Combined Single-Microphone Wiener and MVDR Filtering based on Speech Interframe Correlations and Speech Presence Probability

Dörte Fischer[1], *Simon Doclo*[1], *Emanuël A. P. Habets*[2], *Timo Gerkmann*[3]

[1] University of Oldenburg, Dept. of Medical Physics and Acoustics, Cluster of Excellence "Hearing4all", 26111 Oldenburg, Germany
[2] International Audio Laboratories Erlangen, 91058 Erlangen, Germany
[3] University of Hamburg, Signal Processing, 22527 Hamburg, Germany
Email[1]: {doerte.fischer, simon.doclo}@uni-oldenburg.de
Email[2]: emanuel.habets@audiolabs-erlangen.de
Email[3]: timo.gerkmann@uni-hamburg.de
Web: http://www.sigproc.uni-oldenburg.de

Abstract

For single-microphone noise reduction, a minimum variance distortionless response (MVDR) filter has been recently proposed based on speech correlations of consecutive time frames. This filter is able to keep speech distortion low but compared to conventional approaches achieves less noise reduction. Further, when only having access to the noisy speech, more artifacts in the background noise are audible due to estimation errors of the speech interframe correlations, especially in time-frequency regions where speech is not dominant. Therefore, in this paper we propose to apply the MVDR filter where speech is dominant and the single-channel Wiener filter otherwise, using a weighting based on the speech presence probability. In addition, we modify the decision-directed approach to estimate the *a priori* SNR in a more robust way for short analysis frames. Experimental results show that the proposed scheme achieves a better speech quality compared to the MVDR filter and the single-channel Wiener filter.

1 Introduction

In many speech communication applications, clean speech is affected by additive noise. As a consequence, the speech quality and intelligibility of the target signal decreases with decreasing signal-to-noise ratio (SNR), such that noise reduction algorithms are required. Commonly, single-microphone noise reduction algorithms operate in the short-time Fourier transform (STFT) domain. To obtain an estimate of the clean speech STFT coefficients, typically a multiplicative gain function is applied to the noisy speech signal at each time-frequency point. The most popular examples are the single-channel Wiener filter (WF) [1], the minimum-mean-square error (MMSE) based amplitude estimator [2] and the MMSE log-amplitude estimator [3]. All these approaches assume that consecutive time frames are uncorrelated, such that each time-frequency point can be processed independently. However, it is well known that speech is highly correlated over time and frequency. To incorporate these correlations, Benesty and Huang [4] [5] proposed a single-microphone multi-frame MVDR (MFMVDR) filter exploiting the temporal speech interframe correlations (IFC). For this, the current time frame as well as previous frames are considered. Conceptually, this frame array is similar to a multimicrophone system when interpreting the considered frames as microphone inputs.

The MFMVDR filter achieves impressive results in terms of speech distortions if the speech IFC is perfectly known [4] [5]. Even in a blind implementation where we only have access to the noisy speech signal, the MFMVDR introduces less speech distortion than conventional single-channel algorithms like the Wiener filter [6]. To increase the amount of noise reduction, we recently proposed to combine the MFMVDR with a Wiener post-filter [7], similar to spectral post-filtering for multi-microphone techniques [8]. However, it has been reported that the blindly implemented MFMVDR filter introduces unpleasant artifacts in the background noise [7]. This effect can be mainly observed in time-frequency regions where speech is not dominant, since the estimation of the speech IFC becomes inaccurate and highly variant. To reduce this effect, in this paper we propose to use the WF where speech is not dominant and the MFMVDR where speech is dominant. For this, we apply a soft frequency dependent weighting between the MFMVDR and WF estimates based on the (local) speech presence probability (SPP). With this approach, we are able to benefit both from the good noise reduction performance and the small amount of produced musical tones of the WF while during speech activity the speech distortions can be kept low using the MFMVDR. Furthermore, to reduce the fluctuation of the speech IFC estimation, we propose a modified decision-directed approach (DDA) for short time frames to estimate the *a priori* SNR. It is well-known that the DDA [2] is able to reduce the background noise and musical tones more strongly [9] than other *a priori* SNR estimators like the maximum likelihood (ML) estimator [10]. The evaluation of the proposed algorithm takes place in terms of PESQ [11], where we show that the predicted speech quality performance can be improved with the proposed combination of MFMVDR and Wiener filtering compared to the WF and MFMVDR alone.

