ANALYSIS OF BEAMFORMER DIRECTED SINGLE-CHANNEL NOISE REDUCTION SYSTEM FOR HEARING AID APPLICATIONS

Jesper Jensen[*†] *and Michael Syskind Pedersen*[*]

[*] Oticon A/S, 2765 Smørum, Denmark
[†]Aalborg University, 9220 Aalborg Ø, Denmark

ABSTRACT

We study multi-microphone noise reduction systems consisting of a beamformer and a single-channel (SC) noise reduction stage. In particular, we present and analyse a maximum likelihood (ML) method for jointly estimating the target and noise power spectral densities (psd's) entering the SC filter. We show that the estimators are minimum variance and unbiased, and provide closed-form expressions for their mean-square error (MSE). Furthermore, we show that the MSE of the noise psd estimator is particularly simple: it is independent of target signal characteristics, frequency, and microphone locations. In a hearing aid context, we analyze the performance of the estimators as a function of target angle-of-arrival and frequency. Finally, we demonstrate the advantage of the proposed method in a hearing aid situation with a target speaker in large-crowd noise.

Index Terms— Power spectral density estimation, multi-microphone speech enhancement, noise reduction for hearing aids.

1. INTRODUCTION

Acoustic communication devices, e.g., tele-conferencing systems, mobile phones, portable/wearable hearing devices, etc., must work well in noisy environments. When power, space, and cost constraints allow it, multi-microphone noise reduction systems are useful in this situation, since they offer spatial filtering, in addition to the spectro-temporal filtering allowed by single-microphone systems.

The Multi-channel Wiener Filter (MWF) and its recent extensions is an efficient tool for multi-microphone speech enhancement, e.g. [1, 2]. It is well-known that the MWF can be decomposed into a concatenation of a Minimum Variance Distortion-less Response (MVDR) beamformer and a single-channel Wiener filter (SWF) [3]. In practice, this decomposition is often prefered as it offers implementational advantages and easier control of spatial and spectro-temporal processing artefacts. With this MVDR+SWF representation, the SWF relies on estimates of the psd's of the target and noise signals, respectively, entering the SWF. These psd's may be estimated directly from the output signal of the beamformer [4] using single-channel noise psd tracking algorithms [5–8]. Other approaches use multiple microphone signals to first estimate the inter-microphone noise covariance matrix based on various assumptions wrt. the structure of this matrix, e.g. [9–12]. Afterwards, since the beamformer is linear, it is easy to estimate the noise psd entering the SWF, e.g., [13]. Both these classes of methods focus on the noise psd; an estimate of the target speech psd is found subsequently using the assumption that target and noise processes are independent.

In this paper, we present and analyze a multi-microphone maximum likelihood (ML) method for *jointly* estimating the target and noise psd entering the SWF. The method is general and our theoretical analysis, which we present in the first part of the paper, is valid for any microphone array configuration. In the last half of the paper, our focus is to understand the behavior of the method as a function of target signal direction and frequency, when used for speech enhancement in a behind-the-ear (BTE) hearing aid (HA) application.

While expressions for the ML estimators themselves have already been derived for direction-of-arrival problems, e.g. [14, 15], our analysis and application of them in a microphone array context is new. The contributions of the paper can be summarized as follows. First, we provide an interpretation of the proposed ML approach in terms of a recently introduced method for noise psd tracking [13] - this interpretation is important because a) it explains the proposed ML approach in terms of simple filtering operations, and b) the theoretical results presented in this paper are directly applicable to [13] and other methods based on it, e.g., [16, 17]. Secondly, we present and analyze closed-form expressions for the mean-square error (MSE) achievable with the proposed psd estimators. Finally, we focus on the BTE-HA, where we provide a simulation study of the MSE performance and demonstrate the advantage of the proposed estimators in a multi-microphone speech enhancement system.

