Binaural Wind-Noise Tracking with Steering Preset

Stefan Thaleiser
Ruhr-Universität Bochum,
Dept. Electrical Engineering and Information Technology
44801 Bochum, Germany
stefan.thaleiser@ruhr-uni-bochum.de

Gerald Enzner
Carl von Ossietzky University of Oldenburg
Dept. Medical Physics and Acoustics
26111 Oldenburg, Germany
gerald.enzner@uni-oldenburg.de

***Abstract*—Optimal performance of many speech enhancement methods is bound to an accurate noise power-spectral density (PSD) estimation. While for stationary noises, such as the white Gaussian or car noise, several approaches have proven themselves to perform sufficiently good, non-stationary noise types like the wind noise are more challenging. In the binaural setting and in multichannel systems, the speech-blocking method is essential to recent developments for non-stationary noise estimation. It critically requires information of the acoustic channel transfer function from source to listener. In this paper, we propose such noise-subspace approach for wind-noise PSD estimation, which relies on data-driven blind channel identification in speech presence and on *a-priori* acoustic channel information (i.e., the steering preset) in speech pause, where the smooth transition of both is controlled by *a-priori* SNR. The algorithm is designed for entire online operation based on the current noisy frame input. It improves on straightforward recursive subspace analysis and on established single-channel estimation in the wind-noise scenario, while dealing well with speech presence or babble noise too.**

***Index Terms*—binaural speech enhancement, noise estimation**

I. Introduction and Relation to Prior Art

Acoustic noise reduction is vital to hands-free and hand-held speech communication [1], [2], hearing aids [3] or true-wireless stereo headphones. Typical single-channel algorithms include the spectral-subtraction [4] and numerous minimum-mean-square error approaches mostly operated with decision-directed *a-priori* SNR [5]–[8]. From the range of multiple-input/single-output algorithms, the minimum-variance distortionless response beamformer and the multichannel Wiener filter are frequently used for speech enhancement [9]. The subclass of binaural algorithms for hearing-aids derives with a further constraint for spatial-cue preservation [10]–[14].

Many of the aforementioned filters require an estimation of the time-varying noise PSD, which is generally a challenging task, but more so in wind noise conditions as shown in this paper. In the single-channel domain, the estimators based on speech-presence probability [15] and minimum statistics [16] are landmarks. In multichannel and binaural applications, recent developments have relied on the idea of target blocking for noise estimation. To this end, [17] uses demixing based on blind source separation (BSS) to create speech and noise references. A BSS approach with directional support was proposed in [18]. Another method for target blocking relies on signal-subspace analysis [19], [20] and the formation of its orthogonal complement as the noise subspace [9], [21]. The variant for maximum noise-to-speech-ratio [22] directly optimizes for the noise reference. Signal- and noise-subspace techniques for speech blocking were compared in [14].

Advances in deep neural networks (DNNs) have also shown promising results in terms of online speech enhancement [23]. While seminal approaches still rely on an external noise reference [24], recent end-to-end DNNs obtain a target speech estimate with an internal noise reference [23]. Hardly any approaches deliver an explicit noise PSD estimate [25].

This contribution picks up the signal-subspace analysis for target blocking to create a noise reference. It reveals issues with the online processing of time-varying conditions and the scenario appears hard for fixed parameter tuning. We therefore deliver a sophisticated architecture consisting of two connected recursions for online subspace-analysis with adaptive stepsize based on SNR. Low SNR activates a steering preset for target blocking in front of the listener, while high SNR invokes an identification of the speech transfer function.

The paper is structured as follows: Sec. II declares the binaural signal model and reviews speech-blocking-based noise PSD estimation in batch and online form. Sec. III analyzes a fixed-stepsize online processing for detail, while Sec. IV deduces our evolution using SNR-based control and steering preset. Secs. IV and V demonstrate improvements for wind noise estimation, before Sec. VI concludes.

II. Binaural Signal Model and Speech-Blocking

Fig. 1 introduces our binaural signal model. Clean speech $d_i(n) = s(n) * h_{i,n}$, $i \in \{\mathrm{l}, \mathrm{r}\}$, at left and right ears results from convolution of the single-source signal $s(n)$ at sampling time n with head-related impulse responses (HRIRs) $h_{i,n}$. Noisy observations $y_i(n) = d_i(n) + n_i(n)$ are buffered and after short-time Fourier transformation (STFT) represented with a multiplicative transfer-function as $Y_\mathrm{i}(k,m) = H_\mathrm{i}(k,m)\, S(k,m) + N_\mathrm{i}(k,m)$ at discrete frequency k and frame m. Theses indices are just kept where necessary and otherwise dropped.