The paper is structured as follows. In Section 2 and 3 we define the interframe signal model and briefly review the MFMVDR filter proposed in [4]. In Section 4, we describe the proposed algorithm. We evaluate and conclude our work in section 5 and 6.

2 Signal Model

We consider a single microphone capturing a speech signal that is corrupted by additive noise. In the STFT domain, the noisy complex-valued spectral observation $Y(k,m)$ is given by

$$Y(k,m) = X(k,m) + V(k,m), \tag{1}$$

where $X(k,m)$ denotes the desired speech and $V(k,m)$ the additive noise signal. The indexes k and m denote the frequency bin and time frame, respectively. It is assumed that the speech and noise processes are uncorrelated and that $X(k,m)$ and $V(k,m)$ are complex-valued, zero-mean Gaussian random variables.

The clean speech spectral component $X(k,m)$ is estimated by applying an FIR filter of order $L-1$ with coefficients $H(k,m,l)$ to the noisy speech signal at each time-frequency point as

$$\hat{X}(k,m) = \sum_{l=0}^{L-1} H^*(k,m,l) Y(k,m-l) \tag{2}$$

$$= \boldsymbol{h}^H(k,m)\boldsymbol{y}(k,m). \tag{3}$$

Here, L is the number of consecutive time-frames, * indicates the complex-conjugate operator and ^{H} the Hermitian operator. The vectors $\boldsymbol{h}(k,m)$ and $\boldsymbol{y}(k,m)$ contain the time-varying filter coefficients and the last $L-1$ noisy speech samples, respectively (see Figure 1), i.e.,

$$\boldsymbol{h}(k,m) = [H(k,m,0),\ H(k,m,1),\ \ldots,\ H(k,m,L-1)]^T, \tag{4}$$

$$\boldsymbol{y}(k,m) = [Y(k,m),\ Y(k,m-1),\ \ldots,\ Y(k,m-L+1)]^T.$$

[1]This work was supported by the joint Lower Saxony-Israeli Project ATHENA and the DFG Cluster of Excellence EXC 1077/1 "Hearing4all".

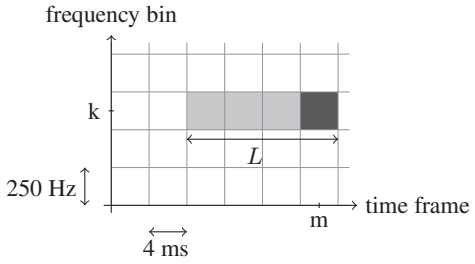


Figure 1: Illustration of the time frames that are taken into account to create the input signal vector $\boldsymbol{y}(k,m)$

According to (1), the L-dimensional vector $\boldsymbol{y}(k,m)$ can be formulated as

$$\boldsymbol{y}(k,m) = \boldsymbol{x}(k,m) + \boldsymbol{v}(k,m), \tag{5}$$

where the clean speech vector $\boldsymbol{x}(k,m)$ and the noise vector $\boldsymbol{v}(k,m)$ are similarly defined as $\boldsymbol{y}(k,m)$ in (4). To take the speech interframe correlations into account, the vector $\boldsymbol{x}(k,m)$ is decomposed into correlated and uncorrelated speech components with respect to the desired signal $X(k,m)$ [4]. Thus, we can define the interframe signal model by rewriting (5) as

$$\begin{aligned}\boldsymbol{y}(k,m) &= \boldsymbol{\rho_x}(k,m)X(k,m) + \boldsymbol{x}'(k,m) + \boldsymbol{v}(k,m), &(6)\\ &= \boldsymbol{\rho_x}(k,m)X(k,m) + \boldsymbol{n}(k,m), &(7)\end{aligned}$$

