Exploiting Multi-Channel Speech Presence Probability in Parametric Multi-Channel Wiener Filter

Saeed Bagheri Daniele Giacobello

SONOS

Introduction and Motivations

- Sonos voice enabled smart speakers

Sonos One

Sonos Beam

Sonos Move

SONOS

INTERSPEECH 2019

Introduction and Motivations

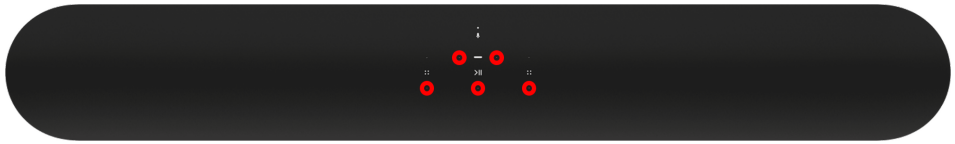
- Sonos voice enabled smart speakers

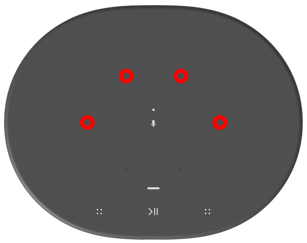
Sonos One

Sonos Beam

Sonos Move

- **Speech enhancement method using microphone arrays?**





SONOS

INTERSPEECH 2019

Introduction and Motivations

- Sonos voice enabled smart speakers

Sonos One

Sonos Beam

Sonos Move

- **Challenges**:
  - Different microphone-array geometries and configurations
  - Different industrial design, form factors, and HW modules
  - Different performance requirements and use cases

SONOS

INTERSPEECH 2019

Introduction and Motivations

- Sonos voice enabled smart speakers

Sonos One

Sonos Beam

Sonos Move

- **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
  - Trade-off noise reduction level to optimize device specific performance metrics
  - Long life-time $\longrightarrow$ fault tolerant!
  - Generalization to distributed applications (e.g., Sonos home sound system) [Doclo et al., 2009]

SONOS

INTERSPEECH 2019

Relevant Work

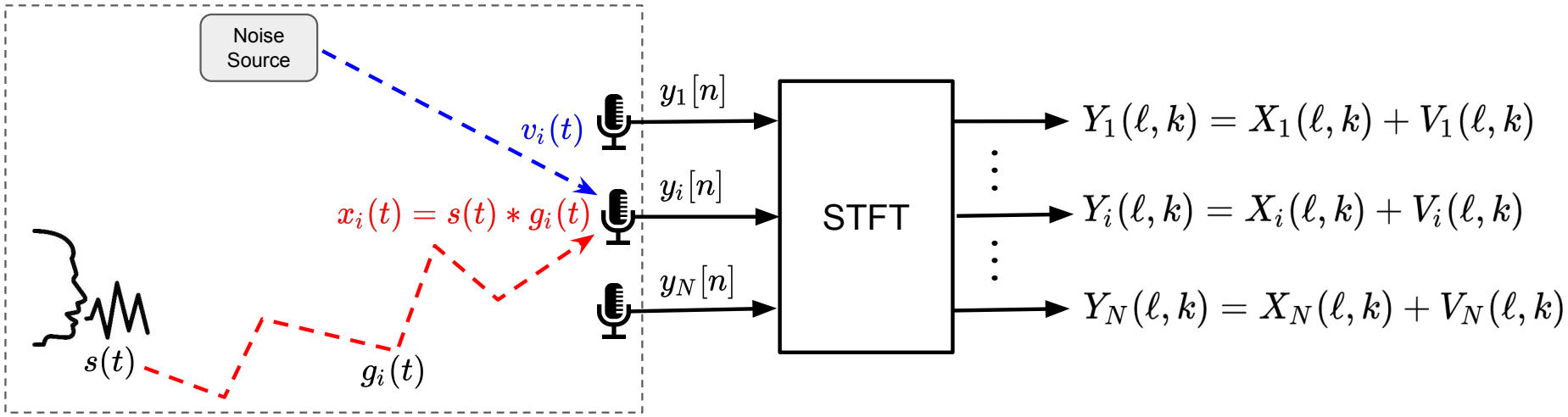
- Adaptive multi-channel noise reduction techniques
  - Frost beamformer [Frost, 1972]
  - Linearly constrained minimum variance (LCMV) beamformers [Er and Cantoni, 1983; Darlington, 1958]
  - Minimum variance distortion-less response (MVDR) beamformer [Capon, 1969]
  - Generalized side-lobe canceler (GSC) [Griffiths and Jim, 1982]
  - Multi-channel Wiener filter (MWF) [Benesty et al., 2008]
- A **common frequency-domain framework** is proposed in [Souden et al., 2010a]
  - MVDR, GSC, and the parametric multi-channel Wiener filter (**PMWF**) are formulated in the same framework
  - Trade-off between noise reduction and speech distortion
  - No assumptions on the geometry of the microphone array