2. SIGNAL MODEL AND ASSUMPTIONS

Let the noisy observation at the mth microphone be given by

$$y_m(n) = x_m(n) + v_m(n), \qquad m = 1, \ldots, M,$$

where $y_m(n)$, $x_m(n)$, and $v_m(n)$ denote the noisy, clean target, and noise signal, respectively, $M > 1$ is the number of microphones, and n is a discrete-time index. We assume that the signals are realizations of zero-mean Gaussian random processes, and that noise and target processes are uncorrelated. A short-time Fourier transform (STFT) applied to each microphone signal leads to STFT coefficients

$$Y_m(l,k) = \sum_{n=0}^{N-1} y_m(n + lD_A) w_A(n) e^{-\frac{2\pi jkn}{N}},$$

where l and k are frame and frequency bin indices, respectively, N is the frame length, D_A is the decimation factor, $w_A(n)$ is the analysis window function, and $j = \sqrt{-1}$. Similar expressions hold for target STFT coefficients $X_m(l,k)$ and noise STFT coefficients $V_m(l,k)$.

We assume that $Y_m(l,k)$ are approximately independent across time l and frequency k[1] , which allows us to treat STFT coefficients with different frequency index k independently, and suppress index k in the notation. For a given frame index l, we can collect the noisy STFT coefficients for each microphone in a vector,

$$Y(l) \triangleq [Y_1(l) \ldots Y_M(l)]^T.$$

[1]This is a standard assumption in speech processing, which is valid when the correlation time of the signal is short compared to the frame length N, and successive frames are spaced sufficiently far apart [5, 18].

Analogous expressions hold for $X(l)$ and $Y(l)$, so that

$$Y(l) = X(l) + V(l).$$

Let $d'(l) = [d'_1(l) \cdots d'_M(l)]^T$ denote the acoustic transfer function from target source to each microphone, let $d(l) = [d'_1(l)/d'_i(l) \cdots d'_M(l)/d'_i(l)]^T$ be the relative acoustic transfer function wrt. the ith (reference) microphone [19], and let $\bar{X}(l)$ be the target STFT coefficient at this microphone. Then,

$$X(l) = \bar{X}(l)d(l).$$

Finally, the noise covariance matrix $C_V(l) = E(V(l)V^H(l))$, which is assumed invertible, can evolve across signal regions with speech activity according to the model introduced in [13],

$$C_V(l) = \lambda_V(l)C_V(l_0), \qquad l > l_0,$$

where l_0 denotes the most recent frame index where speech was absent. For convenience, we scale $C_V(l_0)$ such that element (i,i) equals one, because then $\lambda_V(l)$ is the time-varying psd of the noise process, measured at the reference microphone. Thus, during speech presence, the noise process does not need to be stationary, but the covariance structure is assumed fixed up to a scalar multiplication.

Hence, the covariance matrix of the noisy observation during speech activity can be summarized as

$$C_Y(l) = \lambda_X(l)d(l)d^H(l) + \lambda_V(l)C_V(l_0), \qquad l > l_0, \quad (1)$$

where $d(l)$ is assumed known and $C_V(l_0)$ can be estimated in speech absence signal regions, but the time-varying speech and noise psds, $\lambda_X(l)$ and $\lambda_V(l)$, are unknown and must be estimated.

3. ML ESTIMATION OF TARGET AND NOISE PSDS

We consider ML estimation of $\lambda_X(l)$ and $\lambda_V(l)$ during signal regions with speech activity. From the assumptions above it follows that vector $Y(l)$ obeys a zero-mean (complex, circular symmetric) Gaussian probability density function (pdf), i.e.,

$$f_{Y(l)}(Y(l); \lambda_X(l), \lambda_V(l)) = \mathcal{N}\left(0, C_Y(l)\right).$$

