DIRECTIONAL INTERFERENCE SUPPRESSION USING A SPATIAL RELATIVE TRANSFER FUNCTION FEATURE

Sebastian Braun, Ivan Tashev

Microsoft Research,
Redmond, WA, USA

ABSTRACT

Many speech enhancement systems consist of a beamformer and a spectral suppression postfilter. While it is well understood how to design beamformers to suppress either non-directional or directional interference, their suppression ability is limited by the number, type and positions of microphones. However, spectral postfilters that can further increase the suppression, are usually only designed to suppress non-directional noise. In this work, we propose a spatially selective spectral suppressor addressing directional and non-directional interference. The proposed suppressor is based on the relative transfer function of the target source location. While existing directional suppression techniques are limited to farfield scenarios or certain microphone geometries, we propose a general approach without restrictions on the microphone array and without farfield assumption. We show that the proposed spatial suppressor is able to suppress noise and directional interfering speakers, which substantially improves the performance of speech recognizer, and reduces undesired recognition of interfering talkers.

***Index Terms*—** Spatial filtering, directional gain, postfiltering

1. INTRODUCTION

Most modern devices used for speech communication and sound capturing are equipped with multiple microphones. This allows the use of spatial filtering techniques to, for example, capture the sound only from a certain direction while suppressing sound from other directions. Linear spatial filtering or *beamforming* [1], which combines the microphone signals with a certain weighting to a single signal, is a popular, well studied, and powerful technique in acoustic signal processing and speech enhancement. However in practice, most devices are equipped only with a small number of microphones to save hardware cost and computational complexity. This limits the maximum possible amount of diffuse noise suppression, which strongly depends on the number of microphones. While in theory linear spatial filters can completely cancel coherent directional sources, complete cancellation is rarely achieved in practice due to source localization errors and due to reverberation, which remains audible as it arrives from different directions than the direct path.

To improve the spatial selectivity and suppression performance of spatial filters, spectral suppression postfilters are commonly used as postfilters. Suppression techniques used for stationary noise suppression, e. g. based on some kind of single-channel voice activity detector (VAD) [2, 3, 4] as well as multichannel speech presence probability (SPP) estimators [5] are not spatially selective. Spatial coherence-based methods such as [6, 7, 8, 9, 10] are designed to only suppress non-directional interference, such as diffuse noise, while directional interfering sources, such as undesired talkers, are either captured as desired sound or can not be sufficiently suppressed using these methods. The methods proposed in [11, 12] require prior knowledge of the spatial coherence of an interfering source in addition to the stationary noise coherence.

Some spectral suppressors that are explicitly spatially selective have been proposed in [13, 14], as well as VADs using spatial cues [15, 16] could be used to design spatially selective suppression filters. However, all of these methods have at least one of the following limitations: (i) they exploit only phase differences, but neglect the magnitudes, which might be of high importance in nearfield beamforming applications or specific array geometries with acoustic shading between the microphones; (ii) the microphone geometry is restricted to two microphones, or even to a certain source position, e. g. the broadside direction.

In a farfield scenario, where the steering vector to the desired source is independent of the distance, estimated narrowband direction-of-arrivals (DOAs) could mapped via a target directivity function to a suppression gain to suppress sources arriving outside of the region of interest [17]. However, this does not generalize to a nearfield scenario, where the relative transfer function (RTF) is distance dependent. To deal with this case, DOA estimators using a discrete set of RTFs such as [18, 19] could be used, where the target nearfield RTF could be added to a set of farfield RTFs. However, the choice of an optimal mapping from estimated RTF from the dataset is an open question, which leads to only heuristic solutions. In [20], an array-specific pre-trained spatial dictionary is used.

In this paper, we propose a probabilistic approach to obtain a spectral suppression gain. Our aim is to develop a spatially selective suppressor that neither depends on the DOA, nor requires a full set of RTFs to all possible source locations, which saves computational complexity and simplifies the solution. Our method is formulated generally without restrictions on the microphone array and source positions by taking magnitude and phase differences into account, which is important as we aim at a nearfield beamforming scenario. We are equally interested in suppressing non-directional noise as well as highly coherent directional interfering speakers, by predefining the target source location. The proposed method is evaluated and compared to other suitable methods in a nearfield beamforming scenario using a head-mounted display in the presence of interfering talkers and ambient noise.