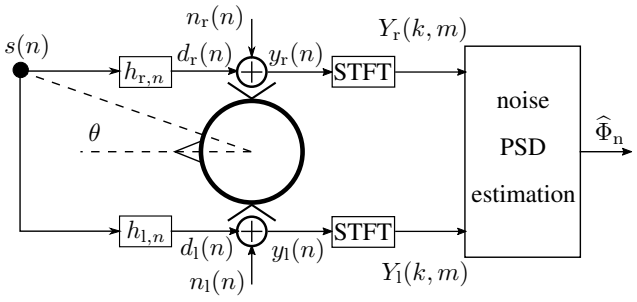


Fig. 1: Binaural signal model with noise PSD estimation.

Considering independent noise, the noisy-speech covariance matrix $\underline{\mathbf{\Phi}}_{\mathrm{yy}} = \underline{\mathbf{\Phi}}_{\mathrm{dd}} + \underline{\mathbf{\Phi}}_{\mathrm{nn}}$ composes of the covariance matrices of clean speech and noise. All covariances are defined similar to $\underline{\mathbf{\Phi}}_{\mathrm{yy}} = \mathrm{E}\{\mathbf{Y}\mathbf{Y}^{\mathrm{H}}\}$ using expectation. Based on the single-source model, the clean-speech covariance $\underline{\mathbf{\Phi}}_{\mathrm{dd}} = \Phi_{\mathrm{s}}\,\mathbf{H}\mathbf{H}^{\mathrm{H}}$ is a rank-1 matrix with the clean-speech PSD Φ_{s} of the source and the acoustic transfer functions (ATF) $\mathbf{H} = (H_{\mathrm{l}}\ H_{\mathrm{r}})^T$. In addition, in our binaural configuration with head shadowing and wind noise, we can use an assumption of uncorrelated left and right noise [26], [27] in form of a diagonal noise covariance $\underline{\mathbf{\Phi}}_{\mathrm{nn}} = \Phi_{\mathrm{n}}\,\underline{\mathbf{I}}$, where $\underline{\mathbf{I}}$ is the identity. The latter is debatable regarding the implication of equal noise PSD Φ_{n} at the microphones, which is currently somewhat necessary for our low-complexity design. The single noise PSD value per frequency is then subject to estimation from the noisy speech $\mathbf{Y} = (Y_{\mathrm{l}}\ Y_{\mathrm{r}})^T$ for application in speech enhancement.

A. Noise Estimation by Dominant-Eigenvector Decomposition

To prepare our derivation of proposed noise PSD estimators using multichannel signals $\mathbf{Y}$, we briefly recall the steps frequently taken for batch estimation of the noise PSD Φ_{n}.

A maximum-SNR problem can be defined [19], [22] as

$$\mathbf{w}_{\mathrm{MaxSNR}} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^{\mathrm{H}}\underline{\mathbf{\Phi}}_{\mathrm{dd}}\mathbf{w}}{\mathbf{w}^{\mathrm{H}}\underline{\mathbf{\Phi}}_{\mathrm{nn}}\mathbf{w}} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^{\mathrm{H}}\underline{\mathbf{\Phi}}_{\mathrm{yy}}\mathbf{w}}{\mathbf{w}^{\mathrm{H}}\underline{\mathbf{\Phi}}_{\mathrm{nn}}\mathbf{w}} \quad (1)$$

for each frequency and with complex weights $\mathbf{w}$. The solution to this cost function obviously can be restated as the dominant generalized eigenvector $\mathbf{w}_{\mathrm{MaxSNR}}$ of the matrix pair $(\underline{\mathbf{\Phi}}_{\mathrm{yy}}, \underline{\mathbf{\Phi}}_{\mathrm{nn}})$, which in turn can be resolved by standard procedures. The resulting weights can be converted [19], [20] to an estimation of normalized ATFs, i.e., $\widehat{\mathbf{H}} = \underline{\mathbf{\Phi}}_{\mathrm{nn}}\mathbf{w}_{\mathrm{MaxSNR}}$ and $||\widehat{\mathbf{H}}|| = 1$. The same end also can be reached directly by a maximum-likelihood blind channel identification [27] instead of SNR maximization and weight conversion.

Using ATFs $\mathbf{H}$, a speech-blocking matrix $\underline{\mathbf{B}}$ is frequently determined in the context of generalized sidelobe cancellation [19], [20], [28] and generally in the field of microphone array processing [9] as the orthogonal complement $\underline{\mathbf{B}} = \underline{\mathbf{I}} - \mathbf{H}(\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H$. The blocking operation then delivers a noise reference $\widetilde{\mathbf{N}} = \underline{\mathbf{B}}^H\mathbf{Y}$ and eventually a noise PSD estimate $\widehat{\Phi}_{\mathrm{n}}$ by element-wise rectification and square-averaging of $\widetilde{\mathbf{N}}$.