Let $\underline{Y}_D(l) = [Y(l-D+1) \cdots Y(l)]$ denote a sequence of D successive observations. Then, since observations $Y_m(l)$ are assumed independent across time l, the pdf of $\underline{Y}_D(l) = [Y(l-D+1) \cdots Y(l)]$ is given by

$$f_{\underline{Y}_{D}(l)}(\underline{Y}_D(l); \lambda_X, \lambda_V) = \prod_{j=l-D+1}^{l} f_{Y(j)}(Y(j); \lambda_X, \lambda_V), \quad (2)$$

under the short-time stationarity assumption $\lambda_V \triangleq \lambda_V(j)$, $\lambda_X \triangleq \lambda_X(j)$, and $d = d(j)$, $j = l-D+1, \ldots, l$. ML estimates $\lambda_{X,ML}(l)$ and $\lambda_{V,ML}(l)$ of $\lambda_X(l)$ and $\lambda_V(l)$ can be found via partial derivatives of Eq. (2) with respect to $\lambda_X(l)$ and $\lambda_V(l)$. As shown in [14], the ML estimate of $\lambda_V(l)$ is given by

$$\lambda_{V,ML}(l) = \frac{1}{M-1}\text{tr}\left(Q_u(l)\hat{C}_Y(l)C_V^{-1}(l_0)\right), \quad (3)$$

where

$$\hat{C}_Y(l) = \frac{1}{D}\sum_{j=l-D+1}^{l} Y(j)Y(j)^H,$$

is the sample covariance matrix of the noisy signal, and

$$Q_u(l) = I - d(l)(d^H(l)C_V^{-1}(l_0)d(l))^{-1}d^H(l)C_V^{-1}(l_0).$$

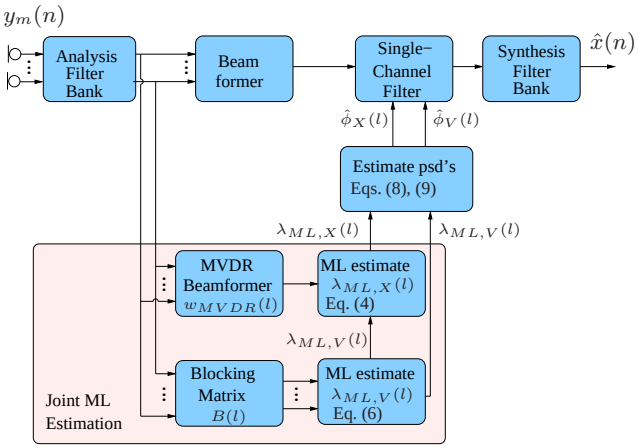

Fig. 1. The proposed ML estimation framework used in a beamformer-directed single-channel noise reduction system.

Furthermore, re-writing the results of [14] slightly, we find

$$\lambda_{X,ML}(l) = w_{MVDR}^H(l)\left(\hat{C}_Y(l) - \lambda_{V,ML}(l)C_V(l_0)\right) \times w_{MVDR}(l), \quad (4)$$

where

$$w_{MVDR}(l) = \frac{C_V^{-1}(l_0)d(l)}{d^H(l)C_V^{-1}(l_0)d(l)} \quad (5)$$

is the weight vector of an MVDR beamformer, e.g., [3].

3.1. Interpretation

The ML estimators in Eqs. (3) and (4) have interesting interpretations. It can be shown that (3) is identical to the ML estimator proposed in [13], where it was interpreted in terms of a blocking matrix $B(l)$; this interpretation may therefore be re-used here. Specifically, let $B(l) \in \mathcal{C}^{M \times M-1}$ denote a blocking matrix whose columns form a basis for the $M-1$ dimensional vector space orthogonal to $d(l)$, so that $d^H(l)B(l) = 0$. Then it may be verified that Eq. (3) can be re-written as the following expression derived in [13],

$$\lambda_{V,ML}(l) = \frac{1}{M-1}\text{tr}\big(\frac{1}{D}\underline{Y}_D^H(l)B(l) \times (B^H(l)C_V(l_0)B(l))^{-1}B^H(l)\underline{Y}_D(l)\big). \quad (6)$$

Eq. (6) may be interpreted as the average variance of the observable noisy vector $Y(l)$, passed through $M-1$ linearly independent target cancelling beamformers, and normalized according to the noise covariance between the outputs of each beamformer.