2. SIGNAL MODEL AND ENHANCEMENT SYSTEM

We assume a general device equipped with M microphones. The m-th microphone signal for $m = \{1, \dots, M\}$ is given in the short-time Fourier transform (STFT) domain by

$$Y_m(k,n) = A_m(k,\mathbf{r}) S_{\mathbf{r}}(k,n) + V_m(k,n) \tag{1}$$

To appear in *Proc. ICASSP2019, May 12-17, 2019, Brighton, UK*

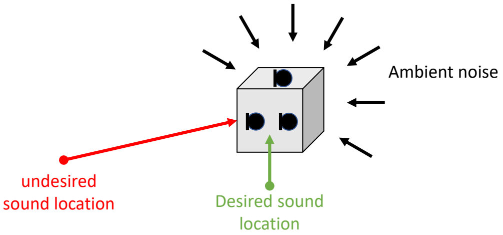

Fig. 1. Description of the acoustic scene. We assume only one dominant directional source per time-frequency bin, either from the desired or an undesired source location.

where $A_m(k,\mathbf{r})$ is the direct path transfer function of some source signal $S_\mathbf{r}(k,n)$ at location $\mathbf{r}$ to the m-th microphone, $V_m(k,n)$ is ambient noise at the m-th microphone, and k and n are the frequency and time indices, respectively. The noise $V_m(k,n)$ can contain noise as well as reverberation. Note that due to the spectral sparsity of speech, the signal model can be valid even for double talk of two sources at different locations by assuming only one dominant source per time-frequency bin. The sound scene model is shown in Fig. 1.

In contrast to acoustic transfer functions, RTFs are easier to estimate. By choosing the first microphone as reference, without loss of generality, the RTFs from the first to the other microphones related to source location $\mathbf{r}$ are given by

$$B_{m,1}(k,\mathbf{r}) = \frac{A_m(k,\mathbf{r})}{A_1(k,\mathbf{r})}. \tag{2}$$

Our goal is to obtain the source signal only from the desired location $\mathbf{r}_\text{d}$ as received by the first microphone, while suppressing noise and sources form other locations, i. e.,

$$X_1(k,n) = \begin{cases} A_1(k,\mathbf{r}_\text{d})S_\text{d}(k,n) & \text{if} \quad \mathbf{r} = \mathbf{r}_\text{d} \\ 0 & \text{if} \quad \mathbf{r} \neq \mathbf{r}_\text{d}. \end{cases} \tag{3}$$

The enhancement system may be a two-stage system, a beamformer followed by a non-linear post-filter, which is a spectral suppressor, as shown in Fig. 2. An estimate of the desired signal (3) is obtained by applying the beamformer weights $W_m(k)$ and the time-varying postfilter weight $W_\text{P}(k,n)$ to the input signals, i. e.,

$$\widehat{X}_1(k,n) = W_\text{P}(k,n) \sum_{m=1}^{M} W_m^*(k)\, Y_m(k,n). \tag{4}$$

The beamformer $W_m(k)$ may be of any type, while all beamformers typically depend on the RTFs. If the desired source location $\mathbf{r}_\text{d}$ is given, the RTFs $B_{m,1}(k,\mathbf{r}_\text{d})$ can be determined e. g. using an analytic sound propagation model, or measuring the acoustic transfer functions in an anechoic chamber. In this work, we choose the time-invariant superdirective minimum variance distortionless response (MVDR) beamformer [1] for $W_m(k)$, while in the following, we propose a method to obtain a spatially selective postfilter $W_\text{P}(k,n)$.

3. PROPOSED RTF-BASED SPATIAL SUPPRESSOR

In this section, we derive a spectral suppressor utilizing the known RTFs of the desired source location and current RTF estimates from the microphone signals. From these two RTFs sets, we build a spatial correlation feature to obtain an indicator for the probability, that the sound in the current time-frequency bin originates from the desired source location $\mathbf{r}_\text{d}$, or from somewhere else.