where, $\boldsymbol{x}'(k,m)$ represents the speech components uncorrelated to the local speech coefficient $X(k,m)$. Since we consider $\boldsymbol{x}'(k,m)$ as an interference, we replaced $\boldsymbol{x}'(k,m)+\boldsymbol{v}(k,m)$ by $\boldsymbol{n}(k,m)$ as the undesired signal vector in (7). The speech interframe coefficient vector $\boldsymbol{\rho_x}(k,m)$, is defined as

$$\boldsymbol{\rho_x}(k,m) = \frac{\mathbb{E}\left[\boldsymbol{x}(k,m)X^*(k,m)\right]}{\mathbb{E}\left[|X(k,m)|^2\right]} = \frac{\boldsymbol{\Phi}_{\boldsymbol{x}X}(k,m)}{\phi_X(k,m)}, \tag{8}$$

with the speech correlation vector $\boldsymbol{\Phi}_{\boldsymbol{x}X}(k,m) = \mathbb{E}[\boldsymbol{x}(k,m)X^*(k,m)]$, the speech power spectral density (PSD) $\phi_X = \mathbb{E}\left[|X(k,m)|^2\right]$ and the operator $\mathbb{E}[\cdot]$ denoting the expectation operator. Due to the normalization, the first element of $\boldsymbol{\rho_x}(k,m)$ is always equal to 1 as $X(k,m)$ is obviously fully correlated with itself.Consequently, the first element of the uncorrelated speech vector $\boldsymbol{x}'(k,m)$ is equal to 0.

3 Multi-Frame MVDR

In this section, we recap the MFMVDR filter presented in [4]. Based on the definition of (3) and (7) the MFMVDR filter can be derived by minimizing the variance of the filtered undesired signal $\boldsymbol{n}(k,m)$ under the constraint that correlated speech components are not distorted, i.e.,

$$\begin{aligned}&\underset{\boldsymbol{h}(k,m)}{\text{argmin}}\ \boldsymbol{h}^H(k,m)\boldsymbol{\Phi_{nn}}(k,m)\boldsymbol{h}(k,m)\\ &\qquad\text{subject to } \boldsymbol{h}^H(k,m)\boldsymbol{\rho_x}(k,m) = 1. \end{aligned}\tag{9}$$

Here, $\boldsymbol{\Phi_{nn}}(k,m) = E[\boldsymbol{n}(k,m)\boldsymbol{n}^H(k,m)]$ denotes the correlation matrix of the undesired signal $\boldsymbol{n}(k,m)$. Solving the problem in (9) leads to the MFMVDR solution [4]

$$\boldsymbol{h}_{\text{MFMVDR}}(k,m) = \frac{\Phi^{-1}_{\boldsymbol{yy}}(k,m)\boldsymbol{\rho_x}(k,m)}{\boldsymbol{\rho_x^H}(k,m)\Phi^{-1}_{\boldsymbol{yy}}(k,m)\boldsymbol{\rho_x}(k,m)}. \tag{10}$$

Applying the MFMVDR filter in (10) to the noisy speech vector $\boldsymbol{y}(k,m)$, the estimated clean speech spectrum $\hat{X}_{\text{MFMVDR}}(k,m)$ is obtained as

$$\hat{X}_{\text{MFMVDR}}(k,m) = \boldsymbol{h}^H_{\text{MFMVDR}}(k,m)\boldsymbol{y}(k,m). \tag{11}$$

In [4] [5] it has been shown that this filter achieves impressive results in terms of speech distortions if the speech correlation coefficient $\boldsymbol{\rho_x}(k,m)$ is perfectly known. Even in a blind implementation where we only have access to the noisy speech signal, it has been shown that the MFMVDR introduces less speech distortions than conventional single-channel algorithms like the Wiener filter [6]. However, applying the MFMVDR filter more artifacts in the background noise are introduced due to inaccurate estimations of $\boldsymbol{\rho_x}(k,m)$ [7].