Furthermore, the ML estimate $\lambda_{X,ML}(l)$, Eq. (4), is simply the variance of the noisy observation $Y(l)$ minus the estimated noise variance, at the output of an MVDR beamformer. Fig. 1 uses this interpretation in a noise reduction system consisting of a beamformer, followed by an SC filter. Here, beamformers $w_{MVDR}(l)$ and $B(l)$ facilitate estimation of $\lambda_X(l)$ and $\lambda_V(l)$, which are used to inform the SC filter about the target and noise psds in its input signal.

4. ANALYSIS OF ESTIMATORS

4.1. Minimum Variance Unbiased (MVU) Estimators

It can be shown that the ML estimators, $\lambda_{S,ML}(l)$ and $\lambda_{V,ML}(l)$, are *minimum variance unbiased* (MVU) estimators (we have omitted

the proof due to space limitations: the unbiasedness property follows easily by computing $E(\lambda_{V,ML}(l))$ and $E(\lambda_{X,ML}(l))$, respectively, while the minimum variance property follows from application of [20, Thm. 3.2]. In other words, no unbiased estimators exist, which achieve an MSE lower than that of $\lambda_{V,ML}(l)$ and $\lambda_{X,ML}(l)$.

4.2. Cramér-Rao Lower Bounds (CRLBs)

The CRLB is a lower bound on the estimation MSE for unbiased estimators. Generally, ML estimators approach the CRLB *asymptotically* in the number of observations D. But when the estimators are MVU, as the proposed ones, they attain the bound not only asymptotically, but for *any* data record length D. This property is important for speech processing, because, due to non-stationarity of the speech or noise signal, D must often be small.

We now derive expressions for the CRLB. To do so, let $\theta = [\lambda_X(l) \;\; \lambda_V(l)]^T$, $\theta_{ML} = [\lambda_{ML,X}(l) \;\; \lambda_{ML,V}(l)]^T$, and let $I(\theta)$ denote the Fisher information matrix (FIM) with elements

$$[I(\theta)]_{ij} = -E\left[\frac{\partial^2 \log f_{\underline{Y}_D(l)}(\underline{Y}_D(l);\theta)}{\partial\theta_i\theta_j}\right]. \qquad (7)$$

Then, because our estimators are MVU, the estimation MSE of the ith element in θ satisfies [20]

$$E[(\theta_{ML,i} - \theta_i)^2] = [I(\theta)^{-1}]_{ii},$$

i.e., the MSE is given by the diagonal elements of the inverse FIM. Evaluating the partial derivaties in Eq. (7), it may be shown that [21]

$$[I(\theta)]_{ij} = D\text{tr}\left[C_Y^{-1}(l;\theta)\frac{\partial C_Y(l;\theta)}{\partial\theta_i}C_Y^{-1}(l;\theta)\frac{\partial C_Y(l;\theta)}{\partial\theta_j}\right],$$

where $C_Y(l;\theta)$ is given by Eq. (1) and we emphasized its dependence on $\theta = [\lambda_X(l) \;\; \lambda_V(l)]^T$. Computing the derivatives $C_Y(l;\theta)/\partial\theta_i$, and applying the matrix-inversion lemma to find expressions for $C_Y^{-1}(l;\theta)$ allows us, after some simplifications, to find expressions for $[I(\theta)^{-1}]_{ii}$.

