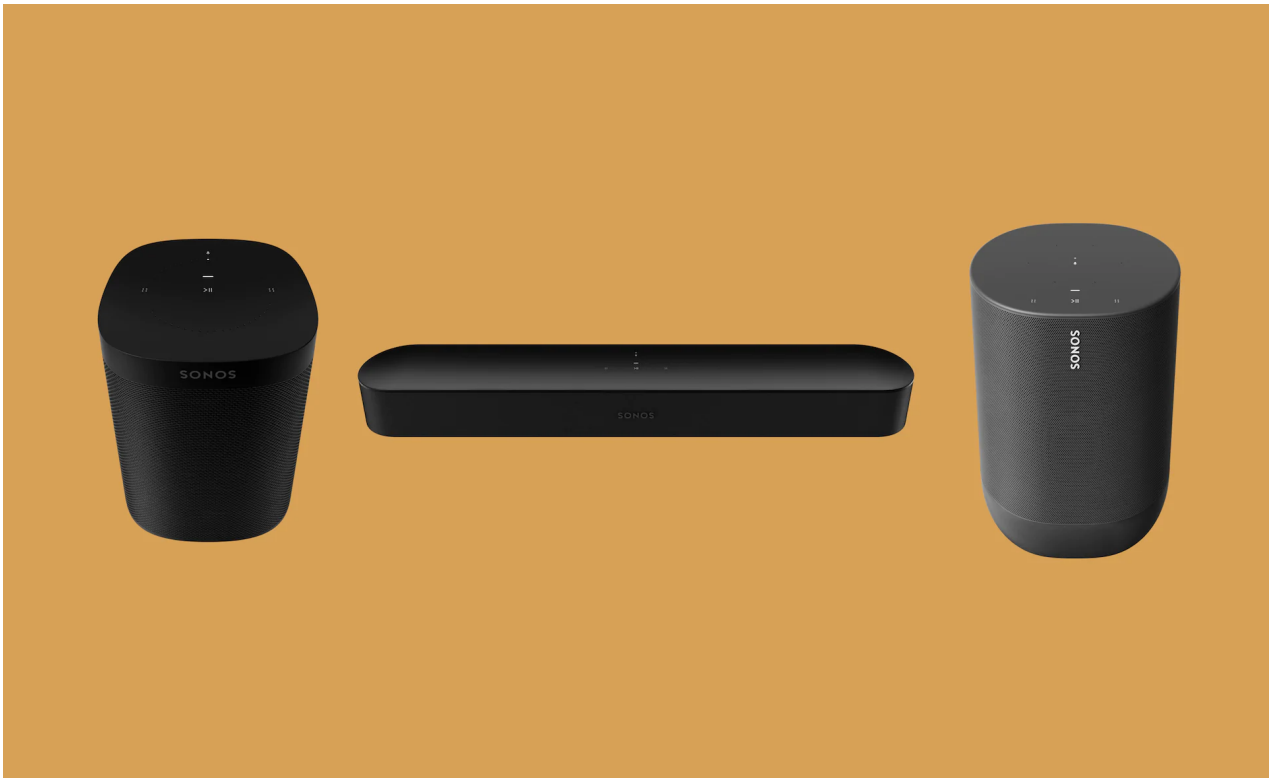


MULTI-CHANNEL NOISE REDUCTION

August 7, 2020

Noise Reduction for Distant Voice Recognition in Smart Speakers

Saeed Sereshki

Principal Audio Research Engineer, Advanced Technology Team

Daniele Giacobello

Distinguished Research Engineer, Advanced Technology Team

Why Noise Reduction?

Multi-channel noise reduction is an integral part of many modern microphone arrays systems with applications ranging from communication systems to human-machine interfaces. The recent advent of smart loudspeakers like the Amazon Echo, Google Home, and Sonos One, has pushed the robustness required in far-field noise reduction, as the user expects the same level of performance in multiple conditions, regardless of their acoustic environments. Differently from single channel approaches, these devices leverage the spatial information captured in different microphone elements to improve the ability to cancel directional and diffused noise. A well designed noise reduction method will enhance the performance of the smart speaker (wake word detection, word error rate, etc) by cleaning the speech signal.

