A DNN based Normalized Time-frequency Weighted Criterion for Robust Wideband DoA Estimation

Kuan-Lin Chen[1], Ching-Hua Lee[1], Bhaskar D. Rao[1], and Harinath Garudadri[2]

[1]Department of Electrical and Computer Engineering, [2]Qualcomm Institute
University of California, San Diego

ICASSP 2023*

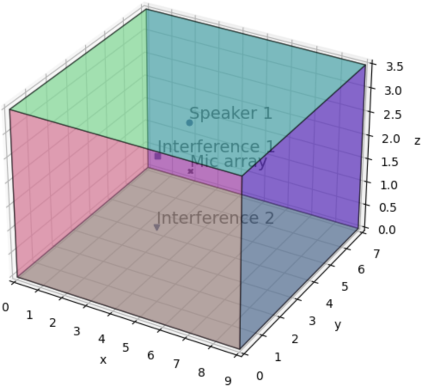
June 7, 2023

*A preprint is available at https://arxiv.org/abs/2302.10147
Code is available at https://github.com/kjason/DnnNormTimeFreq4DoA

Outline

1. Wideband direction of arrival (DoA) estimation
2. The general framework
   - The proposed DNN based normalized T-F weighted criterion
   - Post-processing methods
3. Experiments
   - Post-processing is crucial
   - The best post-processing is criterion-dependent
   - Robustness against a wide range of SIRs
4. References

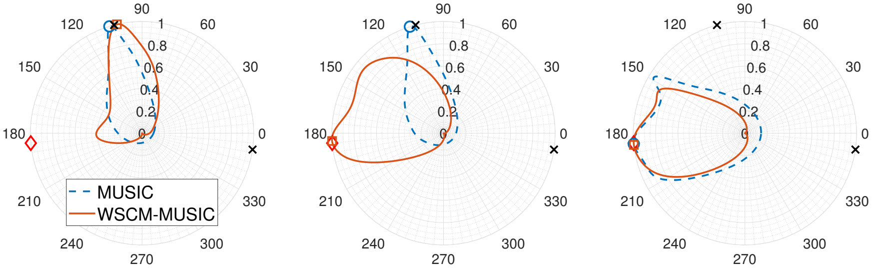
Wideband direction of arrival (DoA) estimation



- Speech source localization.
- Hearing aids and augmented hearing systems (Pisha et al., 2019).
- Many DoA estimation methods now rely on deep learning (Xu et al., 2017; Yang et al., 2017; Wang et al., 2018a; Yang et al., 2019).
- Let us focus on a simple framework using weighted spatial covariance matrices (WSCMs).

A simple approach based on a popular subspace method

- There are **one** speech source and **multiple** interference sources.
- Train a DNN to estimate the ideal ratio mask (IRM) of the speech signal.
- Compare the wideband MUSIC and the WSCM-MUSIC (Xu et al., 2017).



(a) SIR $= -6$ dB. (b) SIR $= 0$ dB. (c) SIR $= 20$ dB.

Figure: $\diamondsuit$ and $\times$ represent the speaker and interference, respectively.

- Other popular methods include the principal vector (Yang et al., 2017; Wang et al., 2018a; Yang et al., 2019) and SRP-PHAT (Pertilä and Cakir, 2017).

A framework based on time-frequency weighted criteria

- A DNN $g : \mathbb{R}^{2\times T\times F} \to \mathbb{R}^{T\times F}$ individually predicts a mask $\mathbf{G}$ for each sensor.
- For each sensor m, pick a post-processing q_m that generates T-F weights

$$\mathbf{W}_m = q_m\left(\mathbf{G}_1, \mathbf{G}_2, \cdots, \mathbf{G}_M\right). \tag{1}$$

- Compute the weighted spatial covariance matrix (WSCM)

$$\boldsymbol{\Phi}(f) = \sum_t \left[\mathbf{w}(t,f) \odot \mathbf{y}(t,f)\right]\left[\mathbf{w}(t,f) \odot \mathbf{y}(t,f)\right]^{\mathsf{H}}. \tag{2}$$