To report the results compactly, assume that an MVDR beamformer is used in the top brach of Fig. 1. Let $\phi_X(l)$ and $\phi_V(l)$ denote the psd of the target and noise components, respectively, at the output of this MVDR filter,

$$\phi_X(l) = \lambda_X(l), \qquad (8)$$

and

$$\phi_V(l) = \frac{\lambda_V(l)}{d^H C_V^{-1}(l_0)d}. \qquad (9)$$

Then, the signal-to-noise ratio (SNR) $\xi(l)$ at the output of the MVDR filter is given by

$$\xi(l) = \frac{\phi_X(l)}{\phi_V(l)} = \frac{\lambda_X(l)}{\lambda_V(l)}d^H C_V^{-1}(l_0)d. \qquad (10)$$

Rather than the absolute CRLB, it is more informative to report the CRLB relative to the squared quantity of interest, i.e., N-CRLB$_{\lambda_X} = E[(\lambda_{X,ML} - \lambda_X)^2]/\lambda_X^2$ and N-CRLB$_{\lambda_V} = E[(\lambda_{V,ML} - \lambda_V)^2]/\lambda_V^2$, respectively. From Eqs. (7)-(10) we find that

$$\text{N-CRLB}_{\lambda_X} = \frac{1}{D}\left(\frac{1+\xi(l)}{\xi(l)}\right)^2 + \frac{1}{D}\frac{1}{M-1}\frac{1}{\xi^2(l)}, \qquad (11)$$

$$\text{N-CRLB}_{\lambda_V} = \frac{1}{D}\frac{1}{M-1}. \qquad (12)$$

From these expressions, the following observations can be made.

1. N-CRLB$_{\lambda_X}$ and N-CRLB$_{\lambda_V}$ are monotonically decreasing in the number of microphones M and observations D.
2. N-CRLB$_{\lambda_X}$ is monotonically decreasing in the reference microphone SNR, $\lambda_X(l)/\lambda_V(l)$, and the MVDR output SNR $\xi(l)$, Eq. (10).
3. Remarkably, N-CRLB$_{\lambda_V}$ is independent of target signal characteristics (power and direction). This may be understood from the fact that the ML estimate of $\lambda_V(l)$ is based on a target-cancelling beamformer (Sec. 3.1), the output of which is unrelated to the target signal. Furthermore, N-CRLB$_{\lambda_V}$ is independent of frequency. Finally, N-CRLB$_{\lambda_V}$ is independent of the noise covariance matrix $C_V(l)$: this implies that N-CRLB$_{\lambda_V}$ is unaffected by changes in the spatial noise distribution as well as the microphone array geometry. A concrete consequence in the HA application example below is that N-CRLB$_{\lambda_V}$ is the same whether the HA microphone array is located behind the ear of a HA user, or it is located in free-field conditions.
4. N-CRLB$_{\lambda_V} = 1/D$ for $M = 2$. It can be shown that this is identical to the N-CRLB if $\lambda_V(l)$ were estimated in a cheating experiment, where the noise component in the noisy input to the SC filter could be observed in isolation.
5. Finally, it can be verified that $\lambda_X(l)$ is "harder" to estimate than $\lambda_V(l)$, because N-CRLB$_{\lambda_X} \geq$ N-CRLB$_{\lambda_V}$.

5. NOISE REDUCTION FOR HEARING AIDS

In this section we study the characteristics of the proposed estimators in a hearing aid (HA) setup, and we demonstrate their use in the beamformer-directed noise reduction system in Fig. 1.

5.1. Acoustic setup and HA configuration

We consider target speech signals impinging on an $M = 2$ BTE-HA microphone array, worn by a HA user. To do so, we mount the BTE-HA behind the left pinna (≈90 degs.) of a head-and-torso-simulator (HATS) (B&K 4128 [22]) in an anechoic chamber and measure impulse responses from 72 possible sound source positions to each microphone. The possible sound source positions are spaced uniformly in a circle with radius 1.5m centered at the HATS.