SONOS

INTERSPEECH 2019

Problem Definition

- **Observation Model** (uncorrelated noise and speech signals)



- **Objective**: Apply a **spatial linear filter** to estimate the speech signal

$$\widehat{X}_i(\ell,k) = \mathbf{h}_i^H(\ell,k) \quad \underbrace{\mathbf{y}(\ell,k)}$$

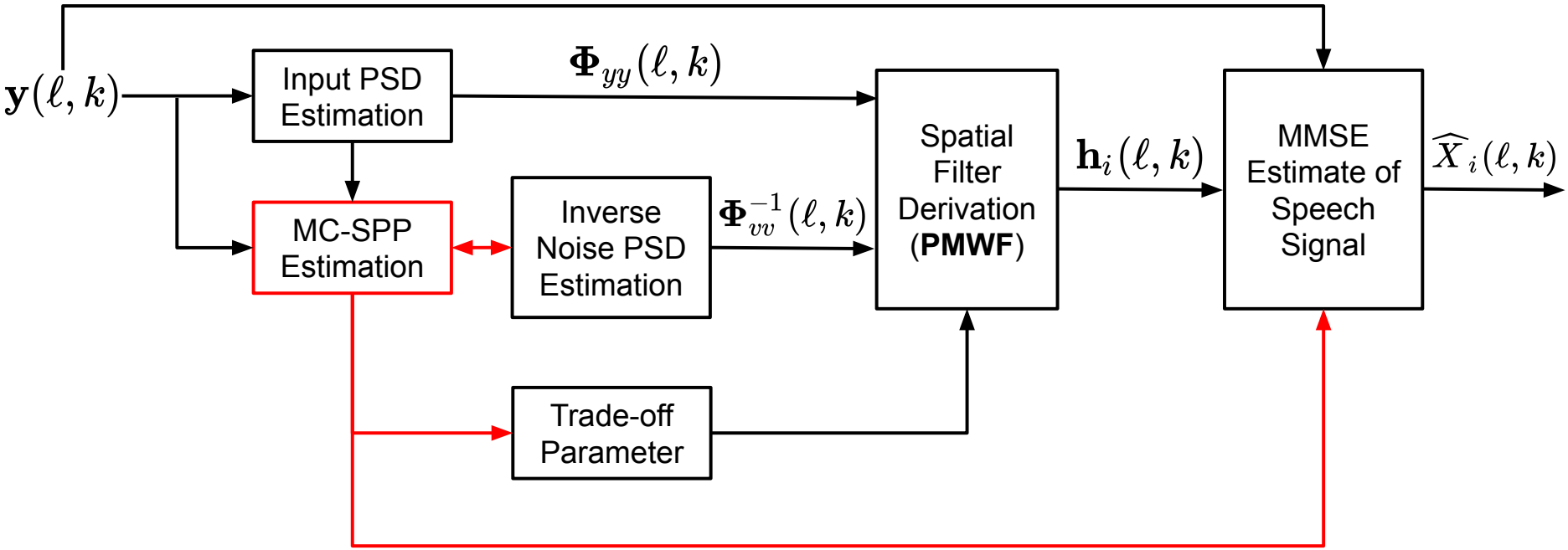
$$\triangleq \begin{bmatrix} Y_1(\ell,k) \\ \vdots \\ Y_N(\ell,k) \end{bmatrix}$$

SONOS

INTERSPEECH 2019

Our Contributions

- A practical implementation of PMWF which incorporates an estimate of the multi-channel speech presence probability (MC-SPP)



Review: Parametric Multi-Channel Wiener Filter

Constraint optimization formulation in frequency domain [Souden et al., 2010a]

$$\underset{\mathbf{h}_i(\ell,k)}{\arg\max} \quad \overbrace{\zeta_{nr}(\mathbf{h}_i(\ell,k))}^{\text{local noise reduction factor}}$$

$$\text{subject to} \quad \underbrace{\nu_{sd}(\mathbf{h}_i(\ell,k))}_{\text{local speech distortion}} \leq \sigma^2(\ell,k)$$

- $\zeta_{nr}(\mathbf{h}_i(\ell,k))$ is a function of noise PSD matrix $\mathbf{\Phi}_{vv}(\ell,k)$
- $\nu_{sd}(\mathbf{h}_i(\ell,k))$ is a function of speech PSD matrix $\mathbf{\Phi}_{xx}(\ell,k)$

SONOS

INTERSPEECH 2019

Review: Parametric Multi-Channel Wiener Filter

PMWF solution [Souden et al., 2010a; Doclo and Moonen, 2002; Spriet et al., 2004]

