Improving robustness of adaptive beamforming for hearing devices

ALASTAIR H. MOORE, PATRICK A. NAYLOR*, AND MIKE BROOKES

Department of Electrical and Electronic Engineering, Imperial College, London, UK

Fixed beamforming for hearing aids is suboptimal due to mismatches in real-world situations between the assumed and encountered sound fields. Adaptive beamforming potentially provides better performance but may degrade it if the characteristics of the signal required by the design procedure are inaccurately estimated. This paper proposes a straightforward but sufficiently rich model for the sound field that can be used to increase the robustness of adaptive beamformer design. A method for estimating the model parameters is also presented. In reverberant acoustic conditions, the proposed method improves performance by > 1 dB even at -16 dB SNR, the lowest signal to noise ratio (SNR) tested. Furthermore, it is shown to be robust in a variety of acoustic conditions which do not conform to the sound field model, and to inaccurate steering of the array.

INTRODUCTION

Current drivers for innovation in microphone array beamforming include the increasing availability of more powerful computational resources, and the increasing significance of several emerging application areas, such as spherical arrays described in Rafaely (2015) and Jarret *et al.* (2017), robot audition as in Tamai *et al.* (2004) and Löllmann *et al.* (2017) and binaural hearing aids discussed in Klasen *et al.* (2007) and Moore *et al.* (2018). The linearly constrained minimum variance (LCMV) family of beamformers are widely used in acoustic beamforming due to their ability to suppress noise without distorting the target signal. The original Capon beamformer (Capon, 1969), or minimum power distortionless response (MPDR) beamformer (van Trees, 2002), minimise the output power given the sample covariance matrix (SCM), whereas the minimum variance distortionless response (MVDR) beamformer design is based on the noise covariance matrix (NCM). Both use a steering vector to set the distortionless constraint on the target signal and, under ideal conditions, they are equivalent. In practice, their sensitivity to errors in the steering vector differs (Cox *et al.*, 1987; Ehrenberg *et al.*, 2010). For the MVDR beamformer, the effect of mis-steering is merely to attenuate the desired signal, whereas for the MPDR signal, cancellation occurs since the distortionless constraint is not matched to the target signal. Furthermore, it is shown in Ehrenberg *et al.* (2010) that an inaccurate estimate of the NCM is preferable to an accurate SCM.

In reverberant environments, even with a perfectly aligned anechoic steering vector,

*Corresponding author: p.naylor@imperial.ac.uk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

coherent reflections originating from the target source cause signal cancellation for the MPDR. Using reverberant relative transfer function (RTF) steering vectors as in Gannot *et al.* (2001), the distortionless constraint preserves the direct path and at least the first few early reflections, reducing the potential for signal cancellation. Effective RTF estimation is an ongoing research problem (Markovich *et al.*, 2009; Markovich-Golan and Gannot, 2015).

Obtaining an estimate of the NCM which is completely uncorrelated with the target speech and is effective in practical applications is difficult to achieve. Informed spatial filtering is a concerted research effort to model the statistics of different signal components from which the total NCM can be obtained (Thiergart and Habets, 2003; Braun and Habets, 2015; Schwartz *et al.*, 2016; Chakrabarty and Habets, 2018; Braun *et al.*, 2018; Moore *et al.*, 2019a). In many cases, estimated hyper-parameters such as the speech presence probability (SPP), coherent to diffuse ratio (CDR), or one or more directions of arrival (DOAs) control when to update each statistic. Inevitably such estimates become less accurate at low signal to noise ratios (SNRs) and in time-varying scenarios which may, for example, lead to target energy leaking into the NCM.

Robust beamformers have been proposed which reduce sensitivity to errors and increase the white noise gain at the expense of reduced directivity, for example in Cox *et al.* (1987) and Li *et al.* (2003). These generally involve diagonal loading of the covariance matrix. Ultimately, to remove all possibility of signal cancellation the conservative approach often adopted in real-world implementations is to design a fixed, super-directive beamformer using an assumed noise model (Bitzer and Simmer, 2001).