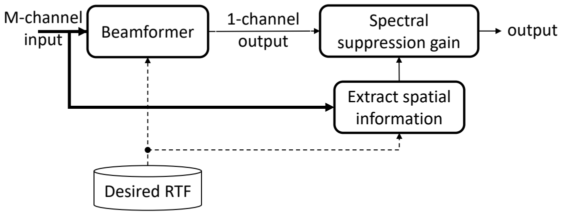

Fig. 2. System consisting of a beamformer and a spatial suppressor utilizing spatial information.

3.1. RTF-based correlation feature

By neglecting the noise term in (1), the RTF can be estimated from the microphone signals in the least-squares sense by

$$\widehat{B}_{m,1}(k,n) = \frac{\mathrm{E}\left\{Y_m(k,n)Y_1^*(k,n)\right\}}{\mathrm{E}\left\{|Y_1(k,n)|^2\right\}} \tag{5}$$

where the expectation operator $\mathrm{E}\{\cdot\}$ can be approximated by first-order recursive smoothing with small time constant. Note that in the presence of noise, the RTF estimate given by (5) is biased. Although there exist a variety of more sophisticated and unbiased RTF estimators [21, 22, 23], we aim to keep the computational complexity low.

Let us define the estimated RTF vector and the *a priori* determined RTF vector related to the desired source location $\mathbf{r}_\text{d}$ as

$$\widehat{\mathbf{b}}(k,n) = \left[\widehat{B}_{2,1}(k,n) \quad \dots \quad \widehat{B}_{M,1}(k,n)\right]^T, \tag{6}$$

$$\mathbf{b}_\text{d}(k) = \left[B_{2,1}(k,\mathbf{r}_\text{d}) \quad \dots \quad B_{M,1}(k,\mathbf{r}_\text{d})\right]^T, \tag{7}$$

which are both vectors of length $M-1$.

As a distance measure between those vectors, we propose the normalized vector inproduct, which can also be interpreted as the cosine of the hermitian angle [24]

$$\Delta = \cos\left\langle \mathbf{b}_\text{d}(k), \widehat{\mathbf{b}}(k,n)\right\rangle = \frac{\Re\left\{\mathbf{b}_\text{d}^H(k)\widehat{\mathbf{b}}(k,n)\right\}}{\|\mathbf{b}_\text{d}(k)\|\,\|\widehat{\mathbf{b}}(k,n)\|}, \tag{8}$$

where $\Re\{\cdot\}$ is the real part operator. Note that $-1 \leq \Delta(k,n) \leq 1$ is bounded. We suspect the feature $\Delta(k,n)$ to be close to one, when the estimated RTF is close to the desired source location, otherwise we expect the cosine angle to be smaller than one, or even negative.

3.2. Probability distribution of the RTF feature

If we assume that additive noise $V_m(k,n)$ is always present, it is unlikely that $\widehat{\mathbf{b}}(k,n)$ and $\mathbf{b}_\text{d}(k,\mathbf{r}_\text{d})$ will match identically. Therefore, we investigate the probability distribution of the feature $\Delta(k,n)$ under the two hypotheses:

- H_d: Speech from the desired target location $\mathbf{r}_\text{d}$ and noise are present.
- H_0: Only noise, or noise plus directional sound from undesired locations different from $\mathbf{r}_\text{d}$ are present.

We are interested in the posterior probability density function (PDF) of $\Delta(k,n)$ during speech activity from the target location $p(\Delta|H_\text{d})$, and during target speech absence $p(\Delta|H_0)$.