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

- Depends on the input and noise PSD matrices
- $\xi(\ell,k) \triangleq \text{tr}\Big\{\mathbf{\Phi}_{vv}^{-1}(\ell,k)\mathbf{\Phi}_{yy}(\ell,k)\Big\} - N \longrightarrow$ Multi-Channel *a priori* SNR
- $\beta(\ell,k)$: trade-off parameter $\longrightarrow$ Inverse of the Lagrange multiplier
  - $\beta = 0 \Longrightarrow$ MVDR
  - $\beta = 1 \Longrightarrow$ MCWF

SONOS

INTERSPEECH 2019

MC-SPP Estimation

Speech and noise are modeled as complex multivariate Gaussian random variables

- MC-SPP expression [Souden et al., 2010b]

$$\underbrace{p(\ell,k)}_{\text{MC-SPP}} = \left\{1 + \frac{q(\ell,k)}{1-q(\ell,k)}[1+\xi(\ell,k)]\exp\left[-\frac{\gamma(\ell,k)}{1+\xi(\ell,k)}\right]\right\}^{-1}$$

- $q(\ell,k)$: the *a priori* speech absence probability [Cohen, 2003; Souden et al., 2011]
- $\xi(\ell,k)$: Multi-Channel *a priori* SNR
- $\gamma(\ell,k) \triangleq \mathbf{y}^H(\ell,k)\left[\mathbf{\Phi}_{vv}^{-1}(\ell,k)\mathbf{\Phi}_{yy}(\ell,k)\mathbf{\Phi}_{vv}^{-1}(\ell,k) - \mathbf{\Phi}_{vv}^{-1}(\ell,k)\right]\mathbf{y}(\ell,k)$

SONOS

INTERSPEECH 2019

MC-SPP Estimation

Speech and noise are modeled as complex multivariate Gaussian random variables

- MC-SPP expression [Souden et al., 2010b]

$$\underbrace{p(\ell,k)}_{\text{MC-SPP}} = \left\{1 + \frac{q(\ell,k)}{1-q(\ell,k)}[1+\xi(\ell,k)]\exp\left[-\frac{\gamma(\ell,k)}{1+\xi(\ell,k)}\right]\right\}^{-1}$$

- $q(\ell,k)$: the *a priori* speech absence probability [Cohen, 2003; Souden et al., 2011]
- $\xi(\ell,k)$: Multi-Channel *a priori* SNR
- $\gamma(\ell,k) \triangleq \mathbf{y}^H(\ell,k)\left[\mathbf{\Phi}_{vv}^{-1}(\ell,k)\mathbf{\Phi}_{yy}(\ell,k)\mathbf{\Phi}_{vv}^{-1}(\ell,k) - \mathbf{\Phi}_{vv}^{-1}(\ell,k)\right]\mathbf{y}(\ell,k)$
- **Smoothing**:
  - $\overline{p}(\ell,k) = \alpha_p\,\overline{p}(\ell-1,k) + (1-\alpha_p)\,p(\ell,k)$
  - $\overline{p}(\ell,k) = \min\left\{\max\left\{\overline{p}(\ell,k),\, p_{\text{min}}\right\},\, p_{\text{max}}\right\}$

SONOS

INTERSPEECH 2019

Utilization of MC-SPP - Part 1

- **Estimation of PSD Matrices**
  - Input PSD matrix: $\widehat{\mathbf{\Phi}}_{yy}(\ell,k) = \alpha_y\, \widehat{\mathbf{\Phi}}_{yy}(\ell-1,k) + (1-\alpha_y)\, \mathbf{y}(\ell,k)\mathbf{y}^H(\ell,k)$
  - Noise PSD matrix: generalization of the IMCRA approach [Cohen, 2003]

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

    $\rightarrow \widetilde{\alpha}_v(\ell,k) \triangleq \alpha_v + (1-\alpha_v)\underbrace{\overline{p}(\ell,k)}_{\text{MC-SPP}}$

Utilization of MC-SPP - Part 1

- **Estimation of PSD Matrices**
  - Input PSD matrix: $\widehat{\mathbf{\Phi}}_{yy}(\ell,k) = \alpha_y\, \widehat{\mathbf{\Phi}}_{yy}(\ell-1,k) + (1-\alpha_y)\, \mathbf{y}(\ell,k)\mathbf{y}^H(\ell,k)$
  - Noise PSD matrix: generalization of the IMCRA approach [Cohen, 2003]

