A DATA-DRIVEN POST-FILTER DESIGN
BASED ON SPATIALLY AND TEMPORALLY SMOOTHED A PRIORI SNR

Huajun Yu, Tim Fingscheidt

Technische Universität Braunschweig, Institute for Communications Technology,
Schleinitzstr. 22, 38106 Braunschweig, Germany
{yu, fingscheidt}@ifn.ing.tu-bs.de

ABSTRACT
A microphone array beamformer combined with a post-filter estimation based on spatial smoothing can deliver good noise attenuation preserving the speech component, however, with disturbing musical tones. On the other hand, the temporal smoothing of the decision-directed (DD) a priori signal-to-noise ratio (SNR) estimation for single-channel noise reduction can suppress musical tones well, however, with speech distortion particularly in speech onset. Based on these facts, we derive a new data-driven multi-channel a priori SNR estimation based on both spatial and temporal smoothing for the use in a beamformer post-filter. The new a priori SNR estimation is able to find an optimum compromise between noise attenuation, quality of the speech component, and musical tones suppression.

Index Terms—a priori SNR estimation, post-filter, speech enhancement

1. INTRODUCTION
A microphone array-based beamformer with post-filter is an attractive means for noise reduction. A typical beamformer is the minimum variance distortionless response (MVDR) design [1]. Unfortunately, using a beamformer alone does not deliver enough noise attenuation, especially for car noise. Therefore, a post-filter has to be applied after the beamformer. McCowan et al. have proposed a post-filter, based on auto and cross power spectral densities (psd) from different channels by considering the noise coherence matrix [2]. Introducing an adaptive smoothing factor for the auto and cross pisd estimation [3], McCowan’s post-filter can further be improved both in terms of noise attenuation and speech distortion. However, musical tones are still clearly perceived.

In contrast, the successful decision-directed (DD) approach to single-channel a priori SNR estimation [4] employs temporal smoothing. By applying the DD approach (i.e., by assigning the smoothing factor for the previous frame component very close to unity), musical tones can be significantly reduced [5]. However, due to the high dependence on the estimate of the previous frame, the DD approach cannot react quickly on an abrupt instantaneous SNR change, leading to speech distortion in speech onset. Motivated by the fact, that the problems of the beamforming post-filter and the DD approach are complementing each other, we present a new a priori SNR estimator which utilizes both spatial and temporal smoothing with the aim of reducing noise under the constraint that musical tones and speech distortion do not exceed certain levels. The spatial smoothing part in the new a priori SNR estimation follows McCowan’s post-filter, but estimates the psd of speech and noise separately (as in [6]), while temporal smoothing uses a relaxed DD approach. Inspired by data-driven approaches to spectral weighting rules [7, 8, 9], we show how an optimal combination of both smoothing processes can be found in a data-driven fashion, resulting in a simple and efficient solution.

The paper is organized as follows: In Section 2 baseline speech enhancement techniques with a priori SNR estimation will be recapitulated. The new a priori SNR estimation using controlled spatial and temporal smoothing will be presented in Section 3. Three instrumental measures are introduced in Section 4, and used in Section 5 to optimize weighting factors of the new a priori SNR estimation approach.

2. BASELINE SPEECH ENHANCEMENT
In the discrete Fourier transform domain, the vector of microphone array signals at frame ℓ and frequency bin ℓ,k can be formulated as $Y(ℓ,k) = S(ℓ,k) · D(ℓ,k) + N(ℓ,k)$, with $N(ℓ,k)$ being the additive noise and $D(ℓ,k)$ being the propagation vector, representing the delays of the desired single-channel source signal $S(ℓ,k)$ for each microphone based on a reference microphone, respectively. After delay compensation, the time-aligned microphone signals can be reformulated as $Y(ℓ,k) = S(ℓ,k) · I + N(ℓ,k)$, with $I$ being the identity matrix.