In this paper, we propose a simple model of the sound field that is sufficiently rich to describe complex scenes and whose parameters can be estimated at low SNRs. We assume that calibration measurements of the array manifold are available and that the steering direction is known. Using this information, a method for estimating the time-varying parameters of the sound field model is proposed. The adequacy of the proposed model and resulting SCM is evaluated in the specific context of MPDR beamforming for binaural hearing aids (HAs) but it can equally be applied to other filter structures and array geometries. In this case, as is customary, a known target direction is realized by fixing the steering direction towards the front of the head and requiring that the listener turn to face the desired talker.

FORMULATION AND PROPOSED MODEL

The time domain signal received at the m^{th} microphone in an array is denoted

$$y_m(t) = \sum_{l=1}^{L} x_{m,l}(t) + v_m(t) \qquad \text{(Eq. 1)}$$

where t is the time index, l is the source index, $x_{m,l}(t)$ is the signal due to the l^{th} source and $v_m(t)$ is sensor noise. In a reverberant enclosure,

$$x_{m,l}(t) = h_{m,l}(t) * s_l(t) \tag{Eq. 2}$$

where $s_l(t)$ is the signal emitted by the l^{th} source, $h_{m,l}(t)$ is the acoustic impulse response (AIR) from the l^{th} source to the m^{th} microphone, and $*$ denotes convolution. Decomposing $h_{m,l}(t)$ into the direct path, $h^{(d)}_{m,l}(t)$, and reflected components, $h^{(r)}_{m,l}(t)$, Eq. 2 can be rewritten

$$x_{m,l}(t) = (h^{(d)}_{m,l}(t) + h^{(r)}_{m,l}(t)) * s_l(t) \tag{Eq. 3}$$

$$= h^{(d)}_{m,l}(t) * s_l(t) + h^{(r)}_{m,l}(t) * s_l(t) \tag{Eq. 4}$$

$$= x^{(d)}_{m,l}(t) + x^{(r)}_{m,l}(t) \tag{Eq. 5}$$

where $x^{(d)}_{m,l}(t)$ and $x^{(r)}_{m,l}(t)$ are the direct path and reflected components of $x_{m,l}(t)$, respectively.

Combining Eq. 1 and Eq. 5, the microphone signals can be written

$$y_m(t) = \sum_{l=1}^{L} x^{(d)}_{m,l}(t) + \sum_{l=1}^{L} x^{(r)}_{m,l}(t) + v_m(t) \tag{Eq. 6}$$

and can equivalently be expressed in the short time Fourier transform (STFT) domain as

$$Y_m(\nu,\ell) = \sum_{l=1}^{L} X^{(d)}_{m,l}(\nu,\ell) + \sum_{l=1}^{L} X^{(r)}_{m,l}(\nu,\ell) + V_m(\nu,\ell) \tag{Eq. 7}$$

where capitalized letters denote the STFT of the quantities denoted by the corresponding lowercase letters in Eq. 6, and ν and ℓ are the frequency and frame indices respectively. Stacking the signals for all M microphones in an array to give, for example, $\mathbf{y}(\ell) = \begin{bmatrix} Y_1(\ell) & \dots & Y_M(\ell) \end{bmatrix}^T$, Eq. 7 then becomes

$$\mathbf{y}(\ell) = \sum_{l=1}^{L} \mathbf{x}^{(d)}_l(\ell) + \sum_{l=1}^{L} \mathbf{x}^{(r)}_l(\ell) + \mathbf{v}(\ell) \tag{Eq. 8}$$

where $(\cdot)^T$ denotes the transpose, and since all frequency bins are processed independently, the dependence on ν has been dropped for clarity.