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

    $\rightarrow\ \widetilde{\alpha}_v(\ell,k) \triangleq \alpha_v + (1-\alpha_v)\underbrace{\overline{p}(\ell,k)}_{\text{MC-SPP}}$
- Both PMWF and MC-SPP need the inverse of the noise PSD matrix
  - Matrix inversion lemma

$$\widehat{\mathbf{\Phi}}_{vv}^{-1}(\ell,k) = \frac{1}{\widetilde{\alpha}_v(\ell,k)}\left[\widehat{\mathbf{\Phi}}_{vv}^{-1}(\ell-1,k) - \frac{\widetilde{\mathbf{y}}(\ell,k)\,\widetilde{\mathbf{y}}^H(\ell,k)}{g(\ell,k)}\right]$$

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

SONOS

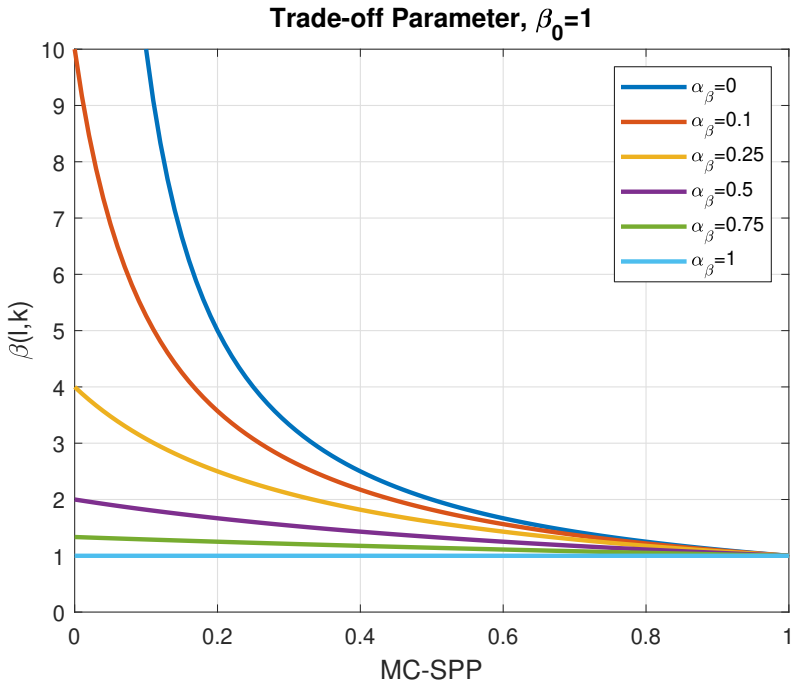
INTERSPEECH 2019

Utilization of MC-SPP - Part 2

- **MC-SPP Controlled trade-off parameter**
  - The *a posteriori* SPP has been used to control the trade-off between noise reduction and speech distortion [Ngo et al., 2009]

$$\beta(\ell, k) = \frac{\beta_0}{\alpha_\beta + (1 - \alpha_\beta)\,\beta_0\;\overline{p}(\ell, k)}$$

  - Outperforms a fixed trade-off parameter
  - Flexible for device and application specific tuning



SONOS

INTERSPEECH 2019

Utilization of MC-SPP - Part 3

- **MMSE estimate of the desired speech signal**

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

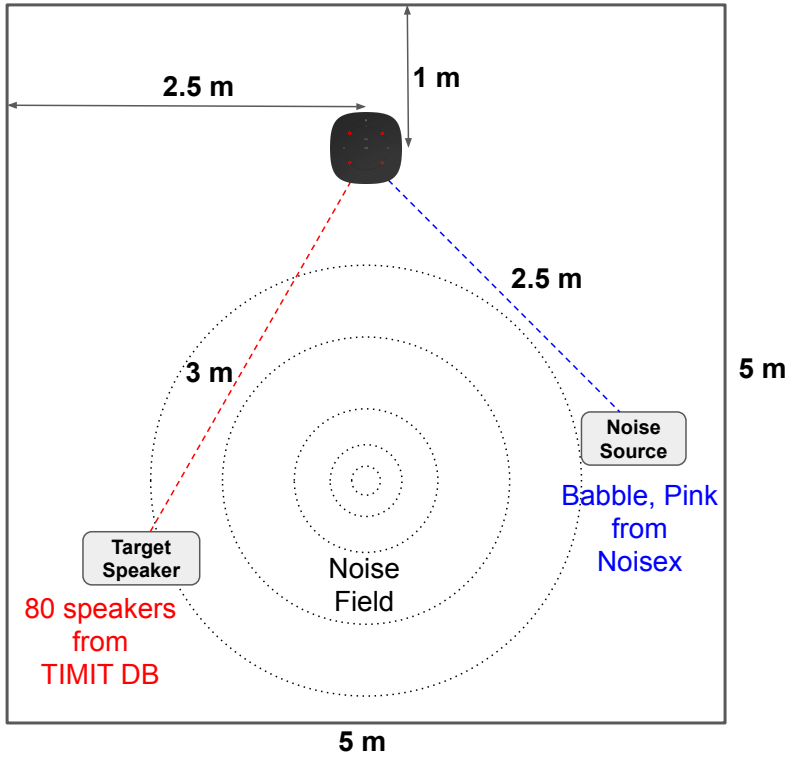
  - Reduces the speech distortion caused due to the estimation error in MC-SPP
  - G_{min} determines the maximum amount of noise suppression
  - G_{min} is tuned to optimize the performance metrics of interest (*e.g.*, word error rate in ASR systems, wake-word detection rate)

