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

IEEE ICASSP 2021

Zhuohuang Zhang[1], Yong Xu[2], Meng Yu[2], Shi–Xiong Zhang[2], Lianwu Chen[2], Dong Yu[2]

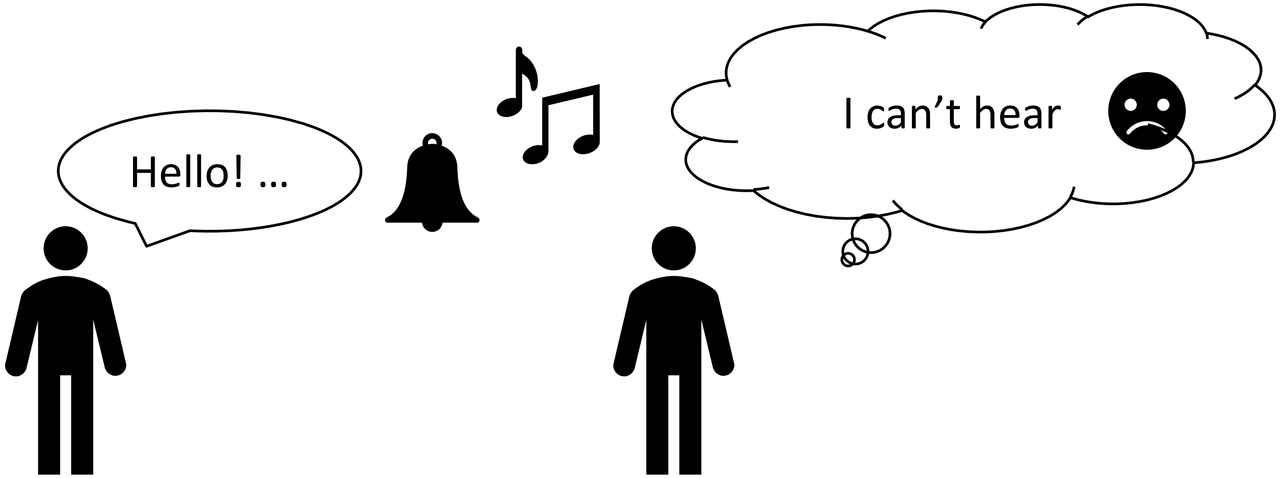
[1] Indiana University Bloomington

[2] Tencent AI Lab

This work was done while Z. Zhang was an intern at Tencent AI Lab, Bellevue, USA

- **Introduction**
- Proposed ADL-MVDR System
- Experimental Setup
- Experimental Results
- Conclusions and Future Work

Introduction



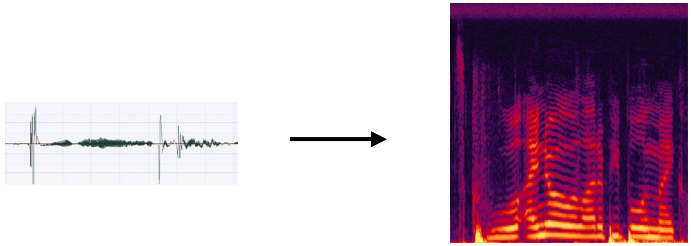
- Goal of Target Speech Separation:

Removing the background noise, interfering sources while maintaining the target speech.

- Applications: front-ends for ASR systems (Du et al. 2014), digital hearing-aids (Van den Bogaert et al., 2019).

Introduction

Prior Studies:

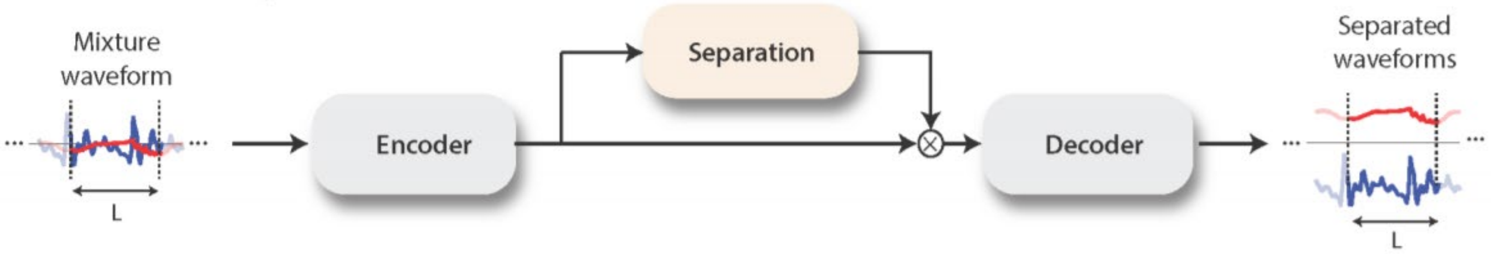


- Time-frequency (TF) domain mask-based systems:

  E.g., DNN-IRM (Wang et al., 2014), DNN-cIRM (Williamson et al., 2015), etc.