The proposed signal model makes four simplifying assumptions: (i) all sources are in the far field, such that $h^{(d)}_{m,l}(t)$ is identical to the response of the array to a plane-wave from the same direction as the l^{th} source, up to a scalar gain and time shift; (ii) the array is sufficiently compact that the RTF to each microphone with respect

to the reference microphone can be represented by a multiplicative constant in the STFT domain (Avargel and Cohen, 2007); (iii) the direct path signals are W-disjoint orthogonal (Yilmaz and Rickard, 2004), such that in each time-frequency bin, a single source is dominant, (iv) the sum of all reflected signals reduces to a diffuse field which, by definition, is isotropic since the incident power from all directions is the same. With these assumptions, Eq. 8 reduces to

$$\dot{\mathbf{y}}(\ell) = \mathbf{a}(\Omega(\ell))\dot{S}_{l(\ell)}(\ell) + \boldsymbol{\gamma}(\ell) + \mathbf{v}(\ell) \quad \text{(Eq. 9)}$$

where $\boldsymbol{\gamma}(\ell)$ is the diffuse noise signal, $l(\ell)$ and $\Omega(\ell)$ are the index and corresponding DOA, respectively, of the dominant source in the ℓ^{th} frame (which may be different at each frequency), $\dot{S}_{l(\ell)}(\ell)$ is the signal due to the dominant source as observed at the arbitrarily selected reference microphone and $\mathbf{a}(\phi)$ is the plane-wave array manifold expressed as the RTF to each microphone with respect to the reference microphone.

The covariance of the microphone signals is

$$\mathbf{R}_{\mathbf{y}}(\ell) = \mathbb{E}\{\mathbf{y}(\ell)\mathbf{y}^H(\ell)\} \quad \text{(Eq. 10)}$$

where $\mathbb{E}\{\cdot\}$ is the expectation operator and $(\cdot)^H$ denotes the conjugate transpose. Using the signal model defined in Eq. 9 and assuming the three terms are uncorrelated

$$\mathbf{R}_{\dot{\mathbf{y}}}(\ell) = \mathbb{E}\{|\dot{S}^2_{l(\ell)}(\ell)|\}\mathbf{a}(\Omega(\ell))\mathbf{a}^H(\Omega(\ell)) + \mathbb{E}\{\boldsymbol{\gamma}(\ell)\boldsymbol{\gamma}^H(\ell)\} + \mathbb{E}\{\mathbf{v}(\ell)\mathbf{v}^H(\ell)\} \quad \text{(Eq. 11)}$$

where $(\cdot)^*$ is the conjugate.

It can now be seen that each term on the right hand side of Eq. 11 can be expressed as the product of a fixed matrix and a scalar parameter. This leads to

$$\mathbf{R}_{\dot{\mathbf{y}}}(\ell) = \sigma_d(\ell)\mathbf{R}_{\mathbf{a}}(\Omega(\ell)) + \sigma_{\gamma}(\ell)\mathbf{R}_{\gamma} + \sigma_v(\ell)\mathbf{R}_{\mathbf{v}} \quad \text{(Eq. 12)}$$

where the covariance is defined by four parameters $\Omega(\ell)$, $\sigma_d(\ell)$, $\sigma_{\gamma}(\ell)$ and $\sigma_v(\ell)$ denoting, respectively, the DOA of the plane-wave component and the powers of the plane-wave, diffuse and sensor noise components.

MODEL PARAMETER ESTIMATION

A method is presented to estimate the parameters of the signal model proposed in Eq. 12, and use them to obtain an estimate of the NCM. The algorithm operates directly in the STFT domain where a recursive estimate, $\hat{\mathbf{R}}_{\mathbf{y}}(\ell)$, of the sample covariance matrix, $\mathbf{R}_{\mathbf{y}}(\ell)$, is obtained as

$$\hat{\mathbf{R}}_{\mathbf{y}}(\ell) = \alpha\hat{\mathbf{R}}_{\mathbf{y}}(\ell-1) + (1-\alpha)\mathbf{y}(\ell)\mathbf{y}^H(\ell) \quad \text{(Eq. 13)}$$

where α defines the time constant.