SONOS

INTERSPEECH 2019

Simulation Setup

Sampling frequency	16 KHz
Microphone array	Sonos One
# of Microphones	4
Frame length	512
Frame overlap	50%
Window function	Hann
T_{60}	300 ms
RIR generation	Image source method

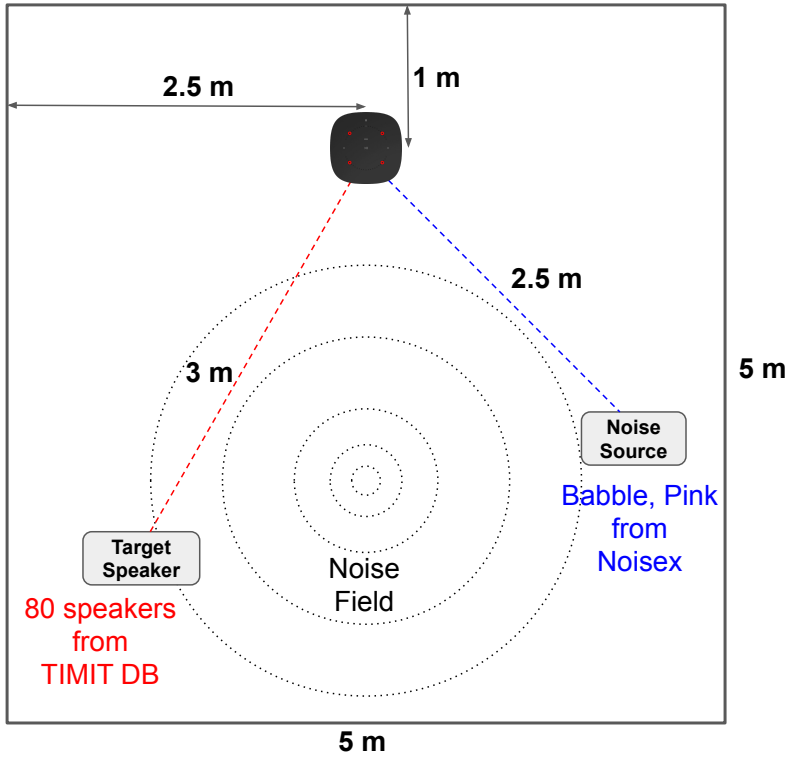


SONOS

INTERSPEECH 2019

Simulation Setup

Sampling frequency	16 KHz
Microphone array	Sonos One
# of Microphones	4
Frame length	512
Frame overlap	50%
Window function	Hann
T_{60}	300 ms
RIR generation	Image source method