- End-to-end systems:

  E.g., TasNet (Luo and Mesgarani, 2018), Conv-TasNet (Luo and Mesgarani, 2019)

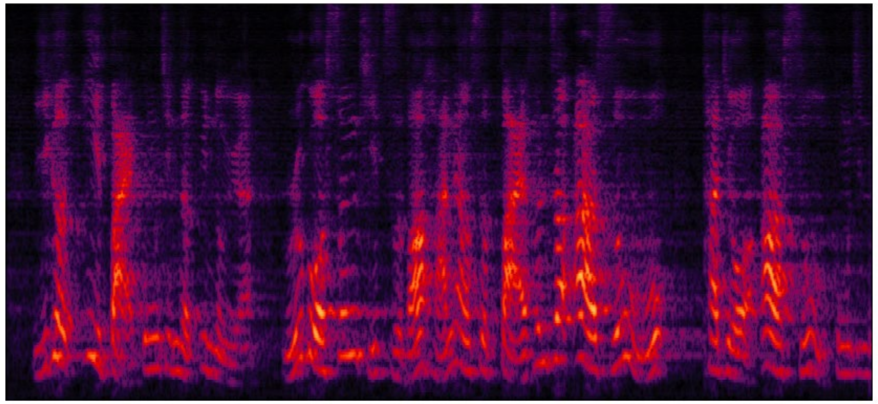


Luo, Yi, and Nima Mesgarani. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation." *IEEE/ACM TASLP* 27.8 (2019): 1256-1266.

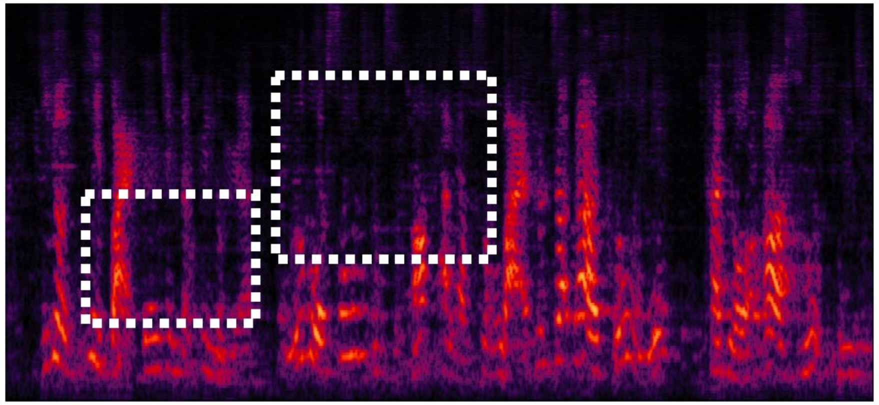
Introduction

Limitations:

- Purely NN-based speech separation systems often lead to large amount of non-linear distortion (e.g., spectral blackholes), which is very harmful to ASR systems.