Microphone signals are generated by convolving the target signal with the relevant impulse responses. The microphone signals are sampled at a frequency of 20 kHz, and passed through an STFT based analysis filterbank using a frame length $N = 128$, a decimation factor $D_A = 64$, and where $w_A(n)$ is a square-root Hanning window. For each frequency channel, time-invariant vectors $d(l) = d$ are determined from white noise sequences played back from the target position in question and captured by the microphones.

5.2. CRLBs for BTE Hearing Aid

To demonstrate experimentally the CRLBs derived in the previous section for the HA situation, we assume that the HA user is exposed to additive, cylindrically isotropic noise. Noise covariance matrices $C_V(l_0)$ for such a noise scenario are estimated from long-duration white noise sequences simultaneously played back from the 72 positions. Finally, constant values of $\lambda_X(l)$ and $\lambda_V(l)$ are chosen to produce an SNR, SNR$_{ref} = \lambda_X(l)/\lambda_V(l)$, of 0 dB at the reference microphone. In Fig. 2 we evaluate N-CRLB$_{\lambda_X}$ as a function of target angle and frequency; as expected from Eq. (12), N-CRLB$_{\lambda_V}$ is constant (not shown). N-CRLB$_{\lambda_X}$ is a complicated function of angle

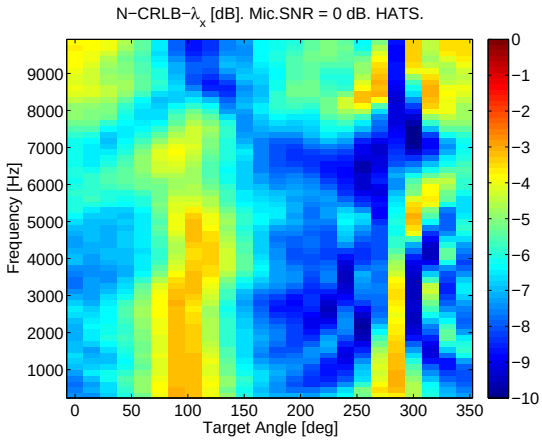

Fig. 2. N-CRLB$_{\lambda_X}$ [dB] vs. frequency and target angle. Reference microphone SNR: 0 dB. $D = 10$. Cylindrically isotropic noise. N-CRLB$_{\lambda_V}$ is independent of target angle and frequency (not shown).

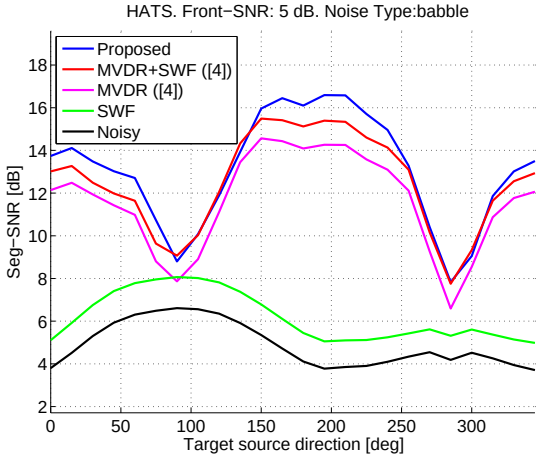

Fig. 3. Seg-SNR vs. target angle for various noise reduction systems in large crowd noise.

and frequency. For targets from the sides (≈90 and ≈270 degs.) estimation accuracy is 4-5 dBs worse than targets from the front/back directions. This can be explained by recalling that $\lambda_{ML,X}(l)$ is based on the noisy signal passed through an MVDR beamformer (Sec. 3.1), which is more efficient for targets located parallel rather than perpendicular to the microphone axis (≈90 and ≈270 degs.).