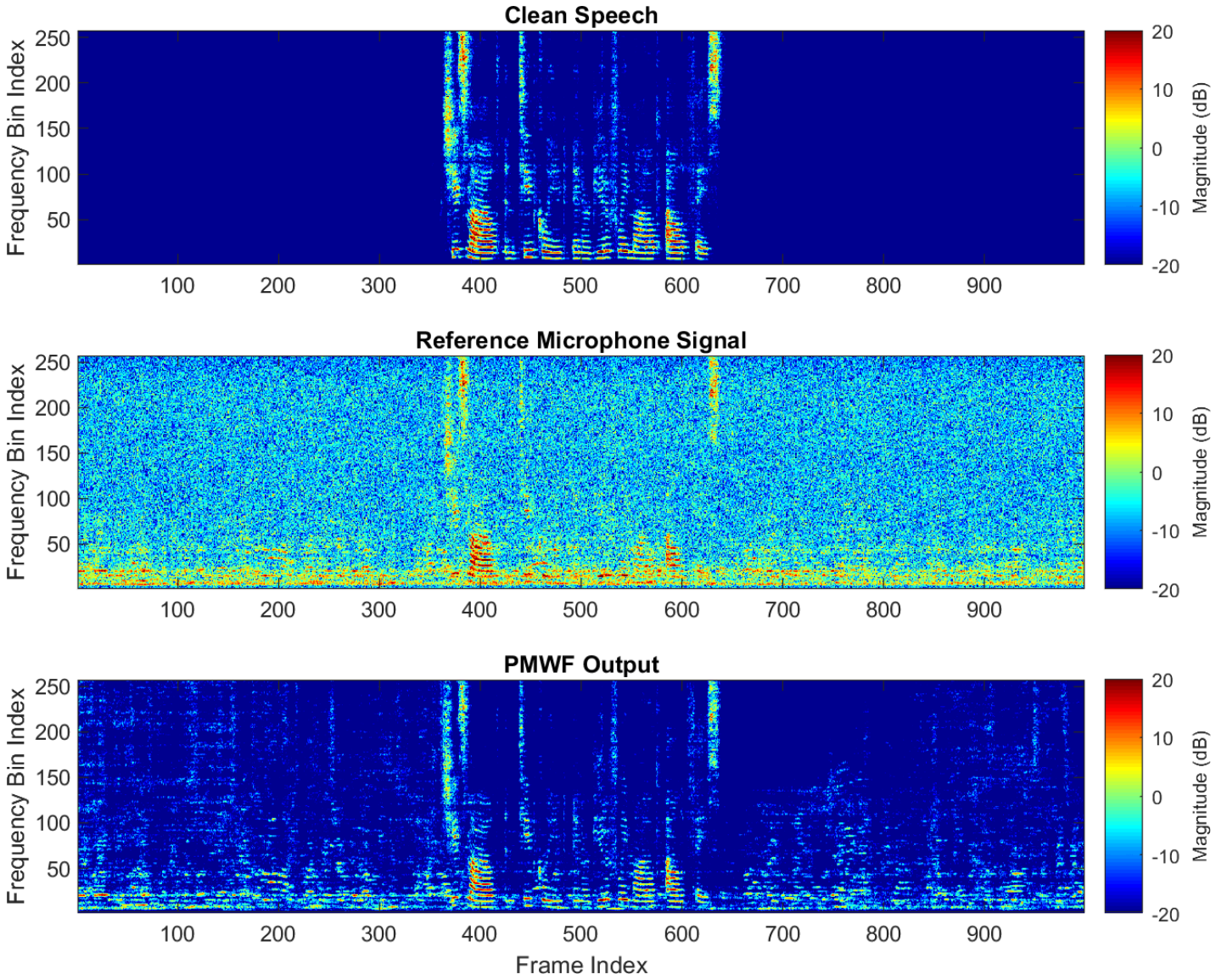
Parameters used to implement the proposed algorithm

$\alpha_v = 0.95$	$\alpha_y = 0.95$	$\alpha_p = 0.1$	$G_{\text{min}} = 0.1$
α_β varies	β_0 varies		
$p_{\text{max}} = 0.99$	$p_{\text{min}} = 0.01$	$q_0 = 0.5$	