In following we present an analysis of the distribution of Δ using three signals for different conditions, that were generated as

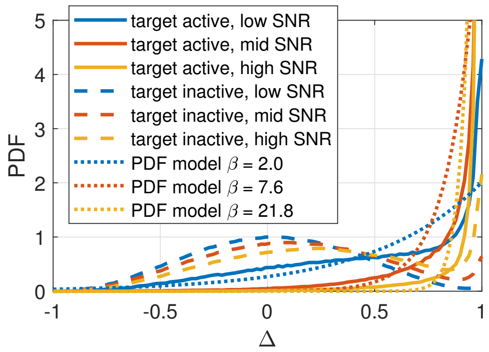

Fig. 3. Probability distributions of RTF feature during target speech and inactive target speech, $p(\Delta|H_\mathrm{d})$ and $p(\Delta|H_0)$, for different SNRs. Fitted distributions for $p(\Delta|H_\mathrm{d})$ are shown as dotted lines.

described in Sec. 4.1. The multichannel test signals contained a speaker at the desired location $\mathbf{r}_\mathrm{d}$, speakers at other undesired locations $\mathbf{r} \neq \mathbf{r}_\mathrm{d}$, and ambient noise with three different signal-to-noise ratios (SNRs): *low*, *mid*, and *high*. Each time-frequency bin was classified as target speech (H_d) or non-target sound (H_0) based on an ideal voice activity detector using an energy threshold on the target clean speech signal. Finally, the normalized histograms of the features (8) for the time-frequency bins in the classes H_d and H_0 were computed. The histograms during active target speech and inactive target speech are shown in Fig. 3 as solid and dashed lines, respectively for three different SNRs.

By considering the histograms in Fig. 3 as estimates of the PDFs $p(\Delta|H_\mathrm{d})$ and $p(\Delta|H_0)$, we can observe that the feature Δ during target speech activity is distributed in some kind of exponential shape, while during inactive target speech its distribution follows a raised cosine distribution. Furthermore, we can observe that the distribution $p(\Delta|H_\mathrm{d})$ depends on the SNR, while $p(\Delta|H_0)$ is almost independent of the SNR. It is worthwhile to note that the distribution of the proposed RTF based feature is independent of the frequency, which is not the case for magnitude or phase-related features as used in [13].

From the observations in Fig. 3, we propose to model the PDFs of the feature Δ during target speech activity by a flipped and shifted exponential distribution

$$p(\Delta|H_\mathrm{d}) = \beta\, e^{\beta(\Delta-1)}, \tag{9}$$

where β is the shape parameter. Strictly speaking, $p(\Delta|H_\mathrm{d})$ should be a truncated PDF as $\Delta \in [-1, 1]$, but we found that the truncation correction terms have negligible influence on the PDF shape within the range of interest, and therefore unnecessarily complicate the parameter estimation. The shape parameter β of the PDF (9) is related to the mean μ_Δ by [25]

$$\beta = (1 - \mu_\Delta)^{-1}. \tag{10}$$

If we assume that during target speech absence, the hermitian angle between estimated and target RTF is uniformly distributed, given (8) the PDF $p(\Delta|H_0)$ naturally follows a cosine shape. From this assumption and from observing Fig. 3, we model the PDF during target speech absence by the raised cosine function

$$p(\Delta|H_0) = \frac{1 + \cos(\pi\Delta)}{2}. \tag{11}$$

3.3. Target speech presence probability

Given the feature $\Delta(k, n)$, the probability that the sound wave in the current time-frequency bin originates from the target location $\mathbf{r}_\mathrm{d}$ is $P(H_\mathrm{d}|\Delta)$. Using Bayes theorem, this probability is given by

$$\begin{aligned} P(H_\mathrm{d}|\Delta) &= \frac{P(H_\mathrm{d})p(\Delta|H_\mathrm{d})}{P(H_0)p(\Delta|H_0) + P(H_\mathrm{d})p(\Delta|H_\mathrm{d})} \\ &= \frac{\Lambda(\Delta)}{\frac{P(H_0)}{P(H_\mathrm{d})} + \Lambda(\Delta)}, \end{aligned} \tag{12}$$

where the log-likelihood ratio is given by

$$\Lambda(\Delta) = \frac{p(\Delta|H_\mathrm{d})}{p(\Delta|H_0)}, \tag{13}$$

and $P(H_\mathrm{d})$ and $P(H_0)$ are the *a priori* probabilities that speech is present or absent, respectively. We assume an equal *a priori* probability ratio of $\frac{P(H_0)}{P(H_\mathrm{d})} = 1$. Given the PDF models (9), (11), the *a priori* probability ratio, and the speech shape parameter β, we can compute the probability (12). In the next section, we show how to estimate the speech shape parameter from the observation.