B. Noise PSD Tracking with Fixed-Stepsize (FS)

Our works seeks a light binaural variant of this dominant-eigenvector-based noise PSD estimation for adaptive online processing in hearing aids. We therefore accomplish a recursive estimation of the potentially time-varying noisy microphone covariance matrix, here with fixed stepsize α_{y}, as

$$\widehat{\underline{\mathbf{\Phi}}}_{\mathrm{yy}}(m) = (1-\alpha_{\mathrm{y}}) \cdot \widehat{\underline{\mathbf{\Phi}}}_{\mathrm{yy}}(m-1) + \alpha_{\mathrm{y}} \cdot \mathbf{Y}(m)\mathbf{Y}^{\mathrm{H}}(m)\,. \quad (2)$$

Temporal updates of its dominant eigenvector are then determined efficiently by the power iteration as $\widehat{\mathbf{w}}_{\mathrm{MaxSNR}}(m) = \widehat{\underline{\mathbf{\Phi}}}_{\mathrm{yy}}(m)\,\widehat{\mathbf{w}}_{\mathrm{MaxSNR}}(m-1)$ constrained to unit-norm $\widehat{\mathbf{w}}_{\mathrm{MaxSNR}}$. With $\underline{\mathbf{\Phi}}_{\mathrm{nn}} \approx \Phi_{\mathrm{n}}\,\underline{\mathbf{I}}$ according to our binaural signal model, we immediately have $\widehat{\mathbf{H}}(m) = \widehat{\mathbf{w}}_{\mathrm{MaxSNR}}(m)$.

For the binaural case at hand, instead of a blocking matrix, a vector $\mathbf{H}_{\perp}(m) = (H_{\mathrm{r}}^*\ -H_{\mathrm{l}}^*)^T$ composed from the elements of $\mathbf{H}(m)$ already meets the requirement of orthogonality with $\mathbf{H}(m)$. On this basis, we can accomplish speech blocking by means of the cross relation operation [14] for obtaining a noise reference as $\widetilde{N} = \widehat{\mathbf{H}}_{\perp}^H\mathbf{Y} = \widehat{H}_{\mathrm{r}}\,Y_{\mathrm{l}} - \widehat{H}_{\mathrm{l}}\,Y_{\mathrm{r}}$.

Eventually, the desired noise PSD estimation can be computed online from the squared rectified noise reference as

$$\widehat{\Phi}_{\mathrm{n}}(m) = (1-\alpha_{\mathrm{CR}}) \cdot \widehat{\Phi}_{\mathrm{n}}(m-1) + \alpha_{\mathrm{CR}} \cdot |\widetilde{N}(m)|^2 \quad (3)$$

with stepsize α_{CR}. In conclusion, this online algorithm termed fixed-stepsize (FS) adaptive estimator is characterized by two hyperparameters, α_{y} and α_{CR}, to control its adaptivity.

TABLE I: Development and test-phase data composition.

phase	target signal	noise type	SNR	data status
dev-1	-	wind	$-\infty$	development
dev-2	speech + COD	sensor	30 dB	development
test-1	speech	wind	5 dB	test
test-2	speech	babble	5 dB	test

III. Analysis of the "FS" Noise-PSD Estimation

For analysis we construct an acoustic scenario with various challenges. A sequence of cases comprising two development and two test phases is illustrated in Fig. 2 and described in Table I. Speech is used from the TIMIT database [29] and wind noise is generated according to [26]. A continuous sensor noise at about 30 dB below the speech level is further added in all phases including initialization. For reference, the time-varying clean speech PSD $\Phi_{\mathrm{d_r}}(m)$ of the right ear and the oracle total-noise PSD $\Phi_{\mathrm{n}}(m)$ are assessed before the mixing by recursive averaging and shown in the diagram. Additionally, a challenging sudden change-of-direction (COD) of the spatial location of the speech signal from $\theta_1 = -45°$ (front left of the listener) to $\theta_2 = 45°$ (front right) is embedded into the dev-2 phase and marked by the dashed vertical line.