SONOS

INTERSPEECH 2019

Example 1: Babble Noise, Input SINR $= 0$ dB



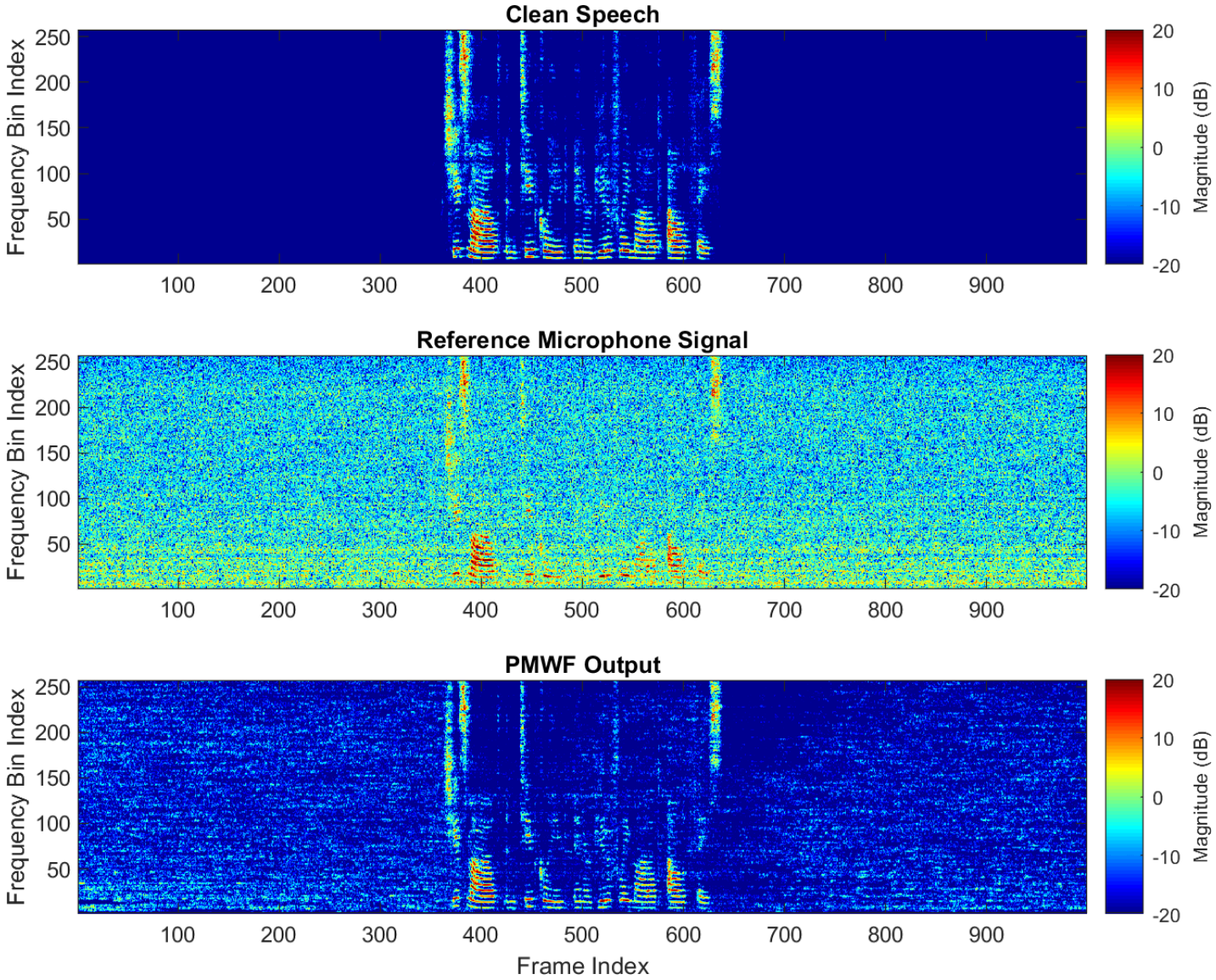
Metrics

ΔSINR	11.85 dB
NR	16.12 dB
SD	-5.82 dB

SONOS

INTERSPEECH 2019

Example 2: Pink Noise, Input SINR = 0 dB



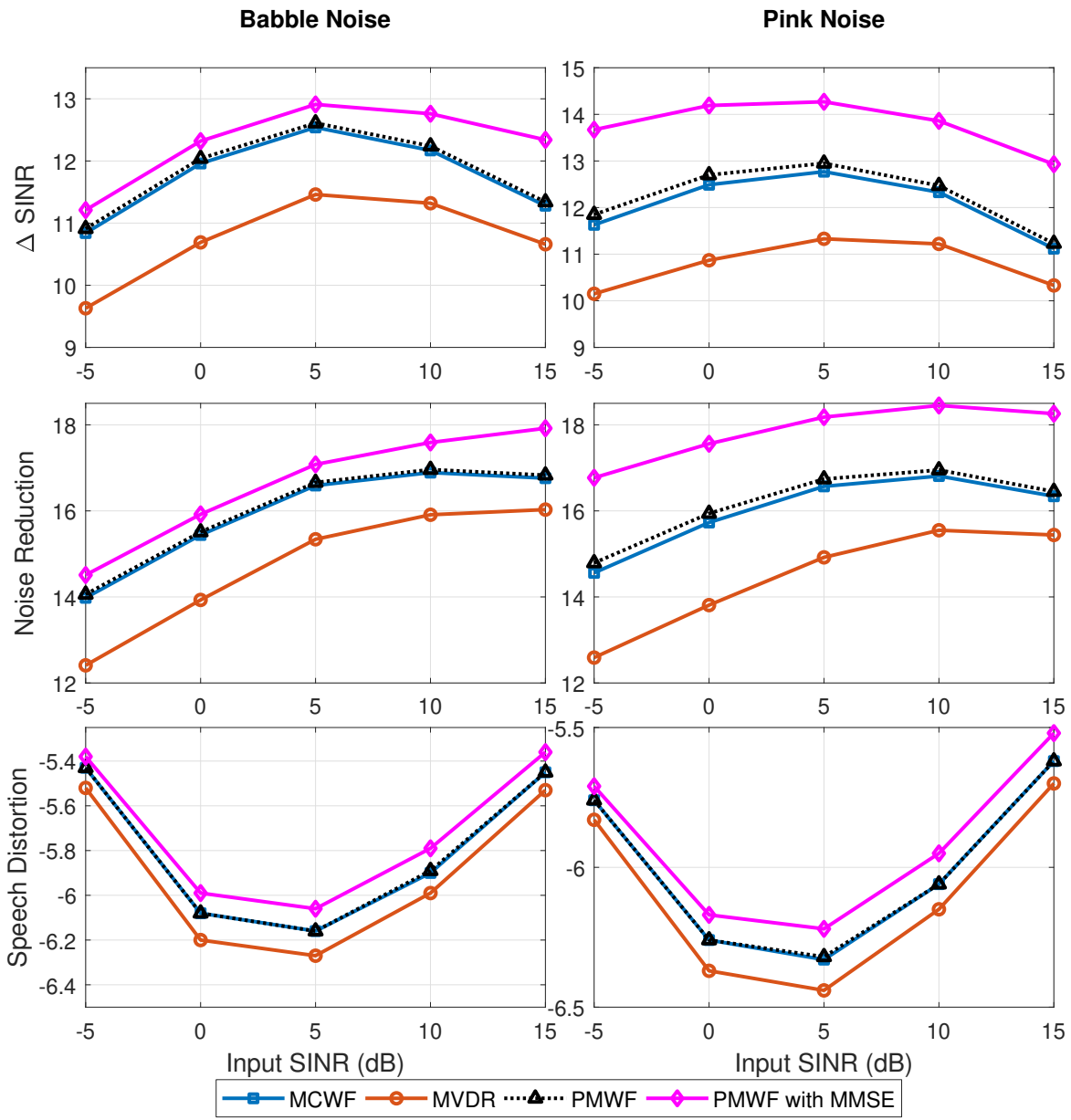
Metrics

ΔSINR	15.06 dB
NR	18.28 dB
SD	−6.49 dB

SONOS

INTERSPEECH 2019

Simulation Results



Configurations:

Test Name	β_0	α_β	MMSE estimate
MVDR	0	1	No
MCWF	1	1	No
PMWF	1	0.75	No
PMWF with MMSE	1	0.75	Yes

PMWF Performance:

- Better NR, and ΔSINR
- Minor increase in SD
- In practice, trade-off NR by tuning β_0 and α_β to optimize the performance metrics of interest (*e.g.*, word error rate in ASR systems, wake-word detection rate)

SONOS

INTERSPEECH 2019

Conclusions

- A **robust** and **scalable** far-field multi-channel noise reduction method
  - Improvement in SINR, and NR performance
  - Trade-off NR level to optimize device-specific performance metrics (PESQ, WWD, and WER)
  - Applicable to different microphone-array geometries
  - Easy to deploy on different devices after proper tuning of hyper-parameters

SONOS

INTERSPEECH 2019

References I

J. Benesty, J. Chen, and Y. Huang. *Microphone array signal processing*. Springer Science & Business Media, 2008.

J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8): 1408–1418, 1969. ISSN 0018-9219. doi: 10.1109/PROC.1969.7278.

I. Cohen. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, 11(5):466–475, 2003. ISSN 1063-6676. doi: 10.1109/TSA.2003.811544.

S. Darlington. Linear least-squares smoothing and prediction, with applications. *The Bell System Technical Journal*, 37(5):1221–1294, 1958. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1958.tb01550.x.

S. Doclo and M. Moonen. GSVD-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Transactions on Signal Processing*, 50(9):2230–2244, Sep. 2002. ISSN 1053-587X. doi: 10.1109/TSP.2002.801937.