To obtain more reliable probability estimates, the final probability $\widehat{P}(H_\mathrm{d}|\Delta)$ is a combination of the frequency-dependent probability $P(H_\mathrm{d}|\Delta)$ and a broadband probability $\overline{P(H_\mathrm{d}|\Delta)}$, that is obtained from the frequency averaged log-likelihood ratio $\overline{\Lambda(\Delta)}$, i. e.

$$\widehat{P}(H_\mathrm{d}|\Delta) = P(H_\mathrm{d}|\Delta)\, \overline{P(H_\mathrm{d}|\Delta)} \tag{14}$$

3.4. Estimating the speech distribution shape

As shown in Fig. 3, the exponential distribution during target speech presence $p(\Delta|H_\mathrm{d})$ depends on the SNR. Therefore, we propose to estimate the shape parameter β online from the data. Using the estimated target speech absence probability $\widehat{P}(H_0|\Delta) = 1 - \widehat{P}(H_\mathrm{d}|\Delta)$ as an adaptive update control, we can estimate the mean of $\Delta(k, n)$ during target speech activity by

$$\begin{aligned} \widehat{\mu}_\Delta(k, n) &= \alpha\, \widehat{P}(H_0|\Delta)\, \widehat{\mu}_\Delta(k, n-1) \\ &\quad + \left[1 - \alpha\, \widehat{P}(H_0|\Delta)\right] \Delta(k, n) \end{aligned} \tag{15}$$

where α is a constant smoothing parameter. We propose to estimate the shape parameter (10) as a frequency-independent value $\widehat{\beta}(n)$ per frame, using a frequency-averaged version of $\widehat{\mu}_\Delta(k, n)$ with (10).

3.5. Enhancement using spatial probability

The spatial probability $P(H_\mathrm{d}|\Delta)$ indicates, how likely the signals at the microphones originate from the desired location. Therefore, the desired signal (3) can be estimated by using the probability directly as suppression gain $W_\mathrm{P}(k, n)$ in (4) [26]. To mitigate speech distortion and artifacts, the postfilter can be limited by the minimal suppression gain $W_{\mathrm{P},\min}$, i. e.,

$$W_\mathrm{P}(k, n) = \max\{\widehat{P}(H_\mathrm{d}|\Delta),\ W_{\mathrm{P},\min}\}. \tag{16}$$

4. EVALUATION

4.1. Experimental setup

We evaluate the proposed method using a mockup device of a head-mounted display. The goal is to capture the speech of a user wearing the microphone-equipped display on his head, while suppressing noise and speech from surrounding speakers as much as possible. The positions of the $M = 5$ microphones and the desired

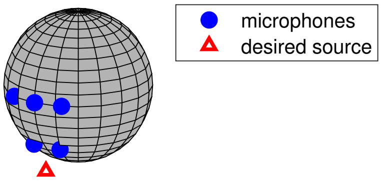

Fig. 4. Model of the microphone setup used to compute the steering vector.

source, which is the user's mouth, are shown in Fig. 4, where the grey sphere indicates the presence of the user's head. By placing the mockup device on a head and torso simulator (dummy head), the target acoustic impulse response (AIR) was measured in an anechoic room using the dummy head's mouth speaker. Additionally, AIRs for interfering speakers were measured by placing the dummy head setup in a reverberant room with about 400 ms reverberation time, and measuring the AIRs from loudspeakers at horizontal angles of $\{0, 45, 90, 135, 180, 220\}^\circ$ one meter distance from the dummy head. Using these AIRs, the microphone signals originating from user and interfering talkers were generated by convolving the AIRs with speech from an internal database. The user speech had 80 dB SPL and the interferer speech 90 dB SPL at their respective mouth positions. Ambient noise recordings, stored in the spatial Ambisonics format, were rendered to the device microphones using a full spherical set of anechoic AIRs of the device. The noise recordings were made on a busy road and in a pub, and were added with $[45, 55, 65, 75]$ dB SPL to the speech. In total, we used 144 audio files each of 3 minutes length.