2.1. Multichannel Spatial Smoothing Approach
It is known that the multichannel Wiener filter can be decomposed into an MVDR beamformer with a Wiener post-filter. The Wiener post-filter can be defined as

$$H_{WF}(ℓ,k) = \frac{E\{ \lvert S(ℓ,k) \rvert^2 \}}{E\{ \lvert N(ℓ,k) \rvert^2 \} + E\{ \lvert S(ℓ,k) \rvert^2 \}}, \quad (1)$$

with $E\{ \lvert N(ℓ,k) \rvert^2 \}$ being the expectation of the squared noise magnitude at the post-filter input.

McCowan et al. have proposed a post-filter estimation [2] taking into account the noise field coherence. The numerator and denominator of (1) are then separately computed by applying a spatial smoothing of the estimated auto and cross pisd $\hat{\phi}_{YY}$ and $\hat{\phi}_{Y_iY_j}$ for all $M$ channels as

$$E\{ \lvert S(ℓ,k) \rvert^2 \} = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \hat{\phi}_{SS}^{(ij)}(ℓ,k), \quad (2)$$

$$E\{ \lvert N(ℓ,k) \rvert^2 \} + E\{ \lvert S(ℓ,k) \rvert^2 \} = \frac{1}{M} \sum_{i=1}^{M} \hat{\phi}_{Y_iY_i}(ℓ,k), \quad (3)$$

with $\hat{\phi}_{SS}^{(ij)}(ℓ,k)$ being estimated as

$$\hat{\phi}_{SS}^{(ij)}(ℓ,k) = \frac{\text{Re} \{ \hat{\phi}_{Y_iY_j}(ℓ,k) \} - \Gamma_{ij}(ℓ,k) \beta_{ij}(ℓ,k)}{1 - \Gamma_{ij}(ℓ,k)}, \quad (4)$$
where \( \beta_{ij}(\ell, k) = \frac{1}{2} \left[ \hat{\phi}_{Y_iY_i}(\ell, k) + \hat{\phi}_{Y_jY_j}(\ell, k) \right] \) and \( P[x] = \max \{ x, 0 \} \). The \( \Re \{ \cdot \} \) operator is used to enforce a real-valued \( \hat{\phi}_{SS}(\ell, k) \), while \( \Gamma_{ij}(k) \) is the real-valued coherence function for the diffuse noise field. The auto psd \( \hat{\phi}_{Y_iY_i}(\ell, k) \) and cross psd \( \hat{\phi}_{Y_jY_j}(\ell, k) \) are estimated recursively with a constant smoothing factor \( \varphi \) and the conjugate complex operator \( \cdot \) as

\[
\hat{\phi}_{Y_iY_i}(\ell, k) = \alpha \hat{\phi}_{Y_iY_i}(\ell-1, k) + (1-\alpha) Y_i^*(\ell, k) Y_i(\ell, k), \\
\hat{\phi}_{Y_jY_j}(\ell, k) = \alpha \hat{\phi}_{Y_jY_j}(\ell-1, k) + (1-\alpha) Y_j^*(\ell, k) Y_j(\ell, k).
\]

Inserting (2) and (3) into (1), McCowan’s post-filter is written as

\[
H_{MC}(\ell, k) = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \hat{\phi}_{SS}(\ell, k),
\]

In our previous works [3], we proposed an adaptive smoothing factor \( \alpha(\ell, k) = \alpha_1 - \alpha_2 \cdot H_{MC}(\ell-1, k) \), \( (\alpha_1, \alpha_2) = (0.8, 0.5) \), for the auto and the cross psd estimation in (5) and (6). In consequence, the noise attenuation performance and the preservation of the quality of the speech component have been significantly improved, while musical tones can still be perceived as in [2].