Microphone array is one of the common components in voice-enabled Sonos smart speakers. The added spatial dimension, inherent to the array spatial aperture, results in more degrees of freedom that allows for noise reduction with low or even no speech distortion.

In our case, we are faced with a few key challenges due to the fact that our approach should work for players with different types of design constraints. On the hardware side, we expect to allow for different microphone-array geometries and configurations as the industrial design pushes for different form factors. Our players are also designed with different user experiences in mind, for example, the Sonos Move was designed as a (mostly) outdoor speaker. Furthermore, at Sonos we work with different voice partners (i.e., Google Assistant and Amazon Alexa) for which performance requirements and use cases might vary.

Taking into account all these challenges, the multi-channel noise reduction technique we employ must satisfy the following objectives:

- *A robust and scalable **far-field** multi-channel noise reduction method*
- *Easy to deploy on different devices, and different **microphone-array geometries***
- *Fast prototyping, testing, and deployment*
- *Ability to trade-off noise reduction level to optimize **device specific** performance metrics*
- *Long life-time → fault tolerant!*
- *Potential for **generalization to distributed applications** (e.g., Sonos home sound system)*

These requirements point to a noise reduction technique which does not depend on the microphone array geometry.

What Are the Options?

Given the inherent spatial nature of the multi-channel noise reduction problem, earlier approaches were greatly influenced by the traditional theory of **beamforming** that was initially

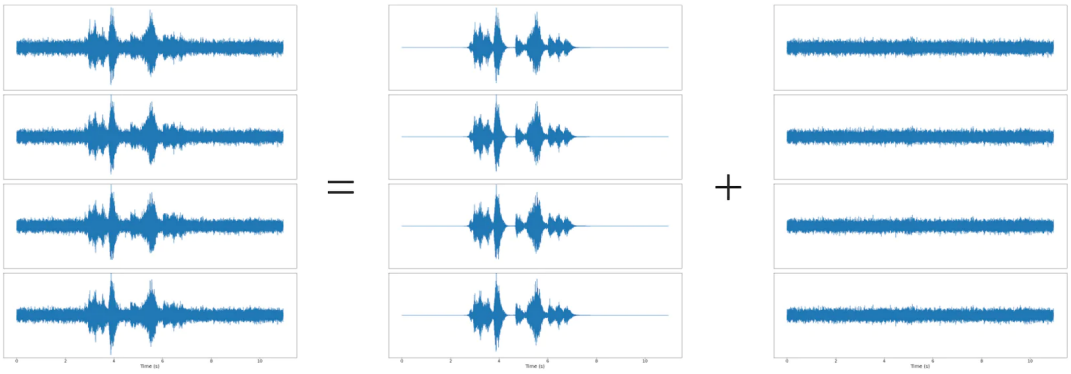
developed for sonar and radar applications using antenna arrays (beamforming). Well-known multi-channel noise reduction techniques include the delay and sum beamformer, minimum variance distortionless response (MVDR) beamformer, also known as Capon beamformer, the linearly constrained minimum variance (LCMV) beamformers, and the **generalized sidelobe canceler** (GSC) (see this page for more details on these beamformers). In all these methods, the general idea is to steer a beam toward the desired speaker while reducing the background noise coming from other directions.

The **multi-channel Wiener filter** (MWF) is another well-known multi-channel noise reduction technique, providing a **minimum mean-squared error** (MMSE) estimate of the speech component in one of the microphone signals. The literature offers several extensions to the traditional MWF. In particular, the MVDR, the GSC, and the **parametric multi-channel Wiener filter** (PMWF) can be formulated into a common frequency-domain framework where a trade-off between noise reduction and speech distortion can be achieved [1]. Furthermore, in contrast to traditional beamforming research, this formulation makes no assumptions on the geometry of the microphone array system. An interesting extension of this approach is that MWF and PMWF can also be formulated as a distributed noise reduction algorithm where the microphone arrays are part of a wireless acoustic sensor network system [2]. This makes the MWF particularly relevant in practical implementation of speech enhancement algorithms in multi-device applications where the relative geometry of multiple sets of microphones is unknown, like multi-device smart loudspeaker systems (e.g., Sonos home sound system).