4 Proposed Algorithm

To reduce the background noise while keeping the musical tones and the speech distortions low, we present a soft frequency weighting of the single-channel WF and MFMVDR filter based on the SPP. Further, we propose a modified DDA for short time frames to reduce fluctuation of the speech IFC estimation and to suppress the background noise and musical noise more strongly. Since we assume to have only access to the noisy speech signal, we need to blindly estimate all required quantities.

4.1 Noisy correlation matrix estimation

The noisy correlation matrix $\boldsymbol{\Phi_{yy}}(k,m)$ is estimated by recursive smoothing [4], i.e.,

$$\hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(k,m) = \lambda\hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(k,m-1) + (1-\lambda)\boldsymbol{y}(k,m)\boldsymbol{y}^{\boldsymbol{H}}(k,m), \tag{12}$$

where λ is the smoothing factor. The first element of the matrix $\hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(k,m)$ corresponds to the noisy speech PSD $\hat{\phi}_Y(k,m)$, i.e., $\hat{\phi}_Y(k,m) = \left[\hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(k,m)\right]_{1,1}$. Before computing the inverse of $\hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(k,m)$, we first perform a matrix regularization [4][6] to improve the robustness of the filter computation, i.e.,

$$\hat{\boldsymbol{\Phi}}^{-1}_{\boldsymbol{yy}}(k,m) = \left(\hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(k,m) + \frac{\delta_{\text{reg}}\,\text{tr}\left[\hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(k,m)\right]}{L}\boldsymbol{I}_{L\times L}\right)^{-1} \tag{13}$$

with a regularization parameter $\delta_{\text{reg}} = 0.04$ as in [4][6]. The operator tr$[\cdot]$ denotes the trace of a matrix and $\boldsymbol{I}_{L\times L}$ is the identity matrix of size $L\times L$.

4.2 Speech IFC estimation

To estimate the clean speech IFC we employ the ML estimator for $\boldsymbol{\rho_x}(k,m)$ proposed in [6], based on the assumption that the noise and speech IFC vectors follow multivariate Gaussian distributions. This ML estimator is given by

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{x}_{\text{ML}}}(k,m) = \frac{\hat{\xi}(k,m)+1}{\hat{\xi}(k,m)}\hat{\boldsymbol{\rho}}_{\boldsymbol{y}}(k,m) - \frac{1}{\hat{\xi}(k,m)}\boldsymbol{\mu}_{\boldsymbol{\rho_v}}. \tag{14}$$

Here, we express the estimated speech IFC $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}_{\text{ML}}}(k,m)$ in terms of the *a-priori* SNR $\xi(k,m) = \frac{\phi_X(k,m)}{\phi_V(k,m)}$ with $\phi_X(k,m)$ and $\phi_V(k,m)$ the speech and noise PSDs, respectively. The vector $\hat{\boldsymbol{\rho}}_{\boldsymbol{y}}(k,m)$ denotes the noisy IFC and is defined similar to the speech IFC in (8). Note that $\boldsymbol{\Phi}_{\boldsymbol{y}Y}(k,m) = \left[\hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(k,m)\right]_{:,1}$, where $[\cdot]_{:,1}$ denotes the first column of a matrix. The parameter $\boldsymbol{\mu}_{\boldsymbol{\rho_v}}$ is the mean of the noise IFC. It is assumed to be given by the frame overlap and the analysis window function [6]. The quantity is fixed for all time-frequency points.