- Optimization criteria:

$$\begin{aligned}
&\text{(MUSIC)} && \max_\theta && \sum_f \frac{1}{\mathbf{v}^{\mathsf{H}}(\theta,f)\mathbf{N}(f)\mathbf{N}^{\mathsf{H}}(f)\mathbf{v}(\theta,f)},\\
&\text{(Principal vector)} && \max_\theta && \sum_f \mathbf{v}^{\mathsf{H}}(\theta,f)\mathbf{p}(f)\mathbf{p}^{\mathsf{H}}(f)\mathbf{v}(\theta,f),\\
&\text{(SRP)} && \max_\theta && \sum_f \mathbf{v}^{\mathsf{H}}(\theta,f)\boldsymbol{\Phi}(f)\mathbf{v}(\theta,f).
\end{aligned} \tag{3}$$

Why these methods are so popular?

- They basically can be applied to arbitrary array geometries.
- The DNN is independent of the microphone array used.
- Only single-channel speech and nonspeech corpora are required for training.

Question 1

Why pick a signal/noise subspace when the estimation of the IRM is accurate?

Question 2

What is the best design for T-F weights? A comparative study seems missing.

- *Binary thresholding (Heymann et al., 2016)*
- *Arithmetic mean (Pertilä and Cakir, 2017)*
- *Hadamard product (Wang et al., 2018b)*
- *And more...*

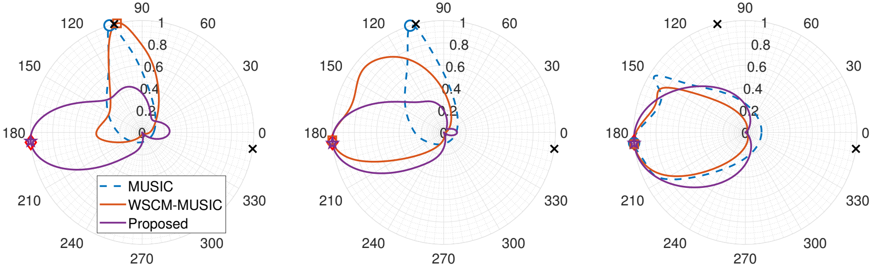
Our contributions

Contribution 1

A simple criterion yields better performance compared to commonly used methods.

Contribution 2

The post-processing that generates T-F weights is crucial and the best strategy is criterion-dependent.



(a) SIR $= -6$ dB. (b) SIR $= 0$ dB. (c) SIR $= 20$ dB.

Figure: $\diamondsuit$ and $\times$ represent the speaker and interference, respectively.

A simple criterion

- No eigenvalue decomposition.
- High-quality snapshots are preferred.
- A normalization of the magnitude of $\mathbf{y}(t, f)$ may prevent the objective function from relying on a single low SINR snapshot.
- We first normalize the filtered snapshot at every T-F bin and then directly match a candidate steering vector to the normalized filtered snapshot, i.e.,

$$\min_{\theta,\mathbf{S}} \quad \sum_f \sum_t \left\| \frac{\mathbf{w}(t,f) \odot \mathbf{y}(t,f)}{\|\mathbf{y}(t,f)\|_2} - s(t,f)\mathbf{v}(\theta,f) \right\|_2^2 . \tag{4}$$

Finding θ is equivalent to solving

$$\max_{\theta} \quad \sum_f \mathbf{v}^{\mathrm{H}}(\theta,f) \sum_t \frac{\tilde{\mathbf{y}}(t,f)\tilde{\mathbf{y}}^{\mathrm{H}}(t,f)}{\|\mathbf{y}(t,f)\|_2^2} \mathbf{v}(\theta,f). \tag{5}$$

where $\tilde{\mathbf{y}}(t,f) = \mathbf{w}(t,f) \odot \mathbf{y}(t,f)$, which is slightly different from the SRP-PHAT (Pertilä and Cakir, 2017; Zhang et al., 2008).

Post-processing that generates T-F weights

Table: Examples of the post-processing function q_m.