The MWF and its extensions, require an accurate estimate of the noise **power spectral density** (PSD). This, in turn, requires a robust estimation of when speech is present. The **speech presence probability** (SPP), and its **multichannel** incarnation (MC-SPP), has been known to offer better performance when incorporated in the noise spectrum estimation [3]. The MC-SPP expression is established for a microphone array with arbitrary geometry under the assumption of Gaussian statistical model for speech and noise. The MC-SPP is then used to extend the single-channel noise PSD estimation to the multi-channel case.

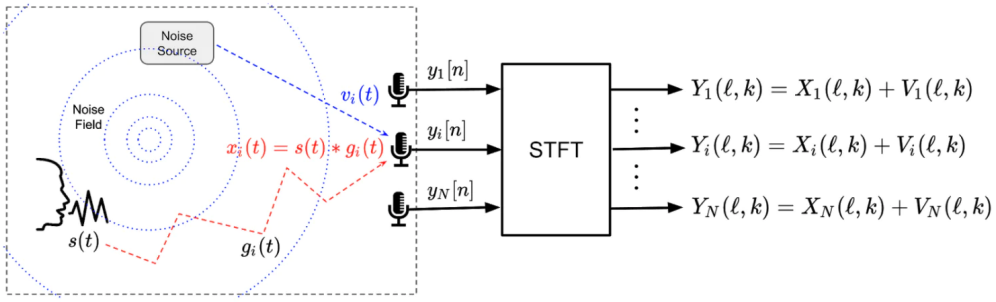
Problem Statement

An example of the signal model at the microphone array (4 microphones in Sonos Beam) is depicted in the following figure. This example is with factory noise where the signal-to-noise ratio (SNR) at the microphone is 5 dB.

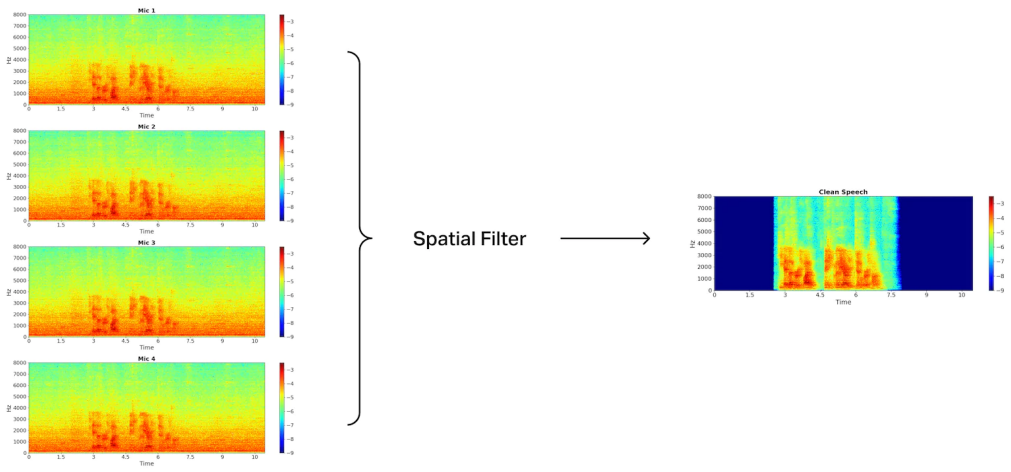


Microphone Signal | Clean Speech | Noise Signal

The spatial filtering idea is to design a filter to exploit the spatial localization to recover the clean speech from the observed microphone signals. The problem of adaptive beamforming (spatial filtering) is schematically and mathematically described in the following figure with N microphones where each microphone input of the array includes an additive mixture of reverberant speech component (or desired signal), and noise. The noise can represent multiple competing point sources or a spatially incoherent noise. The microphone input is transformed using **Short-Time Fourier Transform** (STFT) which is a standard approach in speech processing and enhancement. The analysis in STFT enables us to process the signal in smaller time segments and independently in different frequency bands. The time-domain signal is first broken into overlapping frames (typically 10-20ms). Each frame is multiplied by a windowing function and then the FFT of the windowed signal is provided at the output. In the figure below, ℓ denotes the frame index and k denotes the index of frequency bin. The number of frequency bins depends on the FFT size. In the following, the derivation will focus on a single frequency bin k and frame ℓ. However, all the expressions are used for all the frequency bins.