Clean Reference

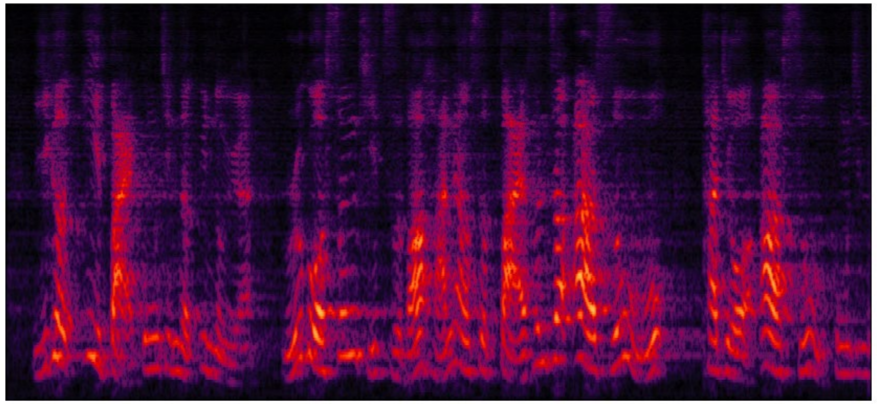


NN Separated Speech

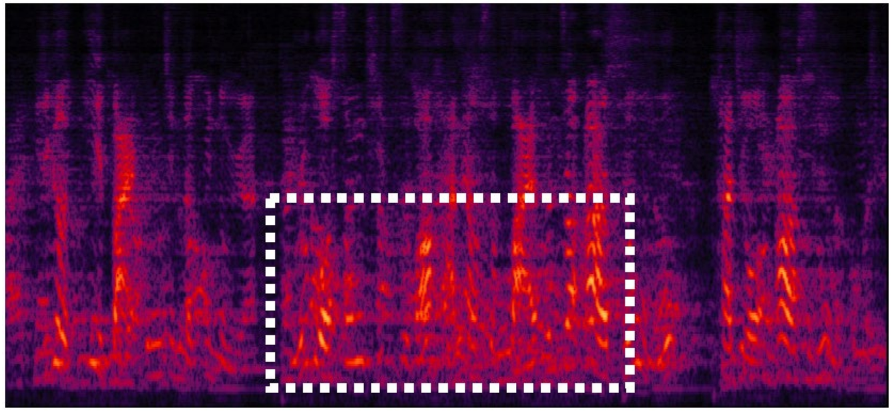
Introduction

Solution (Higuchi et al., 2018): $\mathbf{h}_{\mathbf{MVDR}} = \arg\min_{\mathbf{h}} \mathbf{h}^{\mathrm{H}} \mathbf{\Phi}_{\mathbf{NN}} \mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^{\mathrm{H}} \boldsymbol{v} = 1,$ $\hat{s}_{\mathbf{MVDR}}(t, f) = \mathbf{h}^{\mathrm{H}}(f) \mathbf{Y}(t, f),$

- The minimum variance distortionless response (MVDR) filters can help to introduce less non-linear distortions to the separated speech.
- However, it often results in high level of residual noise (trade-off between noise reduction and distortion).



Clean Reference



MVDR Separated Speech

Introduction

Limitations on conventional MVDR:

1. High level of residual noise mainly due to utterance/chunk-level beamforming weights.

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

2. The steering vector (obtained by applying PCA on speech covariance matrix) and inverse of noise covariance matrix are sometimes numerically unstable when jointly trained with NN.

- Introduction
- **Proposed ADL-MVDR System**
- Experimental Setup
- Experimental Results
- Conclusions and Future Work

Proposed System

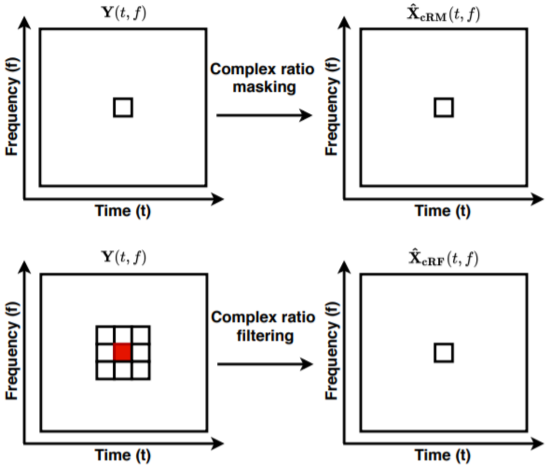
ADL-MVDR:

Replacing matrix inversion and eigenvalue decomposition with two recurrent neural networks (RNNs):

1. Solves the issue of chunk-level beamforming weights and therefore can greatly reduce the residual noise.

2. Also solves the instability issue when jointly trained with NNs.

Proposed System



ADL-MVDR:

We applied a complex ratio filter (cRF) (Mack and Habets, 2019) for speech and noise component estimation.

$$\begin{aligned}\hat{\mathbf{S}}_{\mathbf{cRF}}(t,f) = &\sum_{\tau_1=-L}^{L}\sum_{\tau_2=-K}^{K} \mathrm{cRF}(t+\tau_1, f+\tau_2)\\ &* \mathbf{Y}(t+\tau_1, f+\tau_2),\\ \mathbf{\Phi}_{\mathbf{SS}}(t,f) = &\frac{\hat{\mathbf{S}}_{\mathbf{cRF}}(t,f)\hat{\mathbf{S}}^{\mathrm{H}}_{\mathbf{cRF}}(t,f)}{\sum_{t=1}^{T}\mathrm{cRM}^{\mathrm{H}}_{\mathrm{S}}(t,f)\mathrm{cRM}_{\mathrm{S}}(t,f)},\end{aligned}$$

Then we use RNNs to replace the PCA and matrix inversion given the frame-wise covariance matrices, thus the frame-wise BF weights can be derived:

$$\begin{aligned}\hat{\boldsymbol{v}}(t,f) &= \textbf{GRU-Net}_{\boldsymbol{v}}(\mathbf{\Phi}_{\mathbf{SS}}(t,f)),\\ \hat{\mathbf{\Phi}}^{-1}_{\mathbf{NN}}(t,f) &= \textbf{GRU-Net}_{\mathbf{NN}}(\mathbf{\Phi}_{\mathbf{NN}}(t,f)),\end{aligned}$$