4.3 Proposed *a priori* SNR estimation

It is well known that the DDA [2] is able to efficiently reduce background noise and musical tones [9], by providing smoother estimates than for instance the positively constrained ML estimator of the speech PSD $\phi_X(k,m)$ used in [6] [7]. This ML estimator is given by [10]

$$\hat{\phi}_{X_{\text{ML}}}(k,m) = \max\left[\hat{\phi}_Y(k,m) - \hat{\phi}_V(k,m), 0\right]. \tag{15}$$

ISBN 978-3-8007-4275-2

However, due to the use of short analysis frames (4 ms), which are typically used in MFMVDR filters, and the nonstationarity of speech, outliers in $\hat{\phi}_{X_{\text{ML}}}(k,m)$ and the *a priori* SNR $\xi(k,m)$ negatively affect the speech IFC estimate and may result in annoying artifacts in the processed speech [6] [7]. Thus, to estimate $\xi(k,m)$, we propose a modified DDA. For this, we propose to use temporally smoothed observations and estimates in the DDA, as

$$\hat{\xi}(k,m) = \alpha \frac{\bar{A}(k,m-1)}{\hat{\phi}_V(k,m-1)} + (1-\alpha)\max\left[\bar{\gamma}(k,m)-1,0\right] \quad (16)$$

where α is a weighting factor. The higher α is set, the more noise reduction and less musical noise are obtained, but the more speech distortions are introduced. The smoothed *a-posteriori* SNR $\bar{\gamma}(k,m)$ and the smoothed estimated speech periodogram $\bar{A}(k,m-1)$ are defined as

$$\bar{\gamma}(k,m) = \frac{1}{T}\sum_{t=0}^{T-1} \frac{|Y(k,m-t)|^2}{\hat{\phi}_V(k,m)}, \quad (17)$$

and

$$\bar{A}(k,m) = \frac{1}{T}\sum_{t=0}^{T-1} |\hat{X}(k,m-t)|^2, \quad (18)$$

where T is the length of the smoothing window. The larger T, the less musical noise is produced but the more speech distortions are introduced. Note that for $T=1$, the equation (16) reduces to the traditional DDA in [2].

For estimating the noise PSD $\phi_V(k,m)$, we apply the simple noise PSD estimator proposed in [12], i.e.,

$$\hat{\phi}_V(k,m) = \min\left[\hat{\phi}_Y(k,m), \hat{\phi}_V(k,m-1)\right](1+\epsilon). \quad (19)$$

The parameter ϵ controls the maximum speed and is set to 5 dB/s as in [6] [7].

4.4 Proposed speech spectrum estimation

In [7], we reported that the single-channel WF achieves less artifacts in the background noise than the MFMVDR filter. This is due to estimation errors of the speech IFC, especially in speech pauses. In practice, i.e. when parameters are estimated blindly, the ML estimator of $\boldsymbol{\rho_x}$ in (14) fluctuates strongly in time-frequency regions where speech is not dominant. These fluctuations may result in annoying musical noise in the processed speech. To reduce this effects, we propose to use the MFMVDR where speech is dominant and the single-channel WF otherwise. For this, we apply a frequency dependent SPP to achieve a soft transition between the MFMVDR and the WF. The estimated clean speech spectrum $\hat{X}(k,m)$ is obtained as

$$\hat{X}(k,m) = p(k,m)\hat{X}_{\text{MFMVDR}}(k,m) + (1-p(k,m))\hat{X}_{\text{WF}}(k,m), \quad (20)$$

where $p(k,m)$ is the SPP $P(H_1(k,m)|\bar{\gamma}(k,m))$ with $H_1(k,m)$ as the hypothesis of speech presence. Here, $\hat{X}_{\text{MFMVDR}}(k,m)$ is given by (11) and $\hat{X}_{\text{WF}}(k,m)$ is the processed speech spectrum given by

$$\hat{X}_{\text{WF}}(k,m) = H_{\text{WF}}(k,m)Y(k,m), \quad (21)$$

with the Wiener gain $H_{\text{WF}}(k,m)$ defined as

$$H_{\text{WF}}(k,m) = \max\left(\frac{\hat{\xi}(k,m)}{1+\hat{\xi}(k,m)}, H_{\min}\right). \quad (22)$$

Note that in this case L is equal to 1. The parameter $H_{\min}$ is a lower limit of the Wiener gain to reduce the effect of speech distortions. Further, if the SPP $p(k,m)$ is 1 in (20) the WF will not be considered just like the MFMVDR is neglected when $p(k,m)=0$ for all k,m.