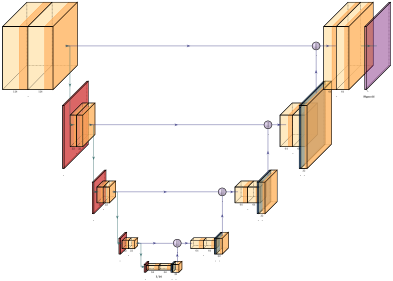
Post-processing	Expression for all $m \in [M]$
Identity (direct masking)	$q_m = \mathbf{G}_m$
Minimum	$[q_m]_{t,f} = \min_{i \in [M]} [\mathbf{G}_i]_{t,f}$
Maximum	$[q_m]_{t,f} = \max_{i \in [M]} [\mathbf{G}_i]_{t,f}$
Arithmetic mean	$q_m = \frac{1}{M} \sum_{i=1}^{M} \mathbf{G}_i$
Arithmetic median	$[q_m]_{t,f} = \text{median}(\{[\mathbf{G}_i]_{t,f}\}_{i=1}^{M})$
Hadamard product	$q_m = \mathbf{G}_1 \odot \mathbf{G}_2 \odot \cdots \odot \mathbf{G}_M$
Geometric mean	$[q_m]_{t,f} = \sqrt[M]{\prod_{i=1}^{M} [\mathbf{G}_i]_{t,f})}$
Binary thresholding (BT)	$[q_m]_{t,f} = 1$, if $[\mathbf{G}_m]_{t,f} > \beta$ $[q_m]_{t,f} = 0$, otherwise

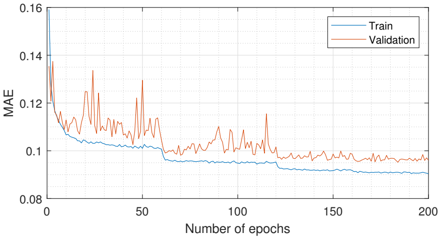
Experimental setup

- TIMIT dataset (Garofolo et al., 1993) and PNL 100 nonspeech sounds (Hu and Wang, 2010) (machine, water, wind, etc).
- Pyroomacoustics (Scheibler et al., 2018).
- Frequency bins corresponding to 50 Hz to 7 kHz are used because this is the frequency band of wideband speech coders (Cox et al., 2009).
- A 9-element rectangular microphone array.
- Simulate a dining environment.†

†Code is available at https://github.com/kjason/DnnNormTimeFreq4DoA

The DNN



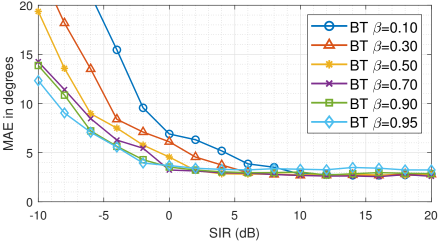


- U-Net (Ronneberger et al., 2015).‡
- Size: 0.67M parameters.
- IRM estimation. ℓ_1 loss.
- SGD with momentum. 200 epochs.§

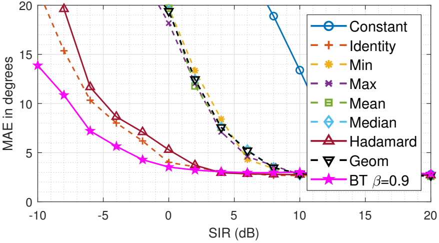
‡PlotNeuralNet https://github.com/HarisIqbal88/PlotNeuralNet
§Code is available at https://github.com/kjason/DnnNormTimeFreq4DoA

Post-processing is crucial (MUSIC)

- Different post-processing functions are evaluated for the DNN based MUSIC.
- $RT_{60} = 0.3$s and SNR $= 20$ dB.
- "Constant" means $w_m(t, f) = 1, \forall(m, t, f)$, leading to original sample SCMs (the signal enhancement model is not used).



(a) BT with different β.



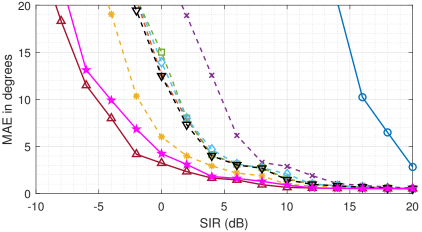
(b) Overall comparison.

Figure: MAE in degrees vs. SIR.