Fig. 2 then demonstrates issues with the FS-based noise PSD estimation $\widehat{\Phi}_{\mathrm{n}}(m)$ for choices of "slow" $\alpha_{\mathrm{y},1} = 0.02$ and "fast" $\alpha_{\mathrm{y},2} = 0.3$, while $\alpha_{\mathrm{CR}} = 10^{-1.5}$ in both cases. All PSDs (oracle and estimated) are depicted per frame index and averaged along the entire frequency range. We observe

- noise PSD underestimation in noisy dev-1, test-1, test-2,
- with the underestimation more pronounced for the faster $\alpha_{\mathrm{y},2}$ due to short-time noise correlation in $\widehat{\underline{\mathbf{\Phi}}}_{\mathrm{yy}}(m)$,
- noise-estimation hangover (thus noise overestimation) at the beginning of the high-SNR speech phase dev-2,
- overshooting noise estimation at the COD with the slow $\alpha_{\mathrm{y},1}$, indicating strong speech leakage into $\widehat{\Phi}_{\mathrm{n}}(m)$.

Any speech enhancement on the basis of the misestimated noise PSD would clearly turn out very unreliable regarding the preservation of desired speech and the suppression of undesired noise. Noise overestimation through dev-2 would result in speech attenuation, while noise underestimation in dev-1 will turn out as unsatisfactory noise filtering.

To improve, we first restrict ourselves to the dev-1 noise phase and systematically evaluate the log-spectral error $\mathrm{LSE} = 10\log_{10}(\Phi_{\mathrm{n}}/\widehat{\Phi}_{\mathrm{n}})$ between oracle and estimated noise PSD (here assessed with equal stepsizes $\alpha_{\mathrm{or}} = \alpha_{\mathrm{CR}}$ for comparability). The mean and standard deviation of the LSE across

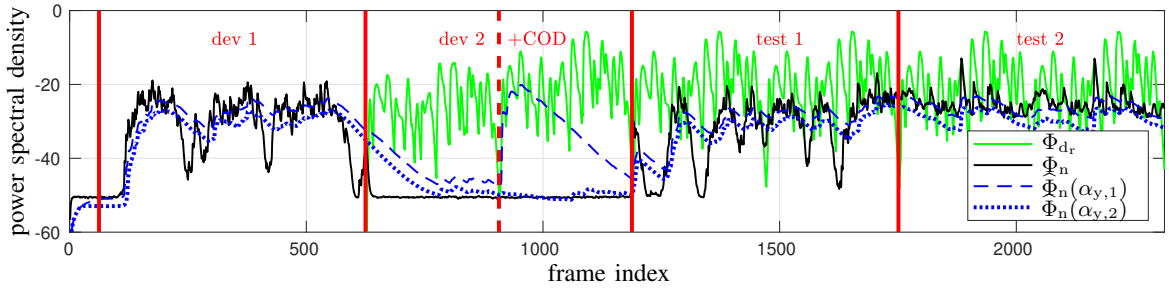


Fig. 2: Issues with noise PSD estimation by the "FS" (fixed-stepsize) online estimator, $\alpha_{\mathrm{CR}} = 10^{-1.5}$.

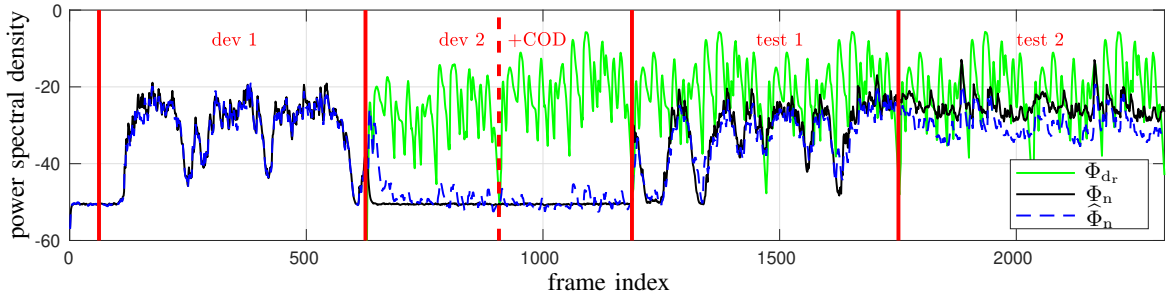


Fig. 3: Improved noise PSD estimation by the "AS" (adaptive-stepsize) online estimator, $\alpha_{\mathrm{y}} = 0.3$.

frames and frequencies for 100 wind-noise realizations are listed in Tabs. II and III, respectively. The stepsize α_{y} of Eq. (2) is varied linearly, whereas α_{CR} of Eq. (3) is varied logarithmically to accommodate their respective effects on the processing. It is apparent from Tab. II that there is a dependency primarily of the bias (mean error) of the noise PSD with α_{y} and not so much with α_{CR}. Conversely, it turns out from Tab. III that the standard deviation of the noise estimation is mainly steered by α_{CR} and hardly by α_{y}.