5.3. Performance Analysis: Noise Reduction in Hearing Aids

Finally, we demonstrate the practical use of the studied ML estimation scheme in a multi-channel enhancement system based on the $M = 2$ BTE-HA described above. The acoustic scene consists of a single target speaker at a variable location in large-crowd noise, generated by convolving different speech signals with each of the 72 pairs of impulse responses and summing the contributions. The noise signals are scaled to realize an SNR of 5 dB at the reference microphone, when the target source originates from the front (0 degs.). Noisy signals are enhanced in the system depicted in Fig. 1, using an MVDR and SWF system. The matrix $\hat{C}_Y(l)$ is found via exponential smoothing with a time constant of 50 ms (corresponding to $D \approx 15$ in the current implementation). Frame indices l_0 are found by thresholding the *a posteriori* SNR, $\zeta(l) = |Y_1(l)|^2/\hat{\lambda}_V(l)$, where $\hat{\lambda}_V(l)$ is the noise psd at the reference microphone, determined by a version of the Minimum Statistics (MS) noise tracker [5] (adapted to the frame length and sample rate used here). Psd estimates $\hat{\phi}_X(l)$ and $\hat{\phi}_V(l)$ - found from $\lambda_{ML,X}(l)$ and $\lambda_{ML,V}(l)$ via Eqs. (8), (9) - are temporally smoothed using the decision-directed approach to produce *a priori* SNR estimates $\xi_{dd}(l)$ [23]. The SWF filter, $g_{SWF}(l) = \xi_{dd}(l)/(\xi_{dd}(l)+1)$, is then applied to the output signal of the MVDR beamformer, to produce an enhanced STFT spectrum. Finally, an enhanced time-domain signal $\hat{x}(n)$ is constructed via an overlap-add procedure [24] using a square-root Hanning synthesis window.

For comparison, we enhance noisy signals in a similar system, but where $C_V(l)$, and subsequently $\phi_V(l)$, are estimated as in [4]. In [4], $C_V(l)$ was estimated in noise-only regions, as determined using the MS noise tracker [5], and fixed in speech presence regions. We also enhance noisy signals using an adaptive MVDR beamformer (Eq. (5)), and a stand-alone SWF, both using noise statistics updated in noise-only regions as determined in [4].

Fig. 3 shows enhancement performance in terms of Segmental SNR (Seg-SNR) [25] as a function of target angle. As expected, Seg-SNR for the noisy signal is maximal for a target direction of ≈90 degrees, i.e., at the left ear where the BTE-HA is mounted. A stand-alone SWF improves Seg-SNR for all target directions. Generally the proposed system performs better than the method in [4], which in turn is better than *MVDR*; we confirmed this order for other distortion measures such as Log-Spectral Distortion [24], and STOI [26] (not shown). Interestingly, the multi-channel methods are better for target directions in the range 20-50 degs., and not at ≈90 degs., where the input Seg-SNR is largest. This optimum angle range can be understood as a tradeoff between the frontal direction (0 degs.), where the beamformer is most efficient (and the estimation errors in $\lambda_{ML,X}(l)$ are smallest with the proposed system, Fig. 2) and 90 degs., where the input SNR is largest. Note also that performance is even better when the target arrives from the rear (180 degs.). This may be explained by the observation that SNR at the rear microphone, which is located behind the pinna, is higher when the target arrives from the rear than from the front. Finally, performance is relatively lower for target angles around 250-330 degrees, because the SNR at the microphones is reduced due to head shadow effects.

6. CONCLUSION

We presented and analyzed a maximum likelihood (ML) method for estimating the power spectral densitities (psd's) of the target and noise signals entering the single-channel (SC) filter in a beamformer-and-SC noise reduction system. We interpret the expressions for the ML estimates in terms of simple filtering operations. Furthermore, we show that the ML estimators are minimum variance, unbiased estimators, and present closed-form expressions for their achievable mean-square error (MSE). In a two-microphone hearing aid context, the MSE for the target psd is complicated and varies by at least 4-5 dB as a function of target direction. The MSE for the noise psd, however, is simple: it is independent of target signal content and direction-of-arrival, frequency, and of microphone locations. Finally, in this hearing aid context, we demonstrate the advantage of the proposed method over a recent method [4] which estimates SC filter characteristics directly from the beamformer output.