4.5 SPP estimation

To estimate the speech spectrum $\hat{X}(k,m)$ using (20), we need to estimate the *a posteriori* SPP $p(k,m) = P(H_1(k,m)|\bar{\gamma}(k,m))$. For this, we employ the SPP estimator based on a smoothed observation $\bar{\gamma}$ proposed in [13], i.e.,

$$p(k,m) = P\left(H_1(k,m)|\bar{\gamma}(k,m)\right) = \frac{\Lambda(k,m)}{1+\Lambda(k,m)}, \quad (23)$$

where Λ is the generalized likelihood ratio (GLR). The GLR is defined as the weighted ratio of the likelihoods of speech presence and speech absence, i.e.,

$$\Lambda(k,m) = \frac{q}{1-q}\left(\frac{1}{1+\xi_{H_1}}\right)^{\frac{\bar{r}}{2}} \exp\left\{\frac{\xi_{H_1}}{1+\xi_{H_1}}\frac{\bar{r}}{2}\bar{\gamma}(k,m)\right\}. \quad (24)$$

In [13], the smoothed *a posteriori* SNR $\bar{\gamma}$ is modeled by a chi-square distribution parameterized by the degrees-of-freedom $\bar{r}$, where for a larger $\bar{r}$ the more smoothing is applied to $\bar{\gamma}$. The parameter ξ_{H_1} denotes a fixed *a priori* SNR which reflects the SNR that is typical if speech were present and $q = P(H_1)$ is the *a priori* SPP. It can be used to bias the GLR to speech presence ($q > 0.5$) or to speech absence ($q < 0.5$). Here, we set $q = 0.5$.

5 Evaluation

In this section, we compare the performance of the proposed SPP-based combination of the MFMVDR filter and the single-channel WF to only applying the MFMVDR filter ($p(k,m)=1$) and the WF ($p(k,m)=0$), as well as to the MFMVDR according to [6] and the classical WF with a larger frame length, corresponding to a higher frequency resolution. The main difference between both considered MFMVDR filters is way of the *a priori* SNR is estimated. For the MFMVDR according to [6], the ML estimator of the speech in PSD (15) is used, whereas for the MFMVDR ($p(k,m)=1$) the proposed modified DDA in (16) is applied.

For all considered techniques (except for the classical WF), we employ a high temporal resolution with a frame length of 4 ms with a frame shift of 1 ms to increase the exploitable IFC. For the classical WF, we use a frame length of 19 ms, an overlap of 50 % and the traditional DDA ($T=1$) with $\alpha = 0.97$. The lower limit h_{min} of the WF in (22) is set to -17 dB. The sampling rate is $f_s = 16$ kHz. For spectral analysis and synthesis, we employ a square-root Hann window.

For the SPP-based combination of the MFMVDR and WF, the required parameters L, λ, T and α were experimentally optimized. For several parameter combinations the PESQ [11] scores were averaged over 30 TIMIT sentences corrupted by white Gaussian noise at 0 dB SNR. The parameter settings with a good compromise between PESQ performance and informal listening impression can be found in Table 1a. For the SPP estimator we used the parameters according to [13] (see Table 1b). However, in contrast to [13], we only considered the local SPP and averaged the speech over 19 ms since this corresponds to the largest considered time window for both the MFMVDR and WF (see Table 1a).

Filter	L	λ	T	α
MFMVDR	18 (21 ms)	0.88	16 (19 ms)	0.65
WF	1 (4 ms)	0.88	16 (19 ms)	0.65