The model parameters can then be found from the solution to the optimization problem

$$\underset{\Omega(\ell),\sigma_d(\ell),\sigma_{\gamma}(\ell),\sigma_v(\ell)}{\arg\min} \{||\hat{\mathbf{R}}_{\mathbf{y}}(\ell) - \mathbf{R}_{\dot{\mathbf{y}}}(\ell)||_F\} \quad \text{(Eq. 14)}$$

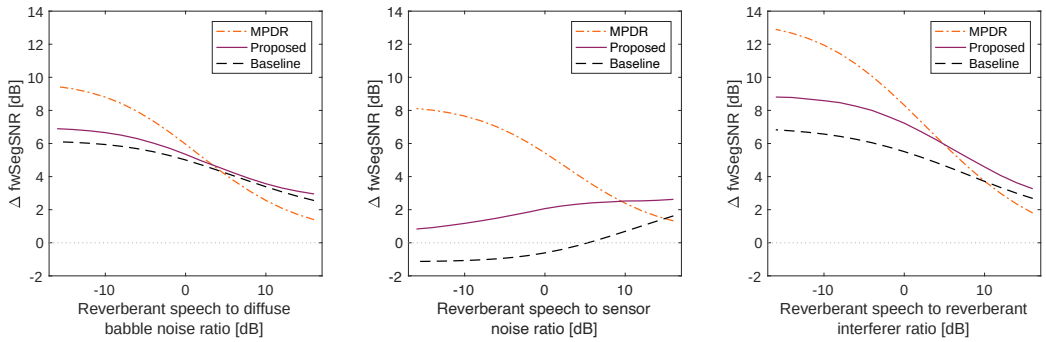


Fig. 1: Improvement in A-weighted segmental SNR as a function of SNR with respect to (a) reverberant babble from 16 directions on circle, (b) sensor noise, (c) interfering speech from −67.5°. In each condition the SNR with respect to the other noise types is 20 dB.

where $||\cdot||_F$ denotes the Frobenius norm. Many approaches to solving Eq. 14 are available. The approach adopted here is to obtain the ordinary least squares solution for $\sigma_d(\ell)$, $\sigma_\gamma(\ell)$ and $\sigma_v(\ell)$ for a candidate set of values of $\Omega(\ell)$, from which the best fit is selected. The final NCM estimate is given by Eq. 12 using the parameter estimates obtained.

SIMULATION EXPERIMENTS

The efficacy of the proposed method in the context of MPDR/MVDR beamforming is evaluated in two experiments. In the first, algorithm performance is assessed as a function of SNR where a single type of noise is dominant. In the second, six different scenarios are considered in which, like real-world acoustic environments, the composition of the sound field is more complicated. Listening examples are available at `https://squaresetsound.com/demos/constrained-covariance-matrix-estimation-2019`.

Microphone signals are simulated for a $7.9 \times 6.0 \times 3.5$ m room with a reverberation time of 250 ms according to Eq. 1 and Eq. 2. Anechoic speech is convolved with hearing aid room impulse responses (HARIRs) measured from a horizontal ring of 16 loudspeakers positioned at azimuth angles, $\phi \in \{0°, 22.5°, \ldots, 337.5°\}$, to a pair of behind the ear (BTE) hearing aids ($M = 4$) worn by a head and torso simulator (HATS) (subject 42; Moore *et al.*, 2019b). Sensor noise is simulated using independent identically distributed Gaussian noise which is filtered to match the spectra of real sensor noise recordings for the microphones used in Moore *et al.* (2019b).