As a result, we might optimize the stepsizes α_{y} and α_{CR} separately for low bias and low standard deviation of the noise estimator. However, it will be difficult to accommodate the conflicting requirements towards the speech phase dev-2 with any choice of the fixed stepsize. While slower stepsize α_{CR} would reduce standard deviation in dev-1, the hangover time after dominant noise in dev-1 will further extend into dev-2 of Fig. 2. Slower stepsize α_{y} would reduce estimation bias in dev-1, but trigger yet harsher speech leakage in the case of rapidly changing ATF due to COD in the middle of dev-2. For overall more satisfactory operation, we must therefore pursue a dynamical solution for controlling stepsize.

TABLE II: Bias of LSE of $\widehat{\Phi}_{\mathrm{n}}$ estimate by FS.

$\alpha_{\mathrm{y}} \setminus \alpha_{\mathrm{CR}}$	$10^{-0.5}$	10^{-1}	$10^{-1.5}$	10^{-2}	$10^{-2.5}$
0.1	1.1 dB	1.0 dB	0.9 dB	0.8 dB	0.8 dB
0.2	2.0 dB	1.9 dB	1.8 dB	1.7 dB	1.6 dB
0.3	2.9 dB	2.7 dB	2.6 dB	2.5 dB	2.4 dB

TABLE III: Standard deviation of LSE of $\widehat{\Phi}_{\mathrm{n}}$ estimate by FS.

$\alpha_{\mathrm{y}} \setminus \alpha_{\mathrm{CR}}$	$10^{-0.5}$	10^{-1}	$10^{-1.5}$	10^{-2}	$10^{-2.5}$
0.1	1.8 dB	1.1 dB	0.7 dB	0.6 dB	0.6 dB
0.2	2.0 dB	1.1 dB	0.7 dB	0.6 dB	0.6 dB
0.3	2.2 dB	1.2 dB	0.8 dB	0.7 dB	0.7 dB

IV. Proposed Adaptive Stepsize "AS" Tracking

Based on the observed challenges with fixed-stepsize online subspace tracking in dynamical ATF conditions and with speech presence/absence or time-varying SNR, we propose an evolution of the noise PSD estimation (with comparable complexity as FS) in which the subspace tracking pertains to further regularizations according to SNR. Specifically,

- we take example of a filtered subspace analysis in [30] to substitute our fixed-stepsize tracking with an adaptive *time-varying stepsize* computation according to SNR
- and we further embed regularization in form of a *steering preset* in low SNR when the data-driven subspace analysis otherwise tends to random or leans on undesired short-time correlation and thus creates estimation variance.

The former intends fast and continual tracking of time-varying ATFs in situations with high SNR (speech presence). The latter is meant to provide for a non-adaptive steering vector $\mathbf{H}_{\mathrm{pre}}$ in cases of very low SNR (speech absence). It refers to any fixed steering, for instance, toward the front of the listener and substitutes the critical data-driven subspace analysis. Smooth transition of both is to be controlled adaptively with SNR.

A. Definition of the Algorithm

Those ambitions regarding improved noise-estimation accuracy can be met by a two-stage system in Fig. 4. A preliminary noise estimate $\widehat{\Phi}'_{\mathrm{n}}$ is here obtained by fixed-stepsize subspace tracking with an FS estimator or might be obtained by another noise estimator suitable for time-varying wind noise conditions at hand. The preliminary noise estimate is then plugged into a spectral-subtraction form of a Wiener filter, i.e.,

$$G = \max\left(\frac{\widehat{\Phi}_{\mathrm{y_l y_l}} + \widehat{\Phi}_{\mathrm{y_r y_r}} - 2 \cdot \xi \cdot \widehat{\Phi}'_{\mathrm{n}}}{\widehat{\Phi}_{\mathrm{y_l y_l}} + \widehat{\Phi}_{\mathrm{y_r y_r}}}, 0\right), \qquad (4)$$

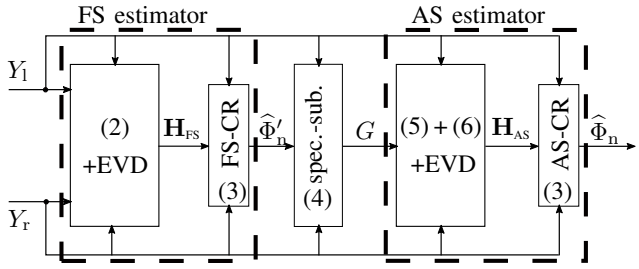


Fig. 4: SNR-adaptive noise PSD estimation architecture.