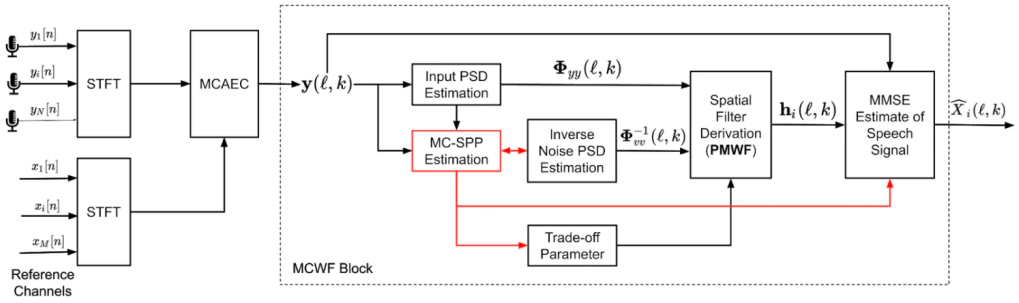


In the above figure, $g_i(t)$ represents the **room impulse response** (RIR). The objective in adaptive beamforming is to apply a spatial linear filter to the microphone inputs in the STFT domain to extract the enhanced speech signal. The input, and clean speech in the STFT domain is shown in the figure below for the same example shown above.



Solution

The block diagram of our practical implementation of PMWF, presented in [5, 6, 7], which incorporates an estimate of the multi-channel speech presence probability (MC-SPP) is depicted in the following figure. In the following, we describe the blocks in this figure in more detail and share the ideas that contribute to the development and implementation of this solution.



We first need to explain how the spatial linear filter is derived. The multi-channel Wiener filter (MCWF) is a linear filter that attempts to enhance the output **signal-to-noise ratio** (SNR) by reducing noise, utilizing the microphone array's input observation. The objective is to reduce the noise and recover the target speech signal in some optimal way (by solving a constraint optimization) where a linear filter is applied to the observation vector. The constraint optimization problem is formed to maximize the local noise reduction factor while limiting the maximum allowable speech distortion below a frequency-dependent threshold. This trade-off between noise reduction and speech distortion naturally satisfies one of our key objectives. In practice, we utilize this capability to tune our method for different players and partners.

The linear filter is derived by solving this optimization problem. In [1], the closed-form solution for the linear spatial filter with respect to microphone i is derived as

$$\mathbf{h}_i(\ell,k) = \frac{\mathbf{\Phi}_{vv}^{-1}(\ell,k)\mathbf{\Phi}_{yy}(\ell,k) - \mathbf{I}_N}{\beta(\ell,k) + \xi(\ell,k)}\mathbf{u}_i.$$

The expression for the optimal linear filter depends on the noise and noisy input PSD matrices $\mathbf{\Phi}_{vv}(\ell,k)$ and $\mathbf{\Phi}_{yy}(\ell,k)$. In this expression, $\xi(\ell,k) \triangleq \mathrm{tr}\{\mathbf{\Phi}_{vv}^{-1}(\ell,k)\mathbf{\Phi}_{yy}(\ell,k)\} - N$ denotes the multi-channel *a priori* signal-to-noise ratio (SNR). $\mathrm{tr}\{\cdot\}$ denotes the trace operator (sum of the elements on the main diagonal of a matrix) and $\beta(\ell,k)$ (positive valued) is a time-frequency dependent factor that allows for tuning the signal distortion and noise reduction at the output of the beamformer. PMWF is uniquely based on second-order statistics, and specifically it only depends on the estimation of the noisy input and the noise PSD matrices. Typically, an averaging time window of 2-3 seconds is used to achieve a reliable estimate of the PSD matrices. This suggests that the noise reduction performance of the PMWF depends on the long-term average of the spectra/spatial characteristics of the speech and the noise sources. In practice, this means that the PMWF works well if the long-term spectral and/or spatial characteristics of the speech and the noise are slowly time-varying.

How to Estimate the PSD Matrices?

As it is evident from the final expression of PMWF, the key to obtain the linear filter is to estimate the PSD matrices. The accuracy of these estimates play a crucial role in the quality of the filter and its final performance. The estimation of the noisy input PSD is relatively straightforward. We use the typical **first order smoothing** (exponential smoothing) to approximate the mathematical expectation and estimate the PSD matrix. The only parameter of importance is the smoothing coefficient α_y which needs to be tuned properly.