### 2.2. Single-Channel Temporal Smoothing Approach

The problem of musical tones has been well addressed within the classical tones. Based on this from the post-filter estimate. Hence, in analogy to the derivation of the clean speech signal psd \( \hat{\phi}_{SS}(\ell, k) \) in (4), the noise psd \( \hat{\phi}_{NN}(\ell, k) \) can be derived as [6]

\[
\hat{\phi}_{NN}(\ell, k) = P \left[ \beta_{ij}(\ell, k) - \Re \left\{ \hat{\phi}_{Y_iY_i}(\ell, k) \right\} \right].
\]

with an overestimation factor \( \alpha_oe \), smoothing factor \( \beta_1 = 1 - \beta_2 \), the enhanced speech signal \( \hat{S}(\ell, k) \) in the previous frame, \( \hat{\phi}_{NN}(\ell-1, k) \) being the noise psd computed via minimum statistics [10] in the previous frame, and \( \gamma(\ell, k) = \frac{\gamma_{NN}(\ell, k)}{\hat{\phi}_{NN}(\ell, k)} \) being the a posteriori SNR. Setting the weighting factor \( \beta_1 \) close to unity, the DD approach indeed yields a strong temporal smoothing for the a priori SNR estimate, which helps to significantly reduce musical tones. Based on this a priori SNR estimate the Wiener filter in (1) can be reformulated according to [11] as

\[
H_{WF}(\ell, k) = \frac{\xi(\ell, k)}{1 + \xi(\ell, k)}.
\]

Unfortunately, the strong temporal smoothing reduces the capability of the DD approach to follow speech onsets, which leads to speech distortion in these cases.

### 3. NEW A PRIORI SNR ESTIMATION

In order to adopt the DD approach in post-filter estimation, it is necessary at first to derive a spatial smoothing based a priori SNR from the post-filter estimate. Hence, in analogy to the derivation of the clean speech signal psd \( \hat{\phi}_{SS}(\ell, k) \) in (4), the noise psd \( \hat{\phi}_{NN}(\ell, k) \) can be derived as [6]

\[
E\{|N(\ell, k)|^2\} = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \hat{\phi}_{SS}^{(i,j)}(\ell, k),
\]

which is a variance-reduced estimate by using spatial smoothing as in (2). Please note, no voice activity detector is needed, since the coherence function \( \Gamma_{ij}(k) \) is used for both speech presence and absence and (10) implicitly yields a good estimate in both cases. In practice, due to the deficiency of the MVDR beamformer at low frequencies [6], (11) ignores any noise reduction effect of the beamformer itself in deriving \( E\{|N(\ell, k)|^2\} \), since car noise has most energy in low frequencies.

Given \( E\{|S(\ell, k)|^2\} \) and \( E\{|N(\ell, k)|^2\} \) being computed via (2) and (11), respectively, we are able to derive a spatial smoothing based a priori SNR estimate as

\[
\xi_{MC}(\ell, k) = \frac{\alpha_oe \cdot \hat{\phi}_{NN}(\ell-1, k)}{2 \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \hat{\phi}_{SS}^{(i,j)}(\ell, k)},
\]

If no \( P[\cdot] \) functions were used, and \( \alpha_oe = 1 \), (12) inserted into (9) would result in McCowan’s post-filter (7). Unlike the strong temporal smoothing in the DD approach (8), the term \( \xi_{MC}(\ell, k) \) is only indirectly based on a temporal smoothing for the auto and the cross psd estimation of microphone signals as given in (5) and (6), which is, however, quite relaxed with its adaptive smoothing factor \( \alpha(\ell, k) \in [0, 3, 0.8] \). In a direct manner, it represents spatial smoothing rather than temporal smoothing. Accordingly, we can also interpret \( \xi_{MC}(\ell, k) \) as an instantaneous a priori SNR estimate, which can follow the transient a priori SNR change much better. Combining the spatial smoothing (12) and the temporal smoothing (8), we propose to compute a new multichannel a priori SNR estimation as follows

\[
\xi_{NEW}(\ell, k) = \beta_1 \cdot \frac{\hat{S}(\ell, k)^2}{\alpha_oe \cdot \hat{\phi}_{NN}(\ell-1, k)} + \beta_2 \cdot P[\gamma(\ell, k) - 1] + \beta_3 \cdot \xi_{MC}(\ell, k),
\]