where the two noisy speech PSDs $\widehat{\Phi}_{\mathrm{y}_i\mathrm{y}_i}$, $i \in \{l, r\}$, can be used from the main diagonal of (2) and computed with equal stepsizes $\alpha_\mathrm{y} = \alpha_\mathrm{CR}$ for successful noise subtraction in the numerator of (4). For the entire system, we can report somewhat better utility of slow adaptation of the preliminary noise PSD and the Wiener filter (e.g., $\alpha_\mathrm{CR} = 10^{-2}$) in order to impose stable *a-priori* SNR conditions for the second, adaptive-stepsize estimator to follow. A noise-overestimation factor ξ (selected according to Tabs. II and III to compensate for estimation bias and variance on the preliminary noise estimate) further enhances the Wiener filter computation.

The spectral filter G then directly supports a second noise PSD estimator to characterize it by an adaptive stepsize for fast and robust tracking. As a further resource for the second stage we prepare a definition of our steering preset as $\mathbf{H}_\mathrm{pre} = (1 \quad 1)^T$ in front of the listener and bring it into covariance form as $\underline{\mathbf{\Phi}}_{\mathrm{yy,pre}} = \mathbf{H}_\mathrm{pre}\mathbf{H}_\mathrm{pre}^H$. On this basis, we can define our adaptive-stepsize (AS) noise estimation module in Fig. 4 by a set of two simple connected recursions, firstly

$$\underline{\widehat{\mathbf{\Phi}}}'_{\mathrm{yy,A}}(m) = (1-\alpha_\mathrm{A})\,\underline{\widehat{\mathbf{\Phi}}}_{\mathrm{yy,A}}(m-1) + \alpha_\mathrm{A}\frac{\mathbf{Y}(m)\mathbf{Y}^\mathrm{H}(m)}{\sigma_y^2} \quad (5)$$

with the adaptive stepsize $\alpha_\mathrm{A} = \alpha_\mathrm{y}\,G$ to activate the current input data $\mathbf{Y}$ in case of higher SNR and to forget the memory term $\underline{\widehat{\mathbf{\Phi}}}_{\mathrm{yy,A}}(m-1)$. This intermediate result $\underline{\widehat{\mathbf{\Phi}}}'_{\mathrm{yy,A}}(m)$ is then handed over to the second recursion

$$\underline{\widehat{\mathbf{\Phi}}}_{\mathrm{yy,A}}(m) = (1-\alpha_\mathrm{pre})\,\underline{\widehat{\mathbf{\Phi}}}'_{\mathrm{yy,A}}(m) + \alpha_\mathrm{pre}\,\underline{\mathbf{\Phi}}_{\mathrm{yy,pre}}\,, \quad (6)$$

which introduces the steering preset as a regularization to the data-driven computation of (5). The stepsize $\alpha_\mathrm{pre} = \alpha_\mathrm{y}\,(1-G)$ is complementary with α_A. Note that the normalization σ_y^2 in (5) is responsible to balance the weight of the input signals with the steering preset for any scaling of the input.

Assuming high SNR, then, specifically, the second recursion will discard the steering preset and thus connect the intermediate output of the first recursion to the final output $\underline{\widehat{\mathbf{\Phi}}}_{\mathrm{yy,A}}(m)$. In this case, with typically a large α_y configured as the maximum stepsize of the entire process, the two connected recursions deliver a rapid update of the data covariance $\underline{\widehat{\mathbf{\Phi}}}_{\mathrm{yy,A}}$ for time-varying ATF conditions.

Assuming low SNR, the first recursion will obviously discard the noisy data $\mathbf{Y}$ and instead pass on the memory term $\underline{\widehat{\mathbf{\Phi}}}_{\mathrm{yy,A}}(m-1)$ to the intermediate output. The second recursion with complementary stepsize will now mix the steering preset into the final output and over time forget the intermediate data-driven input $\underline{\widehat{\mathbf{\Phi}}}'_{\mathrm{yy,A}}(m)$. The steering preset therefore dominates the final output in low SNR conditions as desired. In mixed conditions with mid-range SNR, the two connected recursion will seek a tradeoff between adaptivity to the data and regularization in form the steering preset.

Conventional power iteration for EVD is then performed on the final output $\underline{\widehat{\mathbf{\Phi}}}_{\mathrm{yy,A}}$ and the resulting ATF estimation $\widehat{\mathbf{H}}_{\mathrm{AS}}(m)$ is handed over to cross relation [14] and Eq. (3) for obtaining a noise reference and a noise PSD (cf. Sec. II).

