Performance Comparison of Single Channel Speech Enhancement Using Speech-Distortion Weighted Inter-Frame Wiener Filters

Klaus Brümann, Dörte Fischer, Simon Doclo *
Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4All, University of Oldenburg, Germany
{klaus.bruemann,doerte.fischer,simon.doclo}@uni-oldenburg.de

Introduction

Speech signals recorded in communication devices are frequently corrupted by undesired additive noise. To improve the speech quality, single-microphone noise reduction is often applied in the short-time Fourier transform (STFT) domain. In contrast to single-frame approaches, where a gain is applied to each noisy STFT coefficient independently, multi-frame approaches aim to exploit the speech inter-frame correlation (IFC) [1, 2, 3, 4].

In this paper, we investigate a real-valued speech-distortion weighted Wiener gain (SDW-WG) as well as real- and complex-valued speech-distortion weighted inter-frame Wiener filters (SDW-IFWFs) [1, 4]. These filters incorporate a trade-off between noise reduction and speech distortion. We compare these filters and analyze the influence of the corresponding trade-off parameter. Experimental results for different speech signals, noise types, and signal-to-noise ratios (SNRs) show that the real-valued SDW-IFWF (R-SDW-IFWF) achieves a higher speech quality improvement than the SDW-WG and complex-valued SDW-IFWF (C-SDW-IFWF). Although the SDW-WG applies more noise reduction than the multi-frame approaches, the C-SDW-IFWF introduces less speech distortion as the level of noise reduction is increased.

Problem Statement

In this section, we introduce the single- and multi-frame signal models.

Single-Frame Signal Model

By applying an STFT with analysis window h_F of length F to the noisy microphone signal, the noisy speech coefficient $Y[f,l]$ with time frame l and frequency bin $f \in \left\{-\frac{F}{2}+1, -\frac{F}{2}+2, \ldots, \frac{F}{2}\right\}$ is obtained. The *single-frame signal model* is defined as

$$\boxed{Y[f,l] = S[f,l] + N[f,l]} \tag{1}$$

where $S[f,l]$ and $N[f,l]$ denote the speech and the noise coefficients, respectively. In single-frame approaches the speech coefficient $S[f,l]$ is estimated by applying a (real-valued) gain $G[f,l]$ independently to each noisy speech coefficient, i.e.,

$$\boxed{\hat{S}[f,l] \;=\; G[f,l]\; Y[f,l]} \tag{2}$$

*This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer 352015383 – SFB 1330 B2 and Cluster of Excellence 1077 Hearing4all.

Multi-Frame Signal Model

Similarly to (1), we apply an STFT to the noisy microphone signal with analysis window h_K of length K, to obtain the noisy speech coefficient $Y[k,l]$ with frequency bin $k \in \left\{-\frac{K}{2}+1, -\frac{K}{2}+2, \ldots, \frac{K}{2}\right\}$, which can be decomposed into the speech coefficient $S[k,l]$ and the noise coefficient $N[k,l]$. The noisy speech vector $\boldsymbol{y}[k,l]$ is defined by considering M consecutive time frames, i.e.,

$$\boldsymbol{y}[k,l] = \left[Y[k,l], Y[k,l-1], \ldots, Y[k,l-M+1]\right]^T, \tag{3}$$

where ^{T} denotes the transpose operator. Similarly to (1), this vector can be written as

$$\boldsymbol{y}[k,l] = \boldsymbol{s}[k,l] + \boldsymbol{n}[k,l] \tag{4}$$

where the speech vector $\boldsymbol{s}[k,l]$ and the noise vector $\boldsymbol{n}[k,l]$ are defined similarly as in (3). In multi-frame approaches the speech coefficient $S[k,l]$ is estimated by applying an M-dimensional (complex-valued) finite impulse response (FIR) filter $\boldsymbol{w}[k,l]$ to the noisy speech vector, i.e.,

$$\boxed{\hat{S}[k,l] = \boldsymbol{w}^H[k,l]\; \boldsymbol{y}[k,l]} \tag{5}$$

where ^{H} denotes the Hermitian operator. For conciseness, in the remainder of this paper the indices f, k, and l will be omitted wherever possible.