All beamformers are designed based on a repeated set of AIR measurements for the same hearing aids and HATS (as per subject 42; Moore *et al.*, 2019b) but made on a different day, after complete removal and replacement of the hearing aids from the

mannequin, and the mannequin from the measurement room. These hearing aid head-related impulse responses (HAHRIRs) (subject s28; Moore *et al.*, 2019b) are truncated to remove reflections from the room and so contain only direct-path propagation. The steering vector, $\mathbf{d} = \mathbf{a}(0)$, is defined here as the RTF with respect to the front right microphone for a plane-wave arriving from $\phi = 0$.

The covariance matrix for a diffuse field is assumed to be cylindrically isotropic, since real rooms tend to have more absorption in the floor and/or ceiling compared to the walls (Schwarz *et al.*, 2015). It is computed by discretising

$$\mathbf{R}_\gamma = \int_{\Omega=0}^{2\pi} \mathbf{h}^{(d)}(\Omega)\mathbf{h}^{(d)H}(\Omega)d\Omega \qquad \text{(Eq. 15)}$$

to the 7.5° resolution of the HAHRIRs.

Beamformer weights are calculated according to

$$\mathbf{w} = {\mathbf{R}_\varepsilon}^{-1}\mathbf{d}\left[\mathbf{d}^H{\mathbf{R}_\varepsilon}^{-1}\mathbf{d}\right]^{-1} \qquad \text{(Eq. 16)}$$

where $\mathbf{R}_\varepsilon = \mathbf{R} + \varepsilon\mathbf{I}$, $\mathbf{I}$ is the identity matrix and $\varepsilon \geq 0$ is set to limit the condition number of $\mathbf{R}_\varepsilon$ to $\leq$100. The baseline method assumes cylindrically isotropic noise (i.e., $\mathbf{R} = \mathbf{R}_\gamma$). The proposed method uses the estimated covariance matrix from Eq. 12 (i.e., $\mathbf{R} = \mathbf{R}_{\dot{\mathbf{y}}}(\ell)$ with the parameters from Eq. 14). The robust MPDR method uses the estimated sample covariance matrix from Eq. 13 (i.e., $\mathbf{R} = \hat{\mathbf{R}}_{\mathbf{y}}(\ell)$). It should be noted that the baseline method is signal independent (fixed), whereas the proposed method and MPDR method are adaptive with, respectively, 4 and $M(M+1)/2 = 10$ estimated parameters per time-frequency cell.

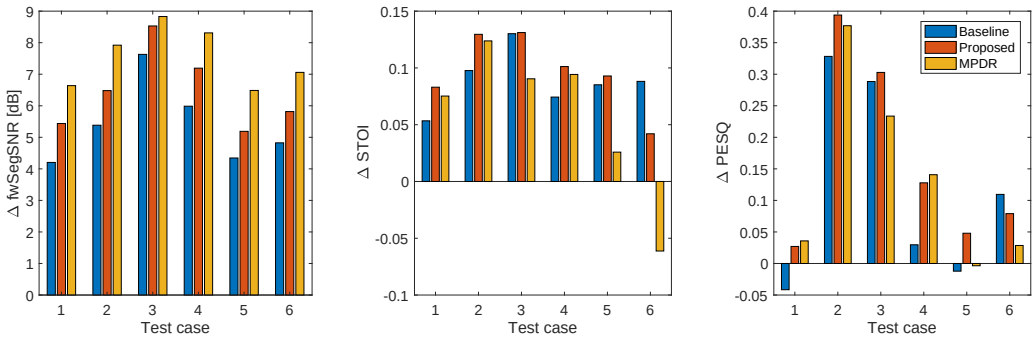
Signals are processed at a sample rate of 20 kHz in the STFT domain with 16 ms frames overlapping by 50 %. The time constant for recursive estimation of $\hat{\mathbf{R}}_{\mathbf{y}}(\ell)$ in Eq. 13 is chosen to be 50 ms in the following experiments.