For the speech enhancement processing, the audio data sampled at 16 kHz was transformed into the frequency domain using a STFT using 50% overlapping Hann windows of 32 ms length (512 samples). The smoothing time constant for RTF estimation in (5) was 25 ms, and 300 ms for the mean estimation in (15). The suppression gain limit in (16) was $W_{\mathrm{P,min}} = -10$ dB, found as the optimal tradeoff between suppression and speech distortion for the given dataset.

The steering vector to the desired location $\mathbf{b}_{\mathrm{d}}(k)$ was computed using an analytic soundfield propagation model, instead of using measured transfer functions of the actual mockup device. Firstly, using a presumably imperfect analytic model shows the robustness of the method to slight deviations of the steering vector, which may occur due to prior made assumptions on the fixed source location. Secondly, the problem of measuring the transfer functions for each individual user wearing the device is alleviated. In our case, the steering vector $\mathbf{b}_{\mathrm{d}}(k)$ was computed by assuming the microphones being placed on a rigid sphere [27] with a diameter of 26 cm as shown in Fig. 4. The microphone positions are indicated as blue dots, and the desired source position as red triangle.

4.2. Results

Due to the data generation process described in Sec. 4.1, the clean user speech as well as text annotations of the speech were available. We computed the perceptual evaluation of speech quality (PESQ) [28] and the C-weighted segmental SNR (SNR$_{\mathrm{C}}$) as perceptually motivated distance metrics between the processed audio and clean speech. Further, the processed speech was fed into a deep neural network-based online speech recognizer trained on clean speech [29], and the word error rate (WER) is reported. The WER consists of three error types: (i) *insertions* are potentially words from undesired interfering talkers; (ii) *deletions* are missed words of the desired talker; (iii) *substitutions* are incorrect recognized words of the desired talker due to noise or distortion.

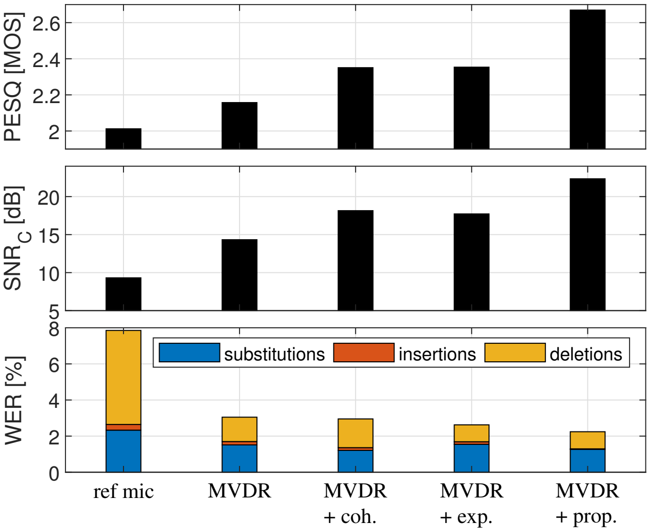

Fig. 5. Results of the MVDR followed by either coherence-based diffuse noise suppressor [30] (*coh.*), power-based expander [31] (*exp.*), or proposed spatial probability-based suppressor. (*prop.*)

Figure 5 shows PESQ, SNR$_{\mathrm{C}}$, and WER for the unprocessed reference microphone, the MVDR beamformer, and the MVDR followed by either one of two existing, or the proposed postfilter. The spatial coherence-based postfilter [30] (*coh.*) is designed to minimize the diffuse noise at the MVDR output. The postfilter proposed in [31] (*exp.*), is based on the level difference before and after the MVDR. The method can be interpreted as a kind of expander based on the signal power ratio between MVDR input and output.