Assuming that the speech and noise signals are uncorrelated, the $M \times M$-dimensional noisy speech correlation matrix $\boldsymbol{R_y} = \mathbb{E}\left[\boldsymbol{y}\boldsymbol{y}^H\right]$, with $\mathbb{E}[\cdot]$ the expectation operator, is given by

$$\boldsymbol{R_y} = \boldsymbol{R_s} + \boldsymbol{R_n}, \tag{6}$$

where $\boldsymbol{R_s} = \mathbb{E}\left[\boldsymbol{s}\boldsymbol{s}^H\right]$ and $\boldsymbol{R_n} = \mathbb{E}\left[\boldsymbol{n}\boldsymbol{n}^H\right]$ denote the speech and noise correlation matrices, respectively.

Considering the speech correlation across time frames, it was proposed in [1] to decompose the speech vector $\boldsymbol{s}$ into a temporally correlated speech component $\boldsymbol{x}$ and a temporally uncorrelated speech component $\boldsymbol{x}'$ with respect to the speech coefficient S, i.e.,

$$\boldsymbol{s} = \boldsymbol{x} + \boldsymbol{x}' = \boldsymbol{\gamma_s} S + \boldsymbol{x}', \tag{7}$$

where $\boldsymbol{\gamma_s}$ denotes the normalized speech IFC vector, which is defined as

$$\boldsymbol{\gamma_s} = \frac{\mathbb{E}\left[\boldsymbol{s}S^*\right]}{\mathbb{E}\left[|S|^2\right]} = \frac{\boldsymbol{r_s}}{\phi_S}, \tag{8}$$

where * denotes the complex-conjugate operator and $\boldsymbol{r_s}$ is the speech IFC vector. Due to the normalization with the speech power spectral density (PSD) $\phi_S = \mathbb{E}\left[|S|^2\right]$, the first element of $\boldsymbol{\gamma_s}$ is equal to 1.

Using (6) and (7), the speech correlation matrix $\boldsymbol{R_s}$ can be decomposed into the rank-1 correlation matrix $\boldsymbol{R_x} = \phi_S \boldsymbol{\gamma_s}\boldsymbol{\gamma_s}^H$ and the correlation matrix $\boldsymbol{R_{x'}} = \mathbb{E}\left[\boldsymbol{x'}\boldsymbol{x'}^H\right]$. Hence, the speech IFC vector $\boldsymbol{r_s}$ and the speech PSD ϕ_S in (8) can be computed as

$$\boldsymbol{r_s} = \boldsymbol{R_x}\boldsymbol{e}, \qquad \phi_S = \boldsymbol{e}^T\boldsymbol{R_x}\boldsymbol{e} \tag{9}$$

with $\boldsymbol{e} = \begin{bmatrix}1, & 0, & \ldots, & 0\end{bmatrix}^T$ an M-dimensional selection vector. Considering the uncorrelated speech component $\boldsymbol{x'}$ in (7) as an interference, we define the undesired signal vector $\boldsymbol{u} = \boldsymbol{x'} + \boldsymbol{n}$ such that the *multi-frame signal model* is given by

$$\boxed{\boldsymbol{y} = \boldsymbol{\gamma_s} S + \boldsymbol{u}} \tag{10}$$

Using (10), the noisy speech correlation matrix in (6) can also be written as

$$\boldsymbol{R_y} = \phi_S \boldsymbol{\gamma_s}\boldsymbol{\gamma_s}^H + \boldsymbol{R_u}, \tag{11}$$

with the undesired correlation matrix $\boldsymbol{R_u} = \boldsymbol{R_{x'}} + \boldsymbol{R_n}$. Similarly to (8) and (9), the normalized noisy speech IFC vector $\boldsymbol{\gamma_y}$ and normalized noise IFC vector $\boldsymbol{\gamma_n}$ are defined as

$$\boldsymbol{\gamma_y} = \frac{\boldsymbol{r_y}}{\phi_Y} = \frac{\boldsymbol{R_y}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{R_y}\boldsymbol{e}}, \qquad \boldsymbol{\gamma_n} = \frac{\boldsymbol{r_n}}{\phi_N} = \frac{\boldsymbol{R_n}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{R_n}\boldsymbol{e}}. \tag{12}$$

Using (6) and (12), it can be easily shown that

$$\phi_Y \boldsymbol{\gamma_y} = \phi_S \boldsymbol{\gamma_s} + \phi_N \boldsymbol{\gamma_n}. \tag{13}$$

Speech-Distortion Weighted Filters

In this section, a SDW-WG is derived using the single-frame signal model and a C-SDW-IFWF and R-SDW-IFWF are derived using the multi-frame signal model. The filters incorporate a trade-off between noise reduction and speech distortion.