with \( \hat{\phi}_{NN}(\ell-1, k) \) being computed by minimum statistics [10] on the basis of the post-filter input signal (WF-NEW). Three weight factors are used: \( \beta_1 \) and \( \beta_2 \) correspond to the temporal smoothing based components, while \( \beta_3 \) is assigned to the spatial smoothing based component, respectively. Fig. 1 shows the possible vectors of weighting factors \( \beta = (\beta_1 \beta_2 \beta_3) \) as a shadowed triangle in a unity cubic space, fulfilling \( \sum_{i=1}^{3} \beta_i = 1 \) and \( \beta_i \geq 0 \). If we set the vector of weighting factors in (13) to \( \beta = (0.98 0.02 0) \) in (13), which is marked as point \( A \) in Fig. 1, we exactly obtain \( \xi_{MC}(\ell, k) \) given in (12). In the same way, setting \( \beta = (0.98 0.02 0) \) in (13), which is marked as point \( B \), yields exactly the classical \( \xi_{DD}(\ell, k) \) as given in (8). In this manner, our new proposed a priori SNR estimation provides a framework, which is capable of continuously combining the benefits of both spatial smoothing and temporal smoothing. Finding the best vector \( \beta \) in a data-driven fashion will be described in Section 5.1.

### 4. INSTRUMENTAL MEASURES

For optimization of \( \beta \) and evaluation of the proposed post-filter approach three instrumental measures are used: The amount of noise attenuation, speech distortion, and musical tones suppression. The reference clean signal \( s(n) \) is chosen as the best clean speech signal from one channel of the multichannel clean speech
recordings. The reference noise signal \( n(n) \) is chosen accordingly from the same channel. By separately processing the multichannel clean speech signals and background noises with a beamformer and one of the evaluated pre-calculated post-filters, we can get the filtered clean signal \( \hat{s}(n) \) and filtered noise signal \( \hat{n}(n) \). The noise attenuation performance is evaluated in terms of \( \Delta SNR = SNR_{out} - SNR_{in} \) with \( SNR_{in} \) being the input signal-to-noise ratio at any microphone (they are approximately equal), and \( SNR_{out} \) being the output signal-to-noise ratio after the post-filter. \( SNR_{out} \) is calculated from \( \hat{s}(n) \) and \( \hat{n}(n) \). Secondly, to assess distortion of the speech component, the perceptual evaluation of speech quality mean opinion score (PESQ-MOS) of \( \hat{s}(n) \) is computed with reference to the clean speech signal \( s(n) \) according to [12].

It has been shown in [13] that the lower the log kurtosis ratio is, the less musical tones will be perceived, which has also been verified with subjective tests in [13]. Therefore, a modified log kurtosis ratio defined as \( \Delta \Psi_{log} = \log \left( \sum_{l,k=1}^{K-1} \sum_{l,k=1}^{K} \mu \cdot \Delta \Psi_{\ell,k} \right) \) based on \( n(n) \) and \( \hat{n}(n) \) is used to measure the amount of musical tones suppression in this paper. For the background noise the gamma distribution model will be applied. The kurtosis \( \Psi_{\ell,k}(n,k) \) of the background noise signal \( n(n) \) can then be computed like in [13] as \( \Psi_{\ell,k}(n,k) = \frac{(\mu+2)(\mu+3)}{\mu(\mu+1)^2} \), with \( \mu = \frac{\sum_{l,k=1}^{K-1} \sum_{l,k=1}^{K} \Delta \Psi_{\ell,k} \cdot \mu}{\sum_{l,k=1}^{K-1} \sum_{l,k=1}^{K} \mu} \).

