# ADL-MVDR: ALL DEEP LEARNING MVDR BEAMFORMER FOR TARGET SPEECH SEPARATION

1240

Zhuohuang Zhang[1,2]; Yong Xu[2]; Meng Yu[2]; Shi-Xiong Zhang[2]; Lianwu Chen[2]; Dong Yu[2]
[1]Indiana University Bloomington; [2]Tencent AI Lab
zhuozhan@iu.edu; {lucayongxu, raymondmyu, auszhang, lianwuchen, dyu}@tencent.com

## Introduction

Speech separation algorithms are often used to separate the target speech from other interfering sources. However, purely neural network (NN) based speech separation systems often cause **nonlinear distortion** that is harmful for automatic speech recognition (ASR) systems. The conventional mask-based minimum variance distortionless response (MVDR) beamformer can be used to minimize the distortion, but comes with high level of **residual noise**. Furthermore, the matrix operations (e.g., matrix inversion) involved in the conventional MVDR solution are sometimes numerically unstable when jointly trained with neural networks.

We propose a novel all deep learning MVDR (ADL-MVDR) framework, where the matrix inversion and eigenvalue decomposition are replaced by two recurrent neural networks (RNNs), to resolve both issues at the same time.

Contributions of this work:
(1). A novel ADL-MVDR framework which can be jointly trained stably with the front-end filter estimator for predicting frame-level beamforming weights.
(2). Replacing the matrix inversion and PCA involved in the MVDR solution with two separate RNNs, instead of utilizing the traditional mathematical approach.
(3). A complex ratio filtering method (Mack and Habets, 2019, denoted as cRF) to further stabilize joint training process and estimate the covariance matrix more accurately.

## Conventional Mask-based MVDR

**MVDR solution (Higuchi et al. 2018; Shimada et al. 2018):**

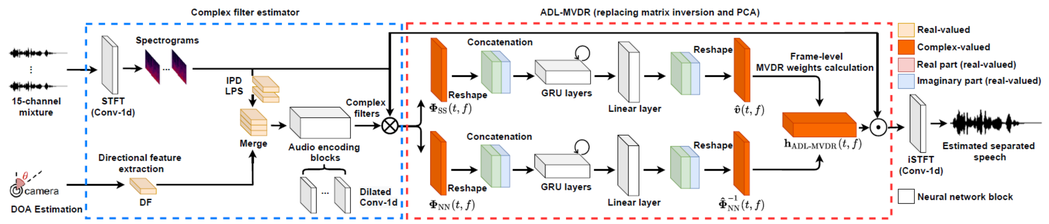
$$\mathbf{h}(t,f) = \frac{\hat{\mathbf{\Phi}}_{\mathrm{NN}}^{-1}(t,f)\hat{\boldsymbol{v}}(t,f)}{\hat{\boldsymbol{v}}^{\mathrm{H}}(t,f)\hat{\mathbf{\Phi}}_{\mathrm{NN}}^{-1}(t,f)\hat{\boldsymbol{v}}(t,f)}, \quad \mathbf{h}(t,f) \in \mathbb{C}^{M}$$

**Covariance matrix (speech part):**

$$\mathbf{\Phi}_{\mathrm{SS}}(f) = \frac{\sum_{t=1}^{T} \hat{\mathbf{S}}_{\mathrm{cRM}}(t,f)\hat{\mathbf{S}}_{\mathrm{cRM}}^{\mathrm{H}}(t,f)}{\sum_{t=1}^{T} \mathbf{M}_{\mathrm{S}}^{\mathrm{H}}(t,f)\mathbf{M}_{\mathrm{S}}(t,f)}$$

where $\hat{\mathbf{S}}_{\mathrm{cRM}}$ is the estimated speech component using a complex ratio mask $\mathbf{M}_{\mathrm{S}}$. The noise covariance matrix can be obtained in a similar way.

## Proposed ADL-MVDR




**Network architecture**

**cRF for covariance matrix estimation:**

$$\hat{\mathbf{S}}_{\mathrm{cRF}} = \sum_{\tau_1=-\mathrm{L}}^{\mathrm{L}} \sum_{\tau_2=-\mathrm{K}}^{\mathrm{K}} \mathbf{F}_{\mathrm{S}}(\mathrm{t}+\tau_1, \mathrm{f}+\tau_2) * \mathbf{Y}(\mathrm{t}+\tau_1, \mathrm{f}+\tau_2),$$

$$\mathbf{\Phi}_{\mathbf{SS}}(t,f) = \frac{\hat{\mathbf{S}}_{\mathrm{cRF}}(t,f)\,\hat{\mathbf{S}}_{\mathrm{cRF}}^{\mathrm{H}}(t,f)}{\sum_{\mathrm{t}=1}^{\mathrm{T}} \mathbf{M}_{\mathrm{S}}^{\mathrm{H}}(t,f)\mathbf{M}_{\mathrm{S}}(t,f)},$$

where the cRF (Mack and Habets, 2019) is equivalent to (2K+1)X(2L+1) number of cRMs that each applied to the corresponding shifted version of the noisy spectrogram.