We observe that all methods substantially improve PESQ, SNR$_{\mathrm{C}}$, and WER. The *coh.* postfilter improves PESQ, SNR$_{\mathrm{C}}$, but achieves no significant WER improvement over the MVDR, as it is assumes only diffuse noise, but no interfering directional talkers. While the *exp.* postfilter improves the WER of the MVDR from 3.0% to 2.6%, the proposed method reduces the WER significantly down to 2.2%. It is remarkable that in contrast to all other methods, the proposed postfilter almost eliminates the word insertions, which are mainly originating from undesired talkers, and yields the best PESQ, SNR$_{\mathrm{C}}$, and overall WER scores.

5. CONCLUSION

We proposed a spectral suppression filter designed to suppress directional and non-directional interference. The proposed approach assumes prior knowledge of the RTF related to the target sound source location, and consequently takes magnitude and phase information into account. The proposed postfilter is derived using a general formulation to extract nearfield and farfield target sources, yields low computational complexity, and does require a DOA estimator. We showed in experiments aiming to extract the nearfield speech of a head-mounted display user, that the proposed suppressor improves the perceptual quality, SNR and WER in the presence of noise and directional interfering talkers. In contrast to existing approaches, especially directional interference can be suppressed more efficiently.

6. REFERENCES

[1] H. L. van Trees, *Optimum Array Processing*, Detection, Estimation and Modulation Theory. Wiley, 2002.

[2] 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.

[3] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

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

[5] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2159–2169, Sep. 2011.

[6] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Am.*, vol. 62, no. 4, pp. 912–915, 1977.

[7] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[8] T. Wolff and M. Buck, "A generalized view on microphone array postfilters," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Tel Aviv, Israel, Aug 2010.

[9] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 1006–1018, June 2015.

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

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

[12] Y. A. Huang, A. Luebs, J. Skoglund, and W. B. Kleijn, "Globally optimized least-squares post-filtering for microphone array speech enhancement," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 380–384.

[13] Ivan Tashev, Michael Seltzer, and Alex Acero, "Microphone array for headset with spatial noise suppressor," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Eindhoven, The Netherlands, 2005.

[14] A. Sugiyama and R. Miyahara, "A directional noise suppressor with a specified beamwidth," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 524–528.

[15] S. M. Kim and H. K. Kim, "Direction-of-arrival based SNR estimation for dual-microphone speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2207–2217, Dec 2014.

[16] J. Rosca, R. Balan, N. P. Fan, C. Beaugeant, and V. Gilg, "Multichannel voice detection in adverse environments," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Sept 2002, pp. 1–4.

[17] O. Thiergart, M. Taseska, and E.A.P. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2182–2196, Dec. 2014.

[18] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[19] S. Braun, W. Zhou, and E. A. P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2015.

[20] N. Ito, S. Araki, M. Delcroix, and T. Nakatani, "Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 681–685.

[21] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, Sept. 2004.

[22] M. Schwab, P. Noll, and T. Sikora, "Noise robust relative transfer function estimation," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Florence, Italy, Sept 2006, pp. 1–5.

[23] R. Varzandeh, M. Taseska, and E. A. P. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *Proc. Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, San Francisco, CA, USA, March 2017, pp. 11–15.

[24] J. L. Coolidge, "Hermitian metrics," *Annals of Mathematics*, vol. 22, no. 1, 1920.

[25] R. S. Tsay, *Analysis of Financial Time Series*, Probability and Statistics. Wiles, New York, USA, 2002.

[26] D. Middleton and R. Esposito, "Simultaneous optimum detection and estimation of signals in noise," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 434–444, May 1968.

[27] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, Academic Press, London, first edition, 1999.

[28] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Feb. 2001.

[29] Frank Seide, Gang Li, and Dong Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Interspeech 2011*. August 2011, International Speech Communication Association.

[30] A. Kuklasinski, S. Doclo, S.H. Jensen, and J. Jensen, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Lisbon, Portugal, Sept. 2014, pp. 61–65.

[31] S. Jovicic and Z. Saric, "Adaptive microphone array free of the desired speaker cancellation combined with postfilter," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 3739–3739, 2008.