SDW-WG

A cost-function for the SDW-WG can be designed which aims to minimize the speech distortion power as well as the output noise power, where the importance of each term can be weighted with a trade-off parameter $\mu \in [0, \infty]$, i.e.,

$$\hat{G} = \underset{G}{\operatorname{argmin}} \left\{ \underbrace{E\left[|GS - S|^2\right]}_{\substack{\text{Speech} \\ \text{distortion power}}} + \mu \underbrace{E\left[|GN|^2\right]}_{\substack{\text{Output} \\ \text{noise power}}} \right\}. \tag{14}$$

Solving this optimization problem leads to the real-valued SDW-WG

$$\boxed{G_{\text{SDW-WG}} = \frac{\xi}{\mu + \xi}} \tag{15}$$

with $\xi = \dfrac{\phi_S}{\phi_N}$ the a-priori SNR.

C-SDW-IFWF

Similarly to (14), the aim of the SDW-IFWF is to minimize the speech distortion power as well as the noise power, weighted with μ, i.e.,

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \left\{ \underbrace{E\left[|\boldsymbol{w}^H\boldsymbol{\gamma_s} - S|^2\right]}_{\substack{\text{Speech} \\ \text{distortion power}}} + \mu \underbrace{E\left[|\boldsymbol{w}^H\boldsymbol{u}|^2\right]}_{\substack{\text{Output} \\ \text{noise power}}} \right\}. \tag{16}$$

Solving this optimization problem leads to the C-SDW-IFWF [1]

$$\boldsymbol{w}_{\text{SDW-IFWF}} = \frac{\boldsymbol{R_y}^{-1}\boldsymbol{\gamma_s}\phi_S}{\mu + (1-\mu)\boldsymbol{\gamma_s}^H\boldsymbol{R_y}^{-1}\boldsymbol{\gamma_s}\phi_S}. \tag{17}$$

In [4], it was reported that this filter can be very sensitive to estimation errors for $\mu > 0$. Since it is well known that decomposing a multi-frame Wiener Filter into a multi-frame minimum-power distortionless-response (MFMPDR) filter and a postfilter leads to more robust results [5], we suggest to decompose the C-SDW-IFWF into an MFMPDR filter [1] and a postfilter

$$\boxed{\boldsymbol{w}_{\text{SDW-IFWF}} = \underbrace{\frac{\boldsymbol{R_y}^{-1}\boldsymbol{\gamma_s}}{\boldsymbol{\gamma_s}^H\boldsymbol{R_y}^{-1}\boldsymbol{\gamma_s}}}_{\boldsymbol{w}_{\text{MFMPDR}}} \underbrace{\frac{\phi_S}{\mu\phi_U^{\text{out}} + \phi_S}}_{\substack{G_{\text{SDW-WG}} \\ \text{Postfilter}}}} \tag{18}$$

where $\phi_U^{\text{out}} = \left(\boldsymbol{\gamma_s}^H\boldsymbol{R_u}^{-1}\boldsymbol{\gamma_s}\right)^{-1}$ denotes the undesired signal PSD at the output of the MFMPDR filter.

R-SDW-IFWF

As in [4], a real-valued, symmetric filter vector

$$\boldsymbol{W}[k,l] = \boldsymbol{D}\boldsymbol{w}[k,l] \tag{19}$$

can be derived, where $\boldsymbol{D}$ is a discrete Fourier transform (DFT) matrix. Assuming that the noisy correlation matrix $\boldsymbol{R_y^{circ}}$ is circulant structured, it can be defined as

$$\boldsymbol{R_y^{circ}}[k,l] = \frac{1}{2M}\boldsymbol{D}^H\boldsymbol{\Phi_y}[k,l]\boldsymbol{D}, \tag{20}$$

where $\boldsymbol{\Phi_y}[k,l]$ is a diagonal matrix containing the neighbouring noisy PSD coefficients around the center frequency of a frequency bin k. The matrix $\boldsymbol{\Phi_y}$ is obtained by windowing the PSDs $\phi_Y[f,l]$ in a filterbank with a $\frac{2M}{O}$ higher frequency-resolution, where O denotes the oversampling factor,