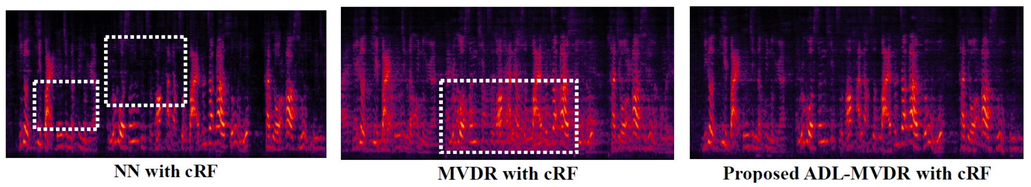
**RNNs for replacing matrix inversion and PCA in MVDR:**
The GRU-Nets can better utilize temporal information from previous frames for estimating statistical terms than conventional frame-wise approaches that are based on heuristic updating factors (Souden et al. 2011; Tammen et al., 2019).

$$\hat{\boldsymbol{v}}(t,f) = \mathbf{GRU\text{-}Net}_{\boldsymbol{v}}(\mathbf{\Phi}_{\mathrm{SS}}(t,f))$$

$$\hat{\mathbf{\Phi}}_{\mathrm{NN}}^{-1}(t,f) = \mathbf{GRU\text{-}Net}_{\mathrm{NN}}(\mathbf{\Phi}_{\mathrm{NN}}(t,f))$$

$$\mathbf{h}(t,f) = \frac{\hat{\mathbf{\Phi}}_{\mathrm{NN}}^{-1}(t,f)\hat{\boldsymbol{v}}(t,f)}{\hat{\boldsymbol{v}}^{\mathrm{H}}(t,f)\hat{\mathbf{\Phi}}_{\mathrm{NN}}^{-1}(t,f)\hat{\boldsymbol{v}}(t,f)}$$

**Experimental Results:**



NN with cRF | MVDR with cRF | Proposed ADL-MVDR with cRF

| Systems/Metrics | PESQ ∈ [−0.5, 4.5] | | | | | | | | Si-SNR (dB) | SDR (dB) | WER (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0-15° | 15-45° | 45-90° | 90-180° | 1spk | 2spk | 3spk | Avg. | Avg. | Avg. | Avg. |
| Reverberant clean (reference) | 4.50 | 4.50 | 4.50 | 4.50 | 4.50 | 4.50 | 4.50 | 4.50 | ∞ | ∞ | 8.26 |
| Noisy Mixture (interfering speech + noise) | 1.88 | 1.88 | 1.98 | 2.03 | 3.55 | 2.02 | 1.77 | 2.16 | 3.39 | 3.50 | 55.14 |
| NN with cRM | 2.72 | 2.92 | 3.09 | 3.07 | 3.96 | 3.02 | 2.74 | 3.07 | 12.23 | 12.73 | 22.49 |
| NN with cRF (3×3) | 2.75 | 2.95 | 3.12 | 3.09 | 3.98 | 3.06 | 2.76 | 3.10 | 12.50 | 13.01 | 22.07 |
| MVDR with cRM [8] | 2.55 | 2.76 | 2.96 | 2.84 | 3.73 | 2.88 | 2.56 | 2.90 | 10.62 | 12.04 | 16.85 |
| MVDR with cRF (3×3) | 2.55 | 2.77 | 2.96 | 2.89 | 3.82 | 2.90 | 2.55 | 2.92 | 11.31 | 12.58 | 15.91 |
| Multi-tap MVDR with cRM (2-tap) [8] | 2.70 | 2.96 | 3.18 | 3.09 | 3.80 | 3.07 | 2.74 | 3.08 | 12.56 | 14.11 | 13.67 |
| Multi-tap MVDR with cRF (2-tap, 3×3) | 2.67 | 2.95 | 3.15 | 3.10 | 3.92 | 3.06 | 2.72 | 3.08 | 12.66 | 14.04 | 13.52 |
| **Proposed ADL-MVDR with cRF (3×3)** | **3.04** | **3.30** | **3.48** | **3.48** | **4.17** | **3.41** | **3.07** | **3.42** | **14.80** | **15.45** | **12.73** |

## Speech Materials

We use our previously reported Mandarin audio-visual dataset (Xu et al., 2020; Tan et al., 2020; Gu et al., 2020) collected from Youtube as the speech corpus. Different from our previous works (Tan et al., 2020; Gu et al., 2020), the lip movement feature is not fed into the model in this study as we focus on the beamforming.

The corpus contains 205500 audio clips (roughly 200 hours) with sampling rate set to 16 kHz. The simulated multi-channel audio data contains sources from different speakers (either target or interfering sources). The audios are further mixed with random cuts of noises recorded indoors and different reverberation conditions (T60s from 0.05 s to 0.7 s) are applied (Tan et al., 2020).

## Results Analysis

- Demos (including real-world recording evaluation) are available at: https://zzhang68.github.io/adlmvdr/
- The proposed ADL-MVDR system achieves significantly better results across all metrics and ASR accuracy than purely NN-based systems.
- The proposed ADL-MVDR system achieves about 17% PESQ improvement over the baseline MVDR system with cRF (i.e., 3.42 vs. 2.92). In terms of ASR accuracy, the proposed ADL-MVDR system outperforms MVDR with cRF by a large margin (i.e., 12.73% vs. 15.91%).
- The NN with cRF achieves better performance in all metrics (e.g., Si-SNR: 12.50 dB vs. 12.23 dB) and ASR accuracy (i.e., 22.07% vs. 22.49%) than NN with cRM. Slight improvements can be found on conventional MVDR systems due to utterance-level weights.

## Future Work

The future of our proposed ADL-MVDR framework is promising and it could be generalized to many speech separation systems. We will further verify this idea on single-channel speech separation and dereverberation tasks

Tencent AI Lab

INDIANA UNIVERSITY