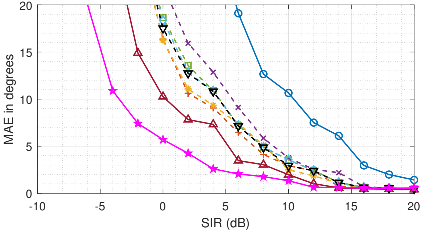
Observation 1

WSCMs can easily become singular when $\beta \geq 0.95$.

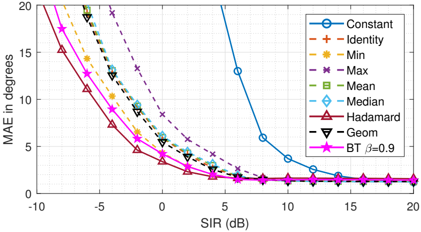
The best post-processing is criterion-dependent



(a) The proposed method.



(b) The principal vector method.



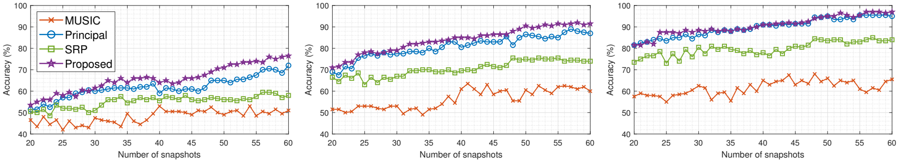
(c) The SRP method.

Figure: MAE in degrees vs. SIR.

How does the proposed method perform?

Table: DoA estimation accuracy. $K = 2$. SNR $= 20$ dB.

RT_{60} (seconds)	0.3			0.9		
SIR (dB)	-6	0	$+6$	-6	0	$+6$
MUSIC	40%	52%	59%	30%	30%	33%
Principal	43%	77%	89%	51%	70%	79%
SRP	33%	59%	75%	28%	37%	40%
Proposed	**54%**	**81%**	**91%**	**59%**	**76%**	**88%**



(a) -6 dB SIR. (b) 0 dB SIR. (c) $+6$ dB SIR.

Figure: Accuracy vs. number of snapshots T. $K = 1$, $RT_{60} = 0.3$s, and SNR $= 20$ dB.

A closer look at the proposed method

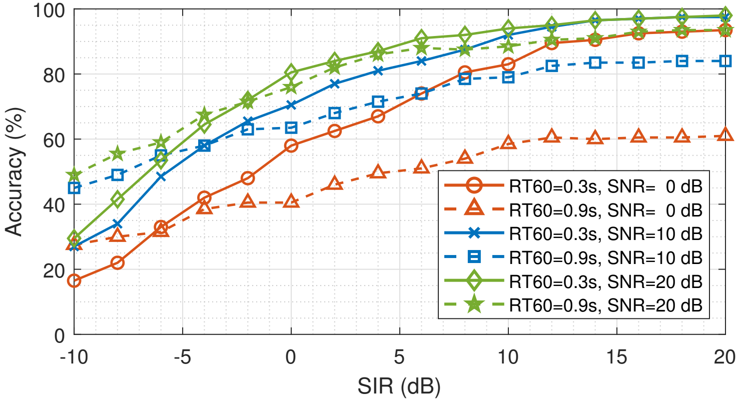
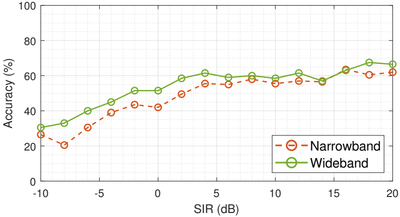
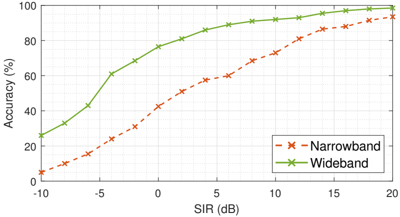


Figure: Evaluation of the proposed method. $K = 2$.

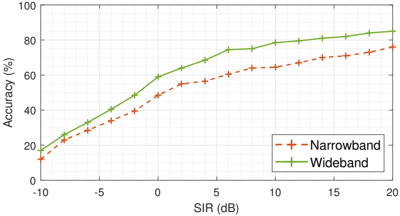
Wideband vs. Narrowband



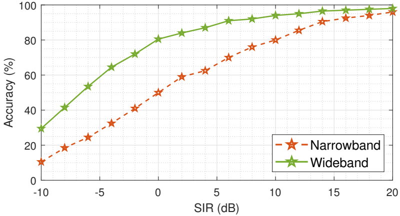
(a) MUSIC.