7. REFERENCES

[1] S. Doclo and M. Moonen, "GSVD-based Optimal Filtering for Single and Multimicrophone Speech Enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, September 2002.

[2] A. Spriet, M. Moonen, and J. Wouters, "Spatially Pre-Processed Speech Distortion Weighted Multi-Channel Wiener Filtering for Noise Reduction," *Signal Processing*, vol. 84, pp. 2367–2387, December 2004.

[3] K. U. Simmer, J. Bitzer, and C. Marro, "Post-Filtering Techniques," in *Microphone Arrays – Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. 2001, Springer Verlag.

[4] X. Zhang and Y. Jia, "A soft decision based noise cross power spectral density estimation for two-microphone speech enhancement systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2005, pp. 813–816.

[5] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech, Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

[6] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Trans. Audio, Speech, Language Processing*, vol. 16, no. 3, pp. 541–553, 2007.

[7] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise psd tracking with low complexity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2010, pp. 4266–4269.

[8] T. Gerkmann and R. C. Hendriks, "Improved mmse-based noise psd tracking using temporal cepstrum smoothing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2012, pp. 105–108.

[9] R. Zelinski, "A Microphone Array With Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1988, vol. 5, pp. 2578–2581.

[10] I. A. McCowan and H. Bourlard, "Microphone Array Post-Filter Based on Noise Field Coherence," *IEEE Trans. Speech, Audio Processing*, vol. 11, no. 6, pp. 709–716, 2003.

[11] S. Lefkimmiatis and P. Maragos, "Optimum post-filter estimation for noise reduction in multichannel speech processing," in *Proc. 14th European Signal Processing Conference*, 2006.

[12] R. C. Hendriks and T. Gerkmann, "Noise Correlation Matrix Estimation for Multi-Microphone Speech Enhancement," *IEEE Trans. Audio, Speech, Language Processing*, vol. 20, pp. 223 – 233, January 2012.

[13] U. Kjems and J. Jensen, "Maximum likelihood noise covariance matrix estimation for multi-microphone speech enhancement," in *Proc. 20th European Signal Processing Conference (Eusipco)*, 2012, pp. 295–299.

[14] H. Ye and R. D. DeGroat, "Maximum Likelihood DOA Estimation and Asymptotic Cramér-Rao Bounds for Additive Unknown Colored Noise," *IEEE Trans. Signal Processing*, vol. 43, no. 4, pp. 938–949, 1995.

[15] P. Stoica and A. Nehorai, "Performance Study of Conditional and Unconditional Direction-of-Arrival Estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 10, pp. 1783–1795, October 1990.

[16] A Kuklasiński, S. Doclo, S.H. Jensen, and J. Jensen, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *Proc. 22nd European Signal Processing Conference (Eusipco)*, Lisbon, Portugal, 2014.

[17] S. Braun and E. A. P. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. 21st European Signal Processing Conference (Eusipco)*, 2013.

[18] D. R. Brillinger, *Time Series: Data Analysis and Theory*, SIAM, Philadelphia, 2001.

[19] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug 2001.

[20] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall signal processing series. Prentice Hall PTR - Upper Saddle River, New Jersey, 1993.

[21] H. Hung and M. Kaveh, "On the statistical sufficiency of the coherently averaged covariance matrix for the estimation of the parameters of wide-band sources," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 1987, pp. 33–36.

[22] Bruel og Kjaer, "Product data. Head and Torso Simulator (HATS) - Type 4128," http://www.bksv.com/products/transducers/ear-simulators/head-and-torso/hats-type-4128c.aspx.

[23] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[24] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.

[25] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. Int. Conf. Spoken Language Processing*, 1998.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE Trans. Audio, Speech, Language Processing*, vol. 19, no. 7, pp. 2125–2136, September 2011.