Experiment 1

The spatial arrangement of sound sources is fixed throughout Experiment 1. The desired source is male speech from $\phi = 0°$, and there are three noise sources: (a) an interferer (male speech) at $\phi = -67.5°$ (to the listener's right); (b) babble noise from sixteen equally-spaced azimuths on the horizontal plane, such that powers of the direct path signals arriving from all azimuth directions are the same; (c) sensor noise.

The levels of the target and interferer speech sources are measured as the average active level in dB of the reverberant signals at the two front microphones, when each sound source is presented from $\phi = 0°$, as defined in ITU-T (1993) and Brookes (1997). Sound presentation from other angles (i.e., interferers) therefore includes the effect of the natural directivity of the head/array geometry. The levels of the noise signals are measured as the average power at the front two microphones.

The level of the desired source is fixed, and in each of three test cases, the effect of varying the level of one, dominant, noise source is assessed whilst keeping the other



(a) **(b)** **(c)**

Fig. 2: Improvement in (a) A-weighted segmental SNR (b) STOI and (c) PESQ for the 6 test cases defined in Table 1.

two fixed at −20 dBr with respect to the desired source. Performance is evaluated in terms of the improvement in frequency-weighted segmental SNR (fwSegSNR) at the reference microphone where the clean target is the direct path component of the desired speech. The improvement in fwSegSNR is an appropriate metric as it quantifies the noise reduction only during periods of target speech activity and changes can be easily interpreted (in dBs) regardless of the specific listening situation.

Figure 1shows that in all cases the proposed method outperforms the baseline. In Figure 1(a), where the dominant noise is babble, the diffuse noise model of the baseline is a reasonably good approximation, and the benefit of the proposed method is smallest. In Figure 1(b), where the dominant noise is uncorrelated between sensors (i.e., spatially white) the baseline method actually reduces the fwSegSNR, which is consistent with the well known trade-off between directivity and white noise gain. In Figure 1(c), where the dominant noise source is reverberant interfering speech, the benefit of the proposed method is most clearly seen with about 1 dB improvement over a wide range of signal to interference ratios (SIRs).

In all cases the MPDR beamformer performs the best at low SNRs but even worse than the baseline at high SNRs. At low SNRs the estimated sample covariance matrix is dominated by noise and so good noise reduction is achieved. In contrast, at high SNRs the estimated sample covariance matrix contains the direct path target and coherent reflections which leads to target cancellation. Experiment 2 investigates the robustness to model violations and employs additional metrics which further highlight the degradation caused by the MPDR method.

Experiment 2

As is well-known, sound fields in real-world situations do not normally conform to the idealised situation of having a single dominant noise type. To evaluate the effect

	Interferer Male @ −67.5°	Interferer Female @ 67.5°	Babble directions	Steering error
1	✓		all	
2	✓	✓	all	
3			22.5°... 157.5°	
4	✓		22.5°... 157.5°	
5	✓		all	7.5°
6	✓		all	15°

Table 1: Test case definitions.

of model violation in complex scenarios, the following additional noise sources are defined: (a) an interferer (female speech) at $\phi = 67.5^\circ$ (to the listener's left) and (b) babble noise only from the seven DOAs to the listener's left (i.e., $22.5^\circ \leq \phi \leq 157.5^\circ$). In this experiment, the levels of the target and all active noise sources are equal, except sensor noise, which is always present at −20 dB with respect to the target. Table 1 defines which noise sources are active in each test case. Test Case 1 has the same spatial arrangement as in Experiment 1, but with the levels of intefering speech and babble being equal. Test Case 2 adds a second interferer. Test Case 3 has non-isotropic babble with no interferers, and Test Case 4 reinstates the male interferer to the right. Test Cases 5 and 6 are the same as Test Case 1 but consider the effect of missteering, where the listener's head is not directly facing the desired source.

In addition to the improvement in fwSegSNR, we also consider the improvements in short-time objective intelligibility measure (STOI) (Taal *et al.*, 2011) and PESQ (ITU-T, 2003).