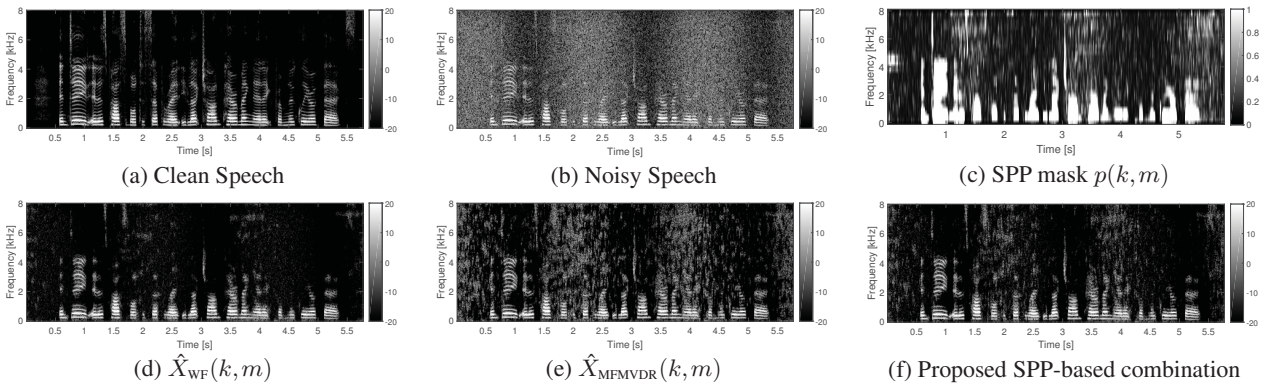
(a)

Δk	Δl +1	N	$\bar{r}$	ξ_{H_1}
1	16 (19 ms)	48	2.11	15 dB

(b)

Table 1: Parameter settings for (a) the proposed SPP-based combination and (b) the (local) SPP estimator based on [13] with a time averaging window of 19 ms.

ISBN 978-3-8007-4275-2



(a) Clean Speech (b) Noisy Speech (c) SPP mask $p(k,m)$

(d) $\hat{X}_{\mathrm{WF}}(k,m)$ (e) $\hat{X}_{\mathrm{MFMVDR}}(k,m)$ (f) Proposed SPP-based combination

Figure 2: Spectrograms of the (a) clean speech, (b) noisy speech, (c) the SPP mask and (d)-(f) the resulting processed speech from a female speaker corrupted by modulated white Gaussian noise at 5 dB SNR.

We first present results for a female speech signal corrupted with modulated white Gaussian noise at 5 dB SNR. Figure 2 depicts the spectrograms of the clean speech, noisy speech and the processed speech of the proposed algorithm, the MFMVDR ($p(k,m) = 1$) and the WF ($p(k,m) = 0$). It can be clearly seen that the WF decreases the background noise but attenuates the speech components resulting in speech distortion. The MFMVDR yields less speech distortion, but also less noise reduction. The proposed SPP-based weighted combination results in a bit less noise reduction than the WF, especially at high frequency bins, but less speech distortions than the WF alone. Informal listening tests confirm the results. Applying the proposed SPP-based combination, clearly less artifacts are audible than with the MFMVDR but slightly more than with the WF. However, the speech sounds less distorted and more natural with the proposed combination than with the WF.

For further analysis, we compare the PESQ [11] improvements of the proposed SPP-based combination to only applying the MFMVDR ($p(k,m) = 1$) and the WF ($p(k,m) = 0$), as well as to the MFMVDR in [6] and the classical WF. Average PESQ improvements were computed over 60 sentences from the TIMIT database [14] spoken by different speakers (5 male, 5 female), corrupted by modulated white Gaussian noise with a modulation frequency of 0.5 Hz, pink and traffic noise.

In Figure 3, the averaged PESQ scores over all evaluated speech and noise files are shown for different SNRs. It can be seen that for a wide range of SNRs the proposed SPP-based combination outperforms only applying the MFMVDR ($p(k,m) = 1$) and the WF ($p(k,m) = 0$). Comparing both MFMVDR implementations, the MFMVDR with $p(k,m) = 1$ exhibits larger improvements than the MFMVDR in [6] where both are considerably worse than the proposed algorithm. The WF with higher frequency resolution performs almost identical to the WF with a lower frequency resolution for SNRs over 0 dB. Considering the average performance at 0 dB SNR, the proposed approach performs 0.04 MOS better than the WF ($p(k,m) = 0$) and 0.02 MOS better than the classical WF, as well as 0.09 and 0.14 MOS better than the MFMVDR [6] and MFMVDR ($p(k,m) = 1$), respectively.