In the same way, the kurtosis \( \Psi_{\ell,k}(s,k) \) can be determined by employing \( \hat{n}(n) \) and \( \hat{n}(n) \), however, computed only in speech absence, in order to get rid of the influence of speech-modulated components in \( \hat{n}(n) \). Evaluated in car noise, musical tones exist mostly in low frequencies. Therefore, the quantities \( \Psi_{\ell,k}(s,k) \) and \( \Psi_{\ell,k}(s,k) \) will be computed here only in frequency components up to 3000 Hz.

Different to [13], the use of \( n(n) \) and \( \hat{n}(n) \) and the definition of \( \nu \) ensure that the measure \( \Delta \Psi_{log} \) is independent of \( \Delta SNR \) and PESQ-MOSs, which is desirable for the optimization of the weighting factors in (13). We now have separate measures for SNR improvement, speech component distortion, and musical tones (as a specific noise distortion).

### 5. Optimization and Evaluation

Optimization and evaluation of the new approach is based on signals acquired from a microphone array with four low-cost microphones having equal distances of 3.6 cm integrated in a car’s head-unit. Multiple recordings for each channel are made separately for clean speech signals and background noises. For our experiment we let the car engine run in idle state with windows closed and air condition level being set to 50%. The input signal-to-noise ratios \( SNR_{in} \) are chosen to be -5 dB, 0 dB, 5 dB, 10 dB, and 15 dB. Seven clean speech signals and noise signals are recorded for each noise condition, respectively, each of 8 s. Thus we obtain 7 × 7 = 49 noisy signals for each \( SNR_{in} \) level, 14 signals will be used to optimize weighting factor vector \( \beta \) and 35 signals are used for evaluation. The sampling frequency \( f_s = 16 \text{ kHz} \) is used for all signals. Signals are windowed by a Hann window of length 512, followed by an FFT with length \( K = 512 \) and a frame shift of 50%.

### 5.1. Optimization of Weighting Factors

The weighting factors \( \beta_1, \beta_2 \), and \( \beta_3 \) in (13) are trained using the \( \Delta SNR \), PESQ-MOSs, and the log kurtosis ratio \( \Delta \Psi_{log} \) as measures of quality. The new Wiener filter (WF-NEW) defined by (13) and (9) is applied with \( \xi_{min} = -15 \text{ dB} \) and \( \alpha_{oe} = 1 \). Two other Wiener filter-based reference post-filters are used for the optimization:

#### 5.1.1. The Siweri post-filter

The simpler reference approach is the Wiener post-filter (WF-DD) using (8) and (9) with \( \alpha_{oe} = 1.4 \) and \( \xi_{min} = -10 \text{ dB} \), meaning \( \beta = (0.98 \ 0.02 \ 0) \). A further multichannel reference, McCowan’s post-filter (MC) as given in (12) and (9) (or, approximately, equally to (7)), using the adaptive smoothing factor as proposed in [3] with \( \xi_{min} = -15 \text{ dB} \) and \( \alpha_{oe} = 1 \), meaning \( \beta = (0 \ 0 \ 0) \).

For the new approach, a two-step optimization is implemented for each \( SNR_{in} \) level: Firstly, a full search of \( 0 \leq \beta_1 \leq 1, \nu = 1, 2, 3, \) with increments of 0.01, and \( \sum_{\nu=1}^{\nu=3} \beta_\nu = 1 \) is performed for each \( SNR_{in} \) level. Three figure-of-merit (FoM) constraints have to be fulfilled: \( \text{FoM}_{F} = \Delta \Psi_{log} - \hat{s}(n,k) < 1 \) and \( \text{FoM}_{MOS} = \frac{\text{PESQ}_{MOS}(WF-NEW)}{\text{PESQ}_{MOS}(MC)} \geq 1 \) and \( \text{FoM}_{SNR} = \frac{\Delta SNR_{WF-NEW}}{\Delta SNR_{MC}} \geq 1 \).

From (14) we find that all weighting vectors fulfilling the three FoM constraints have \( \beta_0 = 0 \) for all \( SNR_{in} \) levels. Please note that since \( \beta_0 = 0 \) the term \( \gamma(\ell) \cdot -1 \) disappeared, which is a nice result showing that the spatial smoothing based \( \xi_{MC}(\ell,k) \) provides a much better instantaneous a priori SNR estimate.