The following expression is used to update the input PSD matrix $\widehat{\mathbf{\Phi}}_{yy}(\ell,k) = \alpha_y \widehat{\mathbf{\Phi}}_{yy}(\ell-1,k) + (1-\alpha_y)\mathbf{y}(\ell,k)\mathbf{y}^H(\ell,k)$. For noise PSD matrix estimation, we need to take into account the speech presence uncertainty. The key idea is to estimate the speech presence probability and use it to adapt the exponential smoothing technique. Assuming that the speech and noise

signals are modelled as complex multivariate Gaussian random variables, the MC-SPP is estimated as follows [4]

$$p(\ell,k) \triangleq \left\{1+\frac{q(\ell,k)}{1-q(\ell,k)}[1+\xi(\ell,k)]\exp\left[-\frac{\gamma(\ell,k)}{1+\xi(\ell,k)}\right]\right\}^{-1},$$

where, $\gamma(\ell,k) \triangleq \mathbf{y}^H(\ell,k)[\mathbf{\Phi}_{vv}^{-1}(\ell,k)\mathbf{\Phi}_{yy}(\ell,k)\mathbf{\Phi}_{vv}^{-1}(\ell,k) - \mathbf{\Phi}_{vv}^{-1}(\ell,k)]\mathbf{y}(\ell,k)$ and $q(\ell,k)$ is the *a priori* speech absence probability. In our implementation, we have used $q(\ell,k) = 0.5$.

The speech presence probability can be further improved by post-processing. A simple approach is to apply exponential smoothing. In addition, in order to avoid stagnation, we make sure that the probabilities do not suck at 0 or 1. So, eventually we have $\overline{p}(\ell,k) = f(\,p(\ell,k))$. Using the speech presence probability, we follow the same approach we took to estimate $\mathbf{\Phi}_{yy}(\ell,k)$. We use recursive averaging with a smoothing parameter α_v when noise is present by including the new observation vector $\mathbf{y}$ in the averaging, while we do not change the estimate when speech is present. Employing this technique, the noise PSD matrix can be estimated as

$$\widehat{\mathbf{\Phi}}_{vv}(\ell,k) = \widetilde{\alpha}_v(\ell,k)\widehat{\mathbf{\Phi}}_{vv}(\ell-1,k) + (1-\widetilde{\alpha}_v(\ell,k))\mathbf{y}(\ell,k)\mathbf{y}^H(\ell,k),$$

where the time-frequency dependent smoothing coefficient depends on the speech presence probability as $\widetilde{\alpha}_v(\ell,k) \triangleq \alpha_v + (1-\alpha_v)\overline{p}(\ell,k)$.

Both expressions for MC-SPP and linear spatial filter require the inverse of the noise PSD matrix. This means that we do not need to explicitly compute and store the noise PSD matrix. Typically, the calculation of matrix inverse is computationally very prohibitive. We note that the update expression for $\widehat{\mathbf{\Phi}}_{vv}(\ell,k)$ includes a rank-1 update in each iteration. As a result, we can use **Woodbury matrix identity** (or Sherman-Morrison formula). This approach directly updates the inverse of the noise PSD matrix which reduces the computational complexity of the algorithm in a practical implementation.

Trade-off Parameter

The spatial filter also requires the trade-off parameter $\beta(\ell,k)$. In the special case that $\beta(\ell,k) = 0$, the proposed filter reduces to the well-known MVDR filter. If $\beta(\ell,k) = 1$, the filter is equivalent to MCWF. The advantage of a tunable $\beta(\ell,k)$ is the ability to control the noise reduction and speech distortion based on a high level performance metric of interest. The goal is to provide a flexible solution which can be tuned for specific applications and devices. We propose to use a MC-SPP controlled version of PMWF. The SPP-controlled PMWF outperforms the traditional MWF that uses a fixed trade-off parameter in terms of noise reduction and speech distortion. The idea is to use small trade-off values when MC-SPP is high to reduce speech distortion, and use larger trade-off values when MC-SPP is low to increase the noise reduction. Putting all these techniques together, we can now compute the PMWF spatial filter.