S. Doclo, S. Doclo, M. Moonen, T. Van den Bogaert, and J. Wouters. Reduced-bandwidth and distributed mwf-based noise reduction algorithms for binaural hearing aids. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1):38–51, Jan 2009. ISSN 1558-7916. doi: 10.1109/TASL.2008.2004291.

M. Er and A. Cantoni. Derivative constraints for broad-band element space antenna array processors. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 31(6):1378–1393, 1983. ISSN 0096-3518. doi: 10.1109/TASSP.1983.1164219.

O. L. Frost. An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60(8): 926–935, 1972. ISSN 0018-9219. doi: 10.1109/PROC.1972.8817.

SONOS

INTERSPEECH 2019

References II

L. Griffiths and C. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30(1):27–34, 1982. ISSN 0018-926X. doi: 10.1109/TAP.1982.1142739.

K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen. Incorporating the conditional speech presence probability in multi-channel wiener filter based noise reduction in hearing aids. *EURASIP Journal on Advances in Signal Processing*, 2009.

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*, 18(2):260–276, 2010a. ISSN 1558-7916. doi: 10.1109/TASL.2009.2025790.

M. Souden, J. Chen, J. Benesty, and S. Affes. Gaussian model-based multichannel speech presence probability. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):1072–1077, 2010b. ISSN 1558-7916. doi: 10.1109/TASL.2009.2035150.

M. Souden, J. Chen, J. Benesty, and S. Affes. An integrated solution for online multichannel noise tracking and reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2159–2169, 2011. ISSN 1558-7916. doi: 10.1109/TASL.2011.2118205.

A. Spriet, M. Moonen, and J. Wouters. Spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction. *Signal Processing*, 84(12):2367 – 2387, 2004. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2004.07.028. URL http://www.sciencedirect.com/science/article/pii/S0165168404002002.

SONOS

INTERSPEECH 2019

Q & A

Thank You For Your Attention!

SONOS

INTERSPEECH 2019