$$\boldsymbol{\Phi_y}[k,l](\tau,\tau) = \frac{1}{O}|H_F[\tau]|^2\phi_Y\left[\frac{2Mk}{O} + \tau, l\right], \quad \tau = -M+1, -M+2, \ldots, M, \tag{21}$$

with H_F the F-point DFT of the zero-padded analysis window h_K and $\boldsymbol{\Phi_y}[k,l](\tau,\tau)$ denotes the τ-th diagonal element of the diagonal matrix $\boldsymbol{\Phi_y}[k,l]$. Similar approximations can be made for the correlation matrices $\boldsymbol{R_s^{circ}}$

and $\boldsymbol{R}_{\boldsymbol{n}}^{\boldsymbol{circ}}$ of speech and noise. Using (19) and (20) in (16), a R-SDW-IFWF can be derived

$$\boxed{\boldsymbol{W}_{\text{SDW-IFWF}} = \frac{\boldsymbol{\Phi}_{\boldsymbol{y}}^{-1}\boldsymbol{\Phi}_{\boldsymbol{s}}\boldsymbol{1}}{\mu + (1-\mu)\dfrac{\boldsymbol{1}^T\boldsymbol{\Phi}_{\boldsymbol{s}}\boldsymbol{\Phi}_{\boldsymbol{y}}^{-1}\boldsymbol{\Phi}_{\boldsymbol{s}}\boldsymbol{1}}{\boldsymbol{1}^T\boldsymbol{\Phi}_{\boldsymbol{s}}\boldsymbol{1}}}} \quad (22)$$

This filter can be rewritten as a gain in the higher frequency resolution filterbank with $F = \frac{2MK}{O}$ frequency bins by applying an overlap procedure

$$G[f,l] = \sum_{\nu=-\frac{O}{2}+1}^{\frac{O}{2}} H_K\left[f' + \frac{F}{K}\nu\right]\boldsymbol{W}\left[\frac{K}{F}(f-f')+\nu, l\right]\left(f' + \frac{F}{K}\nu\right), \quad (23)$$

where H_K is the DFT of the analysis window h_K and

$$f' = \text{mod}\left(f + \frac{F}{K} - 1, \frac{F}{K}\right) - \frac{F}{K} + 1, \quad (24)$$

with mod () the modulo operator.

Parameter Estimation

In this section, we present several estimators for the required parameters of the SDW-WG, C-SDW-IFWF, and R-SDW-IFWF.

Real-Valued Filters

For the SDW-WG, an estimate of ξ is required, which is estimated using the decision-directed approach (DDA) in [6] with ϕ_N estimated as in [7] in the high frequency resolution filterbank F.

For the R-SDW-IFWF, estimates of the speech and the noisy speech PSDs are required. The PSDs are estimated using periodograms in the high frequency resolution filterbank F. The noisy speech periodogram is given by

$$P_Y = |Y|^2. \quad (25)$$

The noisy PSD matrix $\boldsymbol{\Phi}_{\boldsymbol{y}}$ is estimated by replacing ϕ_Y with P_Y in (21). The speech and noise periodograms are estimated by applying a Wiener gain (WG) G_{WG} (which is obtained by setting $\mu = 1$ in (15)) to P_Y, i.e.,

$$\hat{P}_S = G_{\text{WG}}P_Y, \qquad \hat{P}_N = (1 - G_{\text{WG}})P_Y, \quad (26)$$

and the speech and noise PSD matrices $\boldsymbol{\Phi}_{\boldsymbol{s}}$ and $\boldsymbol{\Phi}_{\boldsymbol{n}}$ can be estimated similarly to $\boldsymbol{\Phi}_{\boldsymbol{y}}$, by replacing ϕ_Y with $\hat{P}_S$ or $\hat{P}_N$ in (21), respectively.