Figure 2 shows that in Test Cases 1 to 5, all metrics suggest that the proposed method outperforms the baseline. Only in Test Case 6, where the steering misalignment is 15°, do the STOI and PESQ metrics suggest that performance of the proposed method is degraded. Whilst the MPDR method is effective at reducing the noise, as indicated by its superlative improvement in fwSegSNR, both the STOI metric and informal listening suggest that there is also signal degradation. Consistent with the literature (Li *et al.*, 2003; Ehrenberg *et al.*, 2010), the MPDR beamformer is particularly sensitive to steering errors as seen in Test Cases 5 and 6.

DISCUSSION AND CONCLUSIONS

The proposed model of the sound field as the weighted sum of three idealised components allows a wide range of real-world sound fields to be approximated. By constraining the allowed DOA of the plane-wave component to a fixed set of candidates, the potential for signal cancellation during desired speech activity is minimised. When the desired speech is dominant, provided the steering error is not

too large, it is likely that the DOA coinciding with the look direction is selected, even in the presence of reflections, and so the MVDR's distortionless constraint ensures that no cancellation of the direct path wave-front occurs. The remaining components of the covariance matrix are a combination of diffuse and spatially white noise and so are as benign as a fixed beamformer. When the desired speech is not dominant or absent, the contribution of the plane-wave component allows the estimated covariance matrix to adapt, at least to some extent, to the irregularities of the encountered sound field, improving the attenuation compared to an ideal model of the noise distribution. By continuously adapting the estimated covariance matrix, the method can respond immediately to changes in the acoustic scene. Combining the proposed method with head-tracker informed beam-steering as in Moore *et al.* (2018), it is feasible to relax the requirement for the user to face the target source. Simulation experiments using measured reverberant impulse responses and challenging levels of realistic noise show that the proposed method outperforms a fixed beamformer by $\geq$1 dB over a range of acoustic scenarios and is more robust than a conventional, diagonally loaded MPDR beamformer.

ACKNOWLEDGEMENTS

This work was supported by the UK Engineering and Physical Sciences Research Council [grant number EP/M026698/1].

REFERENCES

Avargel, Y., and Cohen, I. (**2007**), "On multiplicative transfer function approximation in the short-time Fourier transform domain," IEEE Signal Process. Lett., **14**(5), 337-340.

Bitzer, J., and Simmer, K.U. (**2001**), "Superdirective microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, 19-38.

Braun, S., and Habets, E.A.P (**2015**), "A multichannel diffuse power estimator for dereverberation in the presence of multiple sources," EURASIP J. Audio Speech Music Process., vol. 2015, no. 1, p. 34.

Braun, S., Kuklasiński, A., Schwartz, O., Thiergart, O., Habets, E.A.P., Gannot, S., Doclo, S., and Jensen, J. (**2018**), "Evaluation and comparison of late reverberation power spectral density estimators," IEEE/ACM Trans. Audio Speech Lang. Process., **26**(6), 1056-1071.

Brookes, D.M. (**1997**), "VOICEBOX: A speech processing toolbox for MATLAB," 1997–2016. [Online]. Available: `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`

Capon, J. (**1969**), "High resolution frequency-wavenumber spectrum analysis," Proc. IEEE, **57**, 1408-1418.

Chakrabarty, S., and Habets, E.A.P. (**2018**), "A Bayesian approach to informed spatial filtering with robustness against DOA estimation errors," IEEE/ACM Trans.

Audio Speech Lang. Process., **26**(1), 145-160.

Cox, H., Zeskind, R.M., and Owen, M.M. (**1987**), "Robust adaptive beamforming," IEEE Trans. Acoust. Speech Signal Process., **35**(10), 1365-1376.