B. Illustration of the Adaptive Mechanism

We return to the acoustic configuration of Fig. 2 for first judgement regarding the desired improvement. In line with arguments throughout the algorithm definition, we choose the fixed parameters $\alpha_\mathrm{y} = 0.3$ and $\alpha_\mathrm{CR} = 10^{-2}$ in the fixed-stepsize, and $\alpha_\mathrm{y} = \alpha_\mathrm{CR} = 0.3$ in the adaptive-stepsize part of the algorithm. The filter G inbetween as shown in (4) relies on squareroot-Hann windowed FFT processing of length 512 with $50\,\%$ overlap and bias correction $\xi = 2$ and is finally averaged across frequency. Fig. 3 then depicts clearly improved noise estimation in dev-1 and test-1 phases, i.e., the estimated noise PSD coherently following the dynamical evolution of the true noise PSD. The dev-2 phase considerably improves on the former hangover from dev-1 and on the overshooting noise PSD with COD in Fig. 2. The test-2 phase, expectedly, shows a noise underestimation due to the inherent coherence of left and right binaural babble noise signals, which would require a framework with additional noise covariance modeling and computationally more demanding generalized EVD.

V. Expanded Evaluation and Comparison

We rely once more on the acoustic configuration of Figs. 2 and 3 to evaluate across various global SNR (averaged left and right channel SNR) and to compare these contenders:

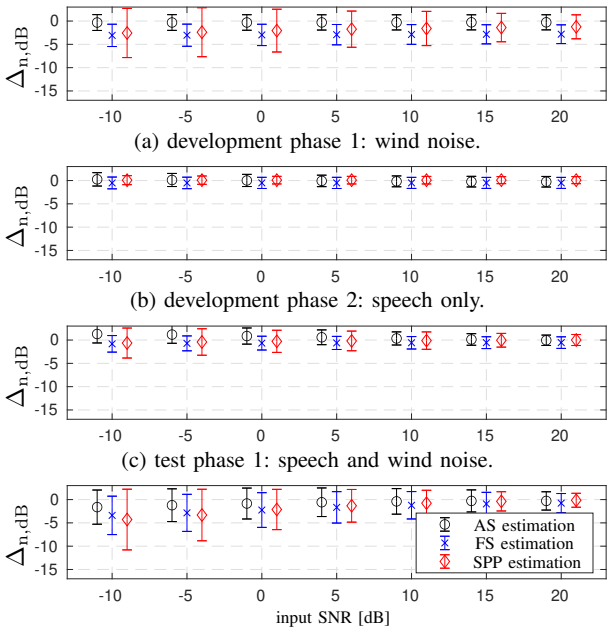
- an $\alpha_\mathrm{y} = 0.3$, $\alpha_\mathrm{CR} = 0.3$ fast fixed-stepsize (FS) recursion
- the proposed adaptive-stepsize (AS) algorithm
- the state-of-the-art 1-channel SPP-based estimator [15].

As a metric for comparison of the noise PSD estimation, we here suggest a speech-conditioned log-spectral error

$$\Delta_\mathrm{n,dB} = \frac{1}{\kappa\,M}\sum_{m=1}^{M}\sum_{k=1}^{\kappa} 10\log_{10}\left(\frac{\widehat{\Phi}_\mathrm{n}(k,m)+\bar{\Phi}_\mathrm{d}(k,m)}{\Phi_\mathrm{n}(k,m)+\bar{\Phi}_\mathrm{d}(k,m)}\right) \quad (7)$$

with the speech PSD $\bar{\Phi}_\mathrm{d} = (\Phi_{\mathrm{d_l}}+\Phi_{\mathrm{d_r}})/2$ to put the noise PSDs appropriately into perspective, for instance, to invalidate small noise estimation errors in the presence of dominant speech. Averaging takes place across all frequencies k, frames m, and 10 realizations of each signal phase. Fig. 5 shows the resulting mean errors (markers) and standard deviations (error bars) for dev-1/2 and test-1/2 phases in the subfigures.

It turns out in the wind-noise dev-1 phase that the 1-channel SPP estimator exhibits large underestimation bias, since its structure of several recursions is not well suited for harsh non-stationarity. The FS noise-subspace algorithm obviously reduces the variance of the estimation at the cost of additional bias. The proposed AS algorithm, in turn, manages to strongly enhance the estimation variance and almost eliminate the bias.



(a) development phase 1: wind noise.

(b) development phase 2: speech only.

(c) test phase 1: speech and wind noise.

(d) test phase 2: speech and babble noise.

Fig. 5: Evaluation of noise PSD estimation related to Fig. 3.