Secondly, to achieve an even better preservation of the speech component quality we propose the adaptation

\[
\beta_\nu(\ell,k) = \beta_\nu + (1 - \beta_\nu) \cdot (1 - q(\ell - 1,k)),
\]

with \( q(\ell - 1,k) \) being the speech absence probability from the previous frame as proposed in [14] and \( 0 \leq \beta_\nu \leq 1 \). Accordingly, we will have \( \beta_\nu(\ell,k) = 1 - \beta_\nu(\ell,k) \) along with \( \beta_\nu(\ell,k) = 0 \), which is shown as region C in Fig. 1. In the same way, a full search of \( \beta_\nu \) with a step increment of 0.01 is performed. Besides satisfying the three FoM constraints, the optimum \( \beta_\nu \) is found by maximizing the figure-of-merit FoM = \( \text{FoM}_{F} + \text{FoM}_{MOS} + \text{FoM}_{SNR} \).

As result we find \( \beta_\nu = 0.94 \), which is optimized for the most crucial level of \( SNR_{in} = -5 \text{ dB} \). The optimal adaptive weighting factor vector for (13) can now be computed according to (14) as \( \beta = (0.06 q(\ell - 1,k) \ 0 \ 0.94 + 0.06(1 - q(\ell - 1,k))) \).

### 5.2. Experimental Results

All schemes within the following evaluation are post-filters after a delay-and-sum beamformer. All approaches WF-DD, MC, and WF-NEW use the setups as described in Section 5.1, with WF-NEW employing the optimum adaptive weighting factor. Please note that different values of \( \alpha_{oe} \) and \( \xi_{min} \) aim at individually optimizing the computed approaches. The results in terms of \( \Delta SNR \) and PESQ-MOSs are depicted in Fig. 2 and the log kurtosis ratio \( \Delta \Psi_{log} \) is presented in Fig. 3. Please note that the comparison of log kurtosis ratios \( \Delta \Psi_{log} \) should be done only within one \( SNR_{in} \) level. For each curve in Fig. 2 the points from the upper left to the lower right correspond to \( SNR_{in} \) values from 15 dB to -5 dB. It can be seen that compared to the MC approach the...
WF-NEW technique shows a much better $\Delta$SNR and suppresses musical tones very efficiently especially for low $\text{SNR}_{\text{in}}$. Only PESQ-MOS$_3$ has been degraded against the MC approach, but the WF-NEW approach still maintains an acceptable score of more than 2.15 even for very low $\text{SNR}_{\text{in}}$ levels. Please note, WF-DD uses a floor of $\xi_{\text{min}} = -10$ dB, so the weighting rule is lower bounded especially for $\text{SNR}_{\text{in}} = -5$ dB. Therefore, PESQ-MOS$_3$ of SNR$_{\text{in}} = -5$ dB is not degraded against that of $\text{SNR}_{\text{in}} = 0$ dB. Although the WF-DD approach delivers the best musical tones attenuation, the WF-NEW approach has achieved overall a somewhat better performance in terms of $\Delta$SNR and particularly in terms of PESQ-MOS$_3$. In summary, the WF-NEW approach based on the newly proposed multichannel \textit{a priori} SNR estimation (13) provides a much better trade-off between noise attenuation, preservation of the speech component quality, and musical tones suppression. The results can be confirmed also by informal listening tests.

6. CONCLUSIONS

In this paper we address a new multichannel \textit{a priori} SNR estimation, which employs both spatial and temporal smoothing. The weighting factors controlling the degree of spatial and temporal smoothing are optimized using car noise training signals. Incorporated in a Wiener post-filter, we show that using this newly estimated \textit{a priori} SNR estimate, one can achieve a much better performance trade-off in terms of noise attenuation, preservation of the speech component quality, and musical tones suppression.

7. REFERENCES