Extracting the Enhanced Speech Signal

Once the spatial filter is derived, the output is computed as $\mathbf{h}_i^H(\ell,k)\,\mathbf{y}(\ell,k)$. The raw PMWF output is further improved by MMSE estimate of the desired speech signal where the speech presence probability estimation is utilized. The MMSE estimate of the desired speech signal can be obtained according to

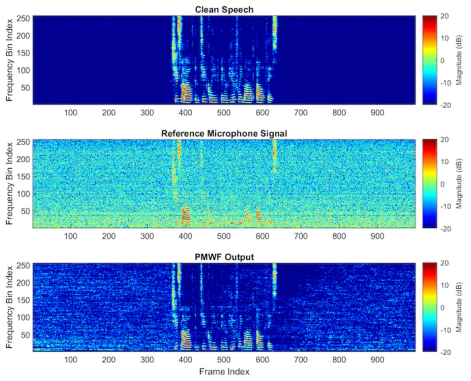
$$\widehat{X}_i(\ell,k) = \overline{p}(\ell,k)\mathbf{h}_i^H(\ell,k)\mathbf{y}(\ell,k) + (1-\overline{p}(\ell,k))G_{\min}Y_i(\ell,k),$$

where the gain factor $G_{\min}$ determines the maximum amount of noise suppression when speech is not present, $\overline{p}(\ell,k) = 0$. The inclusion of minimum gain in the expression reduces the speech distortion caused due to the estimation error of MC-SPP. The parameter $G_{\min}$ can be tuned to optimize the performance metric of interest (e.g. word error rate in automatic speech recognition systems, wake-word detection rate) in the speech acquisition system.

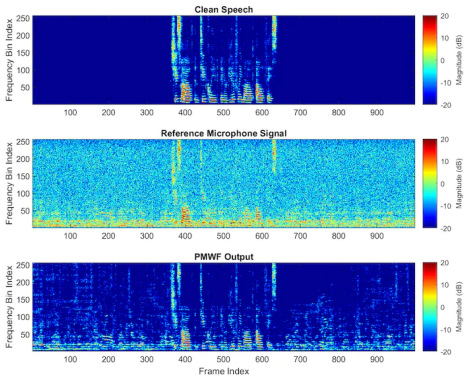
Conclusion

We showed how the MC-SPP affects different aspects of a traditional PMWF formulation, i.e., the estimation of the noise PSD matrix, the control of the trade-off between noise reduction and speech distortion, and the estimate of desired speech signal at the output of the PMWF. The proposed method outperforms traditional beamforming techniques in terms of SINR improvement and speech distortion factor (see [5] for the detailed simulation results).

The following two figures show two examples of the method in action with SNR = 0 dB, arguably one of the most challenging scenarios.



SNR = 0dB, noise type = pink noise



SNR = 0dB, noise type = babble noise

References

[1] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction". IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 2, pp. 260–276, 2010.

[2] A. Bertrand, and M. Moonen. "Distributed adaptive node-specific signal estimation in fully connected sensor networks—Part I: Sequential node updating". IEEE Transactions on Signal Processing, vol. 58, no. 10, pp. 5277-5291, 2010.

[3] T. Gerkmann and R. C. Hendriks. "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay". IEEE Transactions on Audio, Speech, and Language

Processing, vol. 20, no. 4 pp. 1383-1393, 2011.

[4] M. Souden, J. Chen, J. Benesty, and S. Affes, “Gaussian model-based multichannel speech presence probability”. IEEE Transactions on Audio, Speech, and Language Processing , vol. 18, no. 5, pp. 1072–1077, 2010.

[5] S. Bagheri, and D. Giacobello, "Exploiting Multi-Channel Speech Presence Probability in Parametric Multi-Channel Wiener Filter". INTERSPEECH, pp. 101-105, 2019.

[6] S. Bagheri Sereshki, and D. Giacobello. "Linear Filtering for Noise-Suppressed Speech Detection". U.S. Patent Application No. 15/984,073, 2017.

[7] S. Bagheri Sereshki, and D. Giacobello “Linear Filtering for Noise-Suppressed Speech Detection Via Multiple Network Microphone Devices”. US Patent 10,692,518, 2020.