Complex-valued filters

For the C-SDW-IFWF, estimates of $\boldsymbol{R}_{\boldsymbol{x}}$, $\boldsymbol{\gamma}_{\boldsymbol{s}}$, ϕ_S, and $\boldsymbol{R}_{\boldsymbol{u}}$ are required. The noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}$ can be estimated using first-order recursive smoothing as

$$\hat{\boldsymbol{R}}_{\boldsymbol{y}}[k,l] = \lambda\hat{\boldsymbol{R}}_{\boldsymbol{y}}[k,l-1] + (1-\lambda)\boldsymbol{y}[k,l]\boldsymbol{y}^H[k,l] \quad (27)$$

with λ a forgetting factor. The normalized speech IFC vector $\boldsymbol{\gamma}_{\boldsymbol{s}}$ can be estimated as

$$\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}} = \frac{\hat{\phi}_S + \hat{\phi}_N}{\hat{\phi}_S}\hat{\boldsymbol{\gamma}}_{\boldsymbol{y}} - \frac{\hat{\phi}_N}{\hat{\phi}_S}\frac{\hat{\boldsymbol{r}}_{\boldsymbol{n}}}{\hat{\boldsymbol{r}}_{\boldsymbol{n}}(1)} \quad (28)$$

where $\boldsymbol{\gamma}_{\boldsymbol{y}}$ is estimated similarly to (12), using (27). In [5], we proposed to estimate the noise IFC vector $\boldsymbol{r}_{\boldsymbol{n}}$ from the F filterbank using the Wiener-Khinchin theorem similarly to [4]. The theorem states that the correlation of a wide-sense stationary process is given by the inverse DFT (IDFT) of the PSD. Hence, the noise IFC vector $\boldsymbol{r}_{\boldsymbol{n}}$ can be estimated by applying the IDFT to the noise periodograms in $\hat{\boldsymbol{\Phi}}_{\boldsymbol{n}}$, i.e.,

$$\hat{\boldsymbol{r}}_{\boldsymbol{n}}[k,l](m) = \frac{1}{2M}\sum_{\tau=-M+1}^{M}\hat{\boldsymbol{\Phi}}_{\boldsymbol{n}}[k,l](\tau,\tau)e^{-j2\pi\tau m/2M}, \quad m = 0,1,\ldots,M-1. \quad (29)$$

The speech PSD ϕ_S is estimated by applying a WG to the noisy speech, i.e. $\hat{\phi}_S = G_{\text{WG}}\hat{\phi}_Y$, with ξ estimated using the DDA and ϕ_N estimated using [7]. To estimate the output undesired PSD ϕ_U^{out}, the undesired correlation matrix $\boldsymbol{R}_{\boldsymbol{u}}$ is estimated as

$$\hat{\boldsymbol{R}}_{\boldsymbol{u}} = \hat{\boldsymbol{R}}_{\boldsymbol{y}} - \hat{\phi}_S\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^H. \quad (30)$$

Due to estimation errors, $\hat{\boldsymbol{R}}_{\boldsymbol{u}}$ may not be positive semi-definite, thus, we set negative eigenvalues of $\hat{\boldsymbol{R}}_{\boldsymbol{u}}$ to zero.

Experimental Results

In this section, we begin with describing the algorithmic implementation details and then we compare the performance of the presented SDW-WG and SDW-IFWFs in dependence of the trade-off parameter μ.

Implementation and Performance Measures

The performance compared to the noisy speech signal is evaluated in terms of the perceptual evaluation of speech quality (PESQ) [8] improvement and the segmental measures for speech distortion (sSD) and noise reduction (sNR) [9] as well as SNR improvement (ΔsSNR)[2], using the clean speech signal as the reference signal. We used audio material from [10] sampled at 16 kHz. We evaluated the average performance over 105 s of speech material under five different noise conditions (babble, white Gaussian noise (WGN), traffic, modulated WGN, cross-road) at 0 dB and 10 dB input SNRs.

To achieve a high speech correlation, we use an STFT with a frame length of $K = 64$ samples (4 ms) and a frame shift of 16 samples (1 ms) in the low frequency-resolution STFT filterbank. As analysis and synthesis window h_K we use a Hann window. The number of the consecutive time frames is $M = 8$, resulting in 11 ms of analysis data in the low frequency-resolution filterbank. In the high frequency-resolution STFT filterbank, we use a four-times higher frequency-resolution, i.e., a frame length of $F = 256$ samples (16 ms), a frame shift of 16 samples (1 ms), and apply an asymmetric analysis window similarly to [4]. However, h_K is used as the synthesis window to maintain low synthesis delay (3 ms). In both filterbanks, the weighting parameter for the DDA [6] is set to 0.97. To reduce the amount of musical noise, the Wiener gain is limited to -17 dB. The forgetting factor in (27) is experimentally set to $\lambda = 0.9$, resulting in