Since the WF is designed to minimize the mean-squared error between the clean speech signal and the estimated speech, the WF reduces noise well but also results in speech distortions (see Figure 2). The MFMVDR is designed to avoid speech distortion, it leads to less noise reduction than the WF. Applying the proposed combination of both the WF and MFMVDR leads to a bit less noise reduction than the WF (see Figure 2) while the speech quality predicted by PESQ can be improved (see Figure 3). Further, comparing both considered MFMVDR implementations, the MFMVDR ($p(k,m) = 1$) results in a bit better predicted speech quality than the MFMVDR in [6]. Consequently, the proposed *a priori* SNR estimator leads to a smoother estimate of the speech IFC than the ML estimator of the speech PSD. Thus, estimation

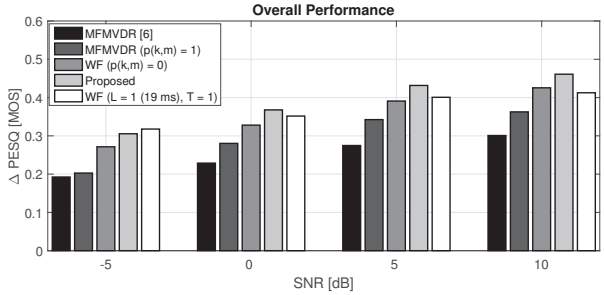


Figure 3: Averaged noise reduction performance of the WF ($p(k,m) = 0$), the MFMVDR ($p(k,m) = 1$) and the proposed SPP-based combination of both as well as the MFMVDR according to [6] an the classical WF with 19 ms analysis frames.

errors of the speech IFC can be reduced and a better speech quality predicted by PESQ can be achieved. In addition, increasing the frequency resolution for the WF leads to a comparable performance as for the WF with lower frequency resolution and the modified DDA. Moreover, combining the WF and MFMVDR estimates with a soft frequency dependent SPP leads to clearly less musical tones and more noise reduction than the MFMVDR and the speech is less distorted and sounds more natural than with the WF.

6 Conclusions

In this paper we consider multi-frame MVDR (MFMVDR) filtering for single-microphone speech enhancement. Recently, it has been shown that the blindly implemented MFMVDR filter introduces less speech distortions than the WF but achieves a lower noise reduction performance and more musical noise [6] [7]. Furthermore, typically the musical noise artifacts in the MFMVDR filtering result from erroneous estimates of the interframe speech correlations in speech absence. To reduce this effect, in this paper we proposed to use the MFMVDR where speech is dominant and the WF otherwise, controlled by an estimate of the speech presence probability. Further, for estimating the *a priori* SNR we modified the decision-directed approach to work robustly with the short signal segments typically used in MFMVDR filtering. The proposed combination of MFMVDR and Wiener filtering achieves a better PESQ score as the WF and MFMVDR alone. Furthermore, with the proposed combination and *a priori* SNR estimation, we achieved clearly less musical tones and more noise reduction than with the MFMVDR alone.

ISBN 978-3-8007-4275-2

References

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

[2] 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, pp. 1109–1121, Dec. 1984.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, pp. 443–445, Apr. 1985.

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

[5] 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, pp. 1256–1269, May 2012.

[6] A. Schasse and R. Martin, "Estimation of subband speech correlations for noise reduction via MVDR processing," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, pp. 1355–1365, Sept. 2014.

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

[8] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, pp. 39–60, Springer, 2001.

[9] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, 1994.

[10] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, pp. 137–145, Apr 1980.

[11] "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.

[12] E. Hänsler and G. Schmidt, *Acoustic echo and noise control: a practical approach*, vol. 40. John Wiley & Sons, 2005.

[13] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 5, pp. 910–919, 2008.

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

ISBN 978-3-8007-4275-2