The same trends translate to the mixed speech-and-noise test-1 phase, where the absolute range of bias and variance is compressed due to speech conditioning of the spectral error metric. In the dev-2 speech phase, estimation bias and variance of all contenders is minor under the considered metric. The babble noise estimation in the test-2 phase shows only slight improvements with AS, however, appears at least robust in our context of optimized wind noise estimation.

VI. Conclusion

Our study was concerned with online noise PSD estimation for binaural speech enhancement in time-varying conditions. Straightforward translation of a common subspace technique to a fixed-stepsize online tracking has revealed considerable issues of bias and variance. We have thus proposed an adaptive-stepsize estimator with SNR-based control and regularization with steering preset. The evaluation proves enhanced performance with moving sources and non-stationary wind noise.

References

[1] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, Wiley, 2004.

[2] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, Wiley, 2006.

[3] J. S. Kates, *Digital Hearing Aids*, Plural Publishing, 2008.

[4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[5] 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, 1984.

[6] P. Wolfe and S. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP J. Appl. Signal Process.*, vol. 910167, no. 10, pp. 1043–1051, 2003.

[7] R. Martin, "Speech enhancement based on minimum mean-square error estimation and super-Gaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[8] J. Erkelens, R. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with genrealized Gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, 2007.

[9] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer, 2008.

[10] B. Cornelis, S. Doclo, T. Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Tr. Audio, Speech, Lang. Proc.*, vol. 18, no. 2, pp. 342–355, 2010.

[11] D. Marquardt and S. Doclo, "Interaural coherence preservation for binaural noise reduction using partial noise estimation and spectral postfiltering," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 26, no. 7, pp. 1261–1274, April 2018.

[12] A. Kamkar-Parsi and M. Bouchard, "Instantaneous binaural target PSD estimation for hearing aid noise reduction in complex acoustic environments," *IEEE Trans. Instrumentation Meas.*, vol. 60, no. 4, pp. 1141–1154, 2011.

[13] M. Jeub, M. Schäfer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, pp. 1732–1745, 2010.

[14] M. Azarpour and G. Enzner, "Binaural noise reduction via cue-preserving MMSE filter and adaptive-blocking-based noise PSD estimation," *EURASIP Jrnl. Adv. Signal Process.*, vol. 2017, no. 49, 2017.

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

[16] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Process.*, vol. 9, pp. 504–512, July 2001.

[17] L. Wang, T. Gerkmann, and S. Doclo, "Noise PSD estimation using blind source seperation in a diffuse noise field," *Int. Workshop on Acoustic Signal Enhancement 2012*, 2012.

[18] K. Reindl, Y. Zheng, A. Schwarz, S. Meier, R. Maas, A. Sehr, and W. Kellermann, "A stereophonic acoustic signal extraction scheme for noisy and reverberant environments," *Computer Speech and Lang.*, vol. 2013, no. 27, pp. 726–745, July 2013.

[19] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Speech, Audio Process.*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[20] A. Krueger, E. Warsitz, and R. Heab-Umbach, "Speech enhancement with a GSC-like structure employing eigevector-based transfer function ratios estimation," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 19, no. 1, pp. 206–218, Jan. 2011.

[21] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE Trans. Speech, Audio Process.*, vol. 26, no. 6, pp. 1056–1071, June 2018.

[22] L. Wang, T. Gerkmann, and S. Doclo, "Noise power spectral density estimation using MaxNSR blocking matrix," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 23, no. 9, pp. 1493–1508, Sep. 2015.

[23] S. Braun, H. Gamper, C. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, pp. 656–660, May 2021.

[24] Y. Xu, J. Du, L. Dai, and C.H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[25] A. Chinaev, J. Heymann, L. Drude, and R. Haeb-Umbach, "Noise-presence-probability-based noise PSD estimation by using DNNs," *ITG Conf. Speech Communication*, Oct. 2016.

[26] C. Nelke and P. Vary, "Measurement, analysis and simulation of wind noise signal for mobile communication devices," *Int. Workshop on Acoustic Signal Enhancement*, 2014.

[27] P. Thüne and G. Enzner, "Maximum-likelihood approach to adaptive multichannel-Wiener postfiltering for wind-noise reduction," *ITG Conf. Speech Communications*, Oct. 2016.

[28] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Speech, Audio Process.*, vol. 17, no. 7, pp. 1420 – 1434, Sep. 2009.

[29] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Comm., Elsevier*, vol. 9, no. 4, pp. 351–356, Aug. 1990.

[30] P. Thüne and G. Enzner, "Maximum-likelihood and maximum-a-posteriori perspectives for blind channel identification on acoustic sensor network data," *ITG Conf. Speech Communication*, Oct. 2018.