(b) The principal vector method.



(c) The SRP method.



(c) The proposed method.

Figure: Summing spatial spectra over the wideband (50 Hz to 7 kHz) is more beneficial than summing them over the narrowband (300 Hz to 3400 Hz).

Conclusion and future work

Takeaway

- The snapshot is first **filtered** and then **normalized**.
- The normalized T-F weighted criterion is **simple** but **effective**.
- Post-processing is important and the best design is **criterion-dependent**.
- Pick a post-processing? Try Hadamard product or BT (with a tuned β).

Future work

- Can the criterion be derived from the maximum likelihood principle under mild assumptions on the noise covariance matrix?
- Do we have the same conclusion for a very different DNN architecture?
- Extension to multiple speech sources and interferences.

References

Cox, R. V., Neto, S. F. D. C., Lamblin, C., and Sherif, M. H. (2009). ITU-T coders for wideband, superwideband, and fullband speech communication [series editorial]. *IEEE Communications Magazine*, 47(10):106–109.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (1993). TIMIT acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*.

Heymann, J., Drude, L., and Haeb-Umbach, R. (2016). Neural network based spectral mask estimation for acoustic beamforming. In *ICASSP*, pages 196–200. IEEE.

Hu, G. and Wang, D. (2010). A tandem algorithm for pitch estimation and voiced speech segregation. *IEEE TASLP*, 18(8):2067–2079.

Pertilä, P. and Cakir, E. (2017). Robust direction estimation with convolutional neural networks based steered response power. In *ICASSP*, pages 6125–6129. IEEE.

Pisha, L., Warchall, J., Zubatiy, T., Hamilton, S., Lee, C.-H., Chockalingam, G., Mercier, P. P., Gupta, R., Rao, B. D., and Garudadri, H. (2019). A wearable, extensible, open-source platform for hearing healthcare research. *IEEE Access*, 7:162083–162101.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer.

Scheibler, R., Bezzam, E., and Dokmanić, I. (2018). Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *ICASSP*, pages 351–355. IEEE.

Wang, Z.-Q., Zhang, X., and Wang, D. (2018a). Robust speaker localization guided by deep learning-based time-frequency masking. *IEEE/ACM TASLP*, 27(1):178–188.

Wang, Z.-Q., Zhang, X., and Wang, D. (2018b). Robust TDOA estimation based on time-frequency masking and deep neural networks. In *Interspeech*, pages 322–326.

Xu, C., Xiao, X., Sun, S., Rao, W., Chng, E. S., and Li, H. (2017). Weighted spatial covariance matrix estimation for MUSIC based TDOA estimation of speech source. In *Interspeech*, pages 1894–1898.

Yang, B., Liu, H., and Pang, C. (2017). Multiple sound source counting and localization based on spatial principal eigenvector. In *Interspeech*, pages 1924–1928.

Yang, B., Liu, H., Pang, C., and Li, X. (2019). Multiple sound source counting and localization based on TF-wise spatial spectrum clustering. *IEEE/ACM TASLP*, 27(8):1241–1255.

Zhang, C., Florêncio, D., and Zhang, Z. (2008). Why does phat work well in low noise, reverberative environments? In *ICASSP*, pages 2565–2568. IEEE.

Q&A and more...

- If you would like to learn more about single-channel speech enhancement...
- Welcome to our poster presentation (SLT-P38.8) tomorrow!

LEVERAGING HETEROSCEDASTIC UNCERTAINTY IN LEARNING COMPLEX SPECTRAL MAPPING FOR SINGLE-CHANNEL SPEECH ENHANCEMENT

Kuan-Lin Chen[12†], *Daniel D. E. Wong*[1*], *Ke Tan*[1], *Buye Xu*[1], *Anurag Kumar*[1], *and Vamsi Krishna Ithapu*[1]

[1]Meta Reality Labs Research
[2]Department of Electrical and Computer Engineering, University of California, San Diego