Ehrenberg, L., Gannot, S., Leshem, A., and Zehavi, E. (**2010**), "Sensitivity analysis of MVDR and MPDR beamformers," Proc. IEEE Conv. Electrical and Electronics Engineers, 416-420.

Gannot, S., Burshtein, D., and Weinstein, E. (**2001**), "Signal enhancement using beamforming and nonstationarity with applications to speech," IEEE Trans. Signal Process., **49**(8), 1614-1626.

ITU-T (**1993**), "Objective measurement of active speech level," Intl. Telecommunications Union (ITU-T), Recommendation P.56, Mar. 1993.

ITU-T (**2003**), "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Intl. Telecommunications Union (ITU-T), Recommendation P.862, Nov. 2003.

Jarrett, D.P., Habets, E.A.P., and Naylor, P.A. (**2017**), *Theory and Applications of Spherical Microphone Array Processing*, ser. Springer Topics in Signal Processing. Springer International Publishing, 2017.

Klasen, T.J., Bogaert, T.V. den, Moonen, M., and Wouters, J. (**2007**), "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," IEEE Trans. Signal Process., **55**(4), 1579-1585.

Li, J., Stoica, P., and Wang, Z. (**2003**), "On robust Capon beamforming and diagonal loading," IEEE Trans. Signal Process., **51**(7), 1702-1715.

Löllmann, H.W., Moore, A.H., Naylor, P.A., Rafaely, B., Horaud, R., Mazel, A., and Kellermann, W. (**2017**), "Microphone array signal processing for robot audition," Proc. HSCMA, 51–55.

Markovich, S., Gannot, S., and Cohen, I. (**2009**), "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," IEEE Trans. Audio, Speech, Lang. Process., **17**(6), 1071-1086.

Markovich-Golan, S. and Gannot, S. (**2015**), "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," Proc. ICASSP, 544–548.

Moore, A.H., Lightburn, L., Xue, W., Naylor, P.A., and Brookes, M. (**2018**), "Binaural mask-informed speech enhancement for hearing aids with head tracking," Proc. IWAENC.

Moore, A.H., Xue, W., Naylor, P.A., and Brookes, M. (**2019**), "Noise covariance matrix estimation for rotating microphone arrays," IEEE/ACM Trans. Audio Speech Lang. Process., **27**(3), 519-530.

Moore, A.H., de Haan, J.M., Pedersen, M.S., Naylor, P.A., Brookes, M., and Jensen, J. (**2019**), "Personalized signal-independent beamforming for binaural hearing aids," J. Acoust. Soc. Am., **145**, 971–2981.

Rafaely, B. (**2015**), *Fundamentals of Spherical Array Processing*, ser. Springer Topics in Signal Processing. Berlin Heidelberg: Springer-Verlag, 2015.

Schwartz, O., Gannot, S., and Habets, E.A.P (**2016**), "Joint estimation of late reverberant and speech power spectral densities in noisy environments using frobenius norm," Proc. EUSIPCO, 1123–1127.

Schwarz, A., and Kellermann, W. (**2015**), "Coherent-to-diffuse power ratio estimation for dereverberation," IEEE/ACM Trans. Audio Speech Lang. Process., **23**(6), 1006–1018.

Taal, C.H., Hendriks, R.C., Heusdens, R., and Jensen, J. (**2011**), "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Process., **19**(7), 2125-2136.

Tamai, Y., Kagami, S., Amemiya, Y., Sasaki, Y., Mizoguchi, H., and Takano, T. (**2004**), "Circular microphone array for robot's audition," Proc. IEEE Sensors, 565–570.

Thiergart, O., and Habets, E.A.P. (**2013**), "An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates," Proc. ICASSP, 659–663.

van Trees, H.L. (**2002**), *Optimum Array Processing*, ser. Detection, Estimation and Modulation Theory. John Wiley & Sons, Inc., 2002.

Yilmaz, O., and Rickard, S. (**2004**), "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Process., **52**(7), 1830-1847.