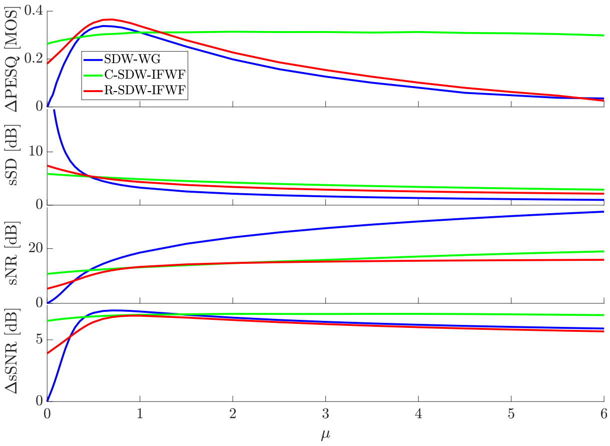

Figure 1: Averaged results at 0 dB SNR.

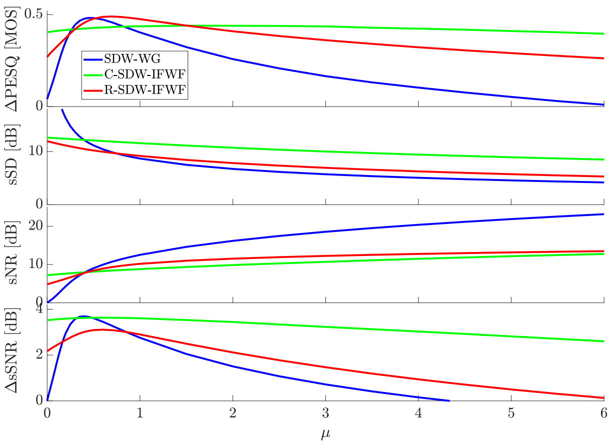

Figure 2: Averaged results at 10 dB SNR.

a smoothing window of 10 ms. Before computing $\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{-1}$ in (18), regularization based on diagonal loading is performed with a regularization parameter of 0.04 as in [2].

Comparison of SDW-IFWFs with SDW-WG

In Figs. 1, 2, the average PESQ, sSD, sNR, and ΔsSNR results at 0 dB and 10 dB SNR are depicted for the SDW-WG, C-SDW-IFWF, and R-SDW-IFWF. The R-SDW-IFWF leads to the highest PESQ improvement with $\mu = 0.7$, followed by the SDW-WG with $\mu = 0.6$ at 0 dB and $\mu = 0.45$ ad 10 dB. The SDW-WG leads to the highest ΔsSNR scores with $\mu = 0.7$ and $\mu = 0.4$ at 0 dB and 10 dB, respectively. The SDW-WG leads to the highest NR scores for increasing μ, however, simultaneously also the lowest SD scores. The C-SDW-IFWF outperforms all filters for all measures except for SD at $\mu = 0$. Only the SDW-WG achieves higher SD scores at $\mu = 0$ since it applies no filtering, leaving the original signal unchanged and therefore the speech undistorted.

CONCLUSION

In this paper, we evaluated the influence of the trade-off parameter in real- and complex-valued speech-distortion weighted filters, using a single- and multi-frame signal model, which balance noise reduction and speech distortion. We compared the performance for different speech and noise signals and signal-to-noise ratios, using practically feasible estimators for the required quantities. We showed that the R-SDW-IFWF achieves the highest speech quality improvement. Although the SDW-WG applies more noise reduction than multi-frame approaches, the C-SDW-IFWF, introduces less speech distortion.

References

[1] 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.

[2] 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.

[3] D. Fischer and S. Doclo, “Sensitivity analysis of the multi-frame MVDR filter for single-microphone speech enhancement,” in *Proc. of Europ. Signal Process. Conf. (EUSIPCO)*, Kos, Greece, Aug. 2017, pp. 603–607.

[4] 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.

[5] D. Fischer, K. Brümann, and S. Doclo, “Comparison of parameter estimation methods for Single-Microphone Multi-Frame wiener filtering,” in *27th European Signal Processing Conference (EUSIPCO)*, submitted.

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

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

[8] “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.

[9] T. Lotter and P. Vary, “Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model,” *EURASIP J. Applied Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, Jan. 2005.

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