'=-M+1}^{M} W_k\Big(\frac{m'}{2M}\Big)^* X_k\Big(\frac{m'}{2M}\Big)e^{2\pi jm'n/(2MR)} \tag{73}$$

with (similar to (39)):

$$X_k\Big(\frac{m'}{2M}\Big) = \frac{1}{R}H_K\Big(\frac{m'}{2MR}\Big)X\Big(\frac{m'}{2MR} + \frac{k}{K}\Big), \quad m' = -M+1, ..., M \tag{74}$$

Inserting (73) and (74) in (72), the synthesis can be realised directly in the HRFB $X[l,i]$:

$$\hat{s}_i[n] = f[n - iR]\frac{1}{L}\sum_{l=0}^{L-1} g\Big(\frac{l}{L}\Big)X\Big(\frac{l}{L}\Big)e^{2\pi ln/L} \tag{75}$$

where $l = m' + \frac{2Mk}{O}$ and $g\big(\frac{l}{L}\big)$ contains a sum of gains corresponding to the frequency sampling $\frac{l}{L}$:

$$g\Big(\frac{l}{L}\Big) = \sum_{o=0}^{O-1} H_K\Big(\frac{l' + \frac{2Mo}{O}}{L}\Big)W_{\frac{O}{2M}(l-l')-o}\Big(\frac{(l' + \frac{2Mo}{O})R}{L}\Big)^* \tag{76}$$

$$l' = \mod(l + M - 1, \frac{2M}{O}) - M + 1 \tag{77}$$

Since $\boldsymbol{W}$ is assumed real, the complex conjugation can be ignored and this corresponds to the "overlap" procedure used in Section IV-B.

2329-9290 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2761699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

REFERENCES

[1] R. Crochiere and L. Rabiner, *Multirate Digital Signal Processing*. Prentice-Hall, 1983.

[2] J. Benesty and Y. Huang, "A single-channel noise reduction mvdr filter," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 273–276.

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

[4] J. Benesty, J. Chen, Y. (Arden) Huang, and T. Gaensler, "Time-domain noise reduction based on an orthogonal decomposition for desired signal extraction," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 452–464, 2012. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/132/1/10.1121/1.4726071

[5] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, Jul 1995.

[6] 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.

[7] 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.

[8] 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 Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 1–11, 2009.

[9] A. Schasse and R. Martin, "Estimation of subband speech correlations for noise reduction via mvdr processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1355–1365, Sept 2014.

[10] A. Schasse, R. Martin, W. Soergel, T. Pilgrim, and H. Puder, "Efficient implementation of single-channel noise reduction for hearing aids using a cascaded filter-bank," in *Speech Communication; 10. ITG Symposium; Proceedings of*, Sept 2012, pp. 1–4.

[11] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Upper Saddle River, New Jersey: Prentice-Hall, 1993.

[12] D. Mauler and R. Martin, "Improved reproduction of stops in noise reduction systems with adaptive windows and nonstationarity detection," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, p. 469480, 2009. [Online]. Available: http://asp.eurasipjournals.com/content/2009/1/469480

[13] ——, "A low delay, variable resolution, perfect reconstruction spectral analysis-synthesis system for speech enhancement," *15th European Signal Processing Conference (EUSIPCO 2007), Poznan, Poland*, 2007.

[14] K. T. Andersen and M. Moonen, "Adaptive time-frequency analysis for noise reduction in an audio filter bank with low delay," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 784–795, April 2016.

[15] 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," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1368–1381, 2011.

[16] J. Benesty, J. Chen, and Y. Huang, "Noncausal (frequency-domain) optimal filters," *Microphone Array Signal Processing*, pp. 115–137, 2008.

[17] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel wiener filter algorithms for noise reduction with application in cochlear implants," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 4, pp. 785–799, April 2014.

[18] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*. New York: Institute of Electrical and Electronics Engineers, 2000, originally published: New York : Macmillan, 1993.

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

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

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

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

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

[24] 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, Apr 1979.

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

[26] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*. New York, USA: Wiley, 1999.

[27] R. M. Gray, *Toeplitz and Circulant Matrices: A Review*. Hanover, MA, USA: now Publishers Inc., 2006.

[28] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

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

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



Kristian Timm Andersen (S'13) was born in Herlev, Denmark, in 1983. He received the M.Sc. degree in electrical engineering from the Technical University of Denmark, Denmark, in 2008. He is currently pursuing the PhD degree at the Electrical Engineering Department of KU Leuven, and is a member of KU Leuven? STADIUS group. His research interests include audio signal processing, acoustical signal enhancement, and real-time audio applications.



Marc Moonen (M'94, SM'06, F'07) is a Full Professor at the Electrical Engineering Department of KU Leuven, where he is heading a research team working in the area of numerical algorithms and signal processing for digital communications, wireless communications, DSL and audio signal processing.

He received the 1994 KU Leuven Research Council Award, the 1997 Alcatel Bell (Belgium) Award (with Piet Vandaele), the 2004 Alcatel Bell (Belgium) Award (with Raphael Cendrillon), and was a 1997 Laureate of the Belgium Royal Academy of Science. He received journal best paper awards from the IEEE Transactions on Signal Processing (with Geert Leus and with Daniele Giacobello) and from Elsevier Signal Processing (with Simon Doclo).

He was chairman of the IEEE Benelux Signal Processing Chapter (1998-2002), a member of the IEEE Signal Processing Society Technical Committee on Signal Processing for Communications, and President of EURASIP (European Association for Signal Processing, 2007-2008 and 2011-2012).

He has served as Editor-in-Chief for the EURASIP Journal on Applied Signal Processing (2003-2005), Area Editor for Feature Articles in IEEE Signal Processing Magazine (2012-2014), and has been a member of the editorial board of IEEE Transactions on Circuits and Systems II, IEEE Signal Processing Magazine, Integration-the VLSI Journal, EURASIP Journal on Wireless Communications and Networking, and Signal Processing. He is currently a member of the editorial board of EURASIP Journal on Advances in Signal Processing.

2329-9290 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.