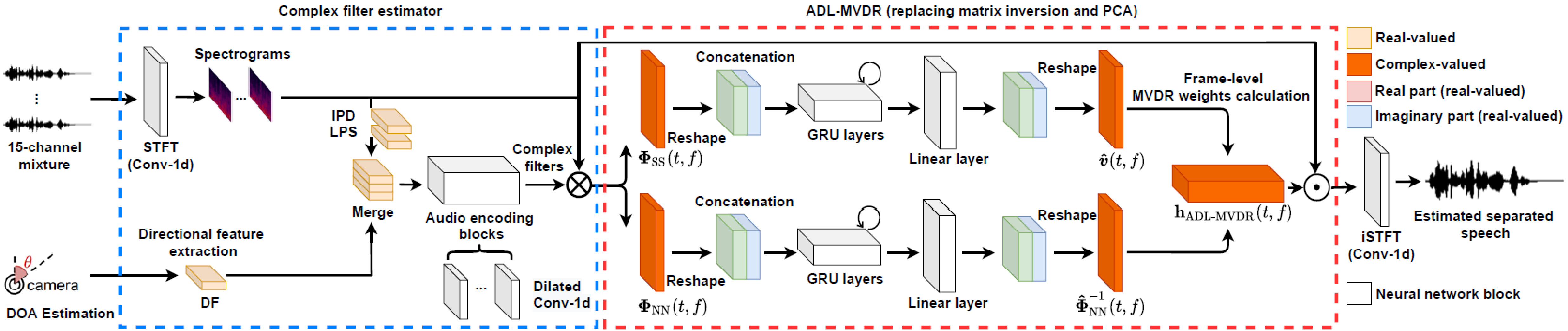
ADL-MVDR:

Frame-level MVDR weights estimation:

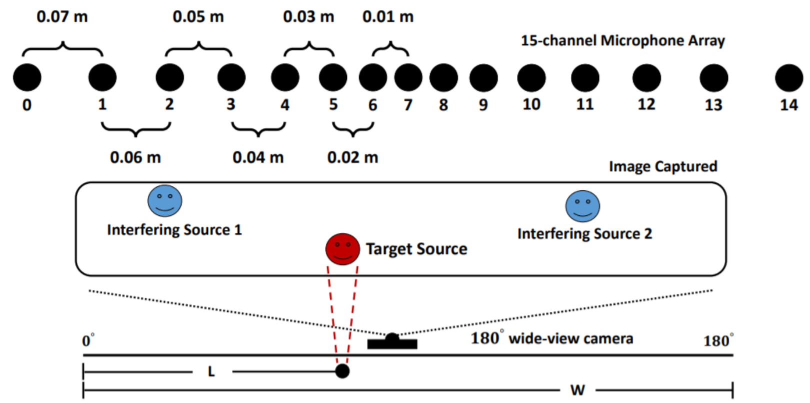
$$\boxed{v(t,f) = \mathbf{RNN}(\boldsymbol{\Phi}_{\boldsymbol{SS}}(t,f)) \qquad \boldsymbol{\Phi}_{NN}^{-1}(t,f) = \mathbf{RNN}(\boldsymbol{\Phi}_{\boldsymbol{NN}}(t,f))}$$

$$w_{ADL-MVDR}(t,f) = \frac{\boldsymbol{\Phi}_{NN}^{-1}(t,f)\boldsymbol{v}(t,f)}{\boldsymbol{v}^{H}\ (t,f)\boldsymbol{\Phi}_{NN}^{-1}(t,f)\boldsymbol{v}(t,f)}$$

Network Structure:



DOA Estimation:



- Introduction
- Proposed ADL-MVDR System
- **Experimental Setup**
- Experimental Results
- Conclusions and Future Work

Experimental Setup

Dataset (Tan et al., 2020; Gu et al., 2020): A Mandarin audio-visual speech corpus collected from Youtube.

The corpus contains 205500 audio clips (roughly 200 hours) with sampling rate set to 16 kHz.

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.

190,000 speech utterances in the training set, 15,000 utterances in the validation set, and another 500 utterances in the testing set.

Experimental Setup

- The interaural phase difference (IPD) and log-power spectra (LPS) features from the 15-channel microphone recorded mixture is merged with a location guided directional feature (DF) as input to the audio encoding network (a Conv-TasNet Variant (Luo and Mesgarani, 2019; Tan et al., 2020)).

- 512-point STFT together with a 32 ms Hann window and 16 ms step size to extract audio features.

- The batch size and audio chunk size are set to 12 and 4 s.

- The system is trained to minimize the time-domain Si-SNR loss.

$$\mathcal{L}_{Si\text{-}SNR} = -20log_{10}\frac{\|\hat{x} - \alpha \cdot x\|}{\|\alpha \cdot x\|},$$

$$\alpha = \frac{\hat{x}^T x}{x^T x},$$

- Introduction

- Proposed ADL-MVDR System

- Experimental Setup

- **Experimental Results**

- Conclusions and Future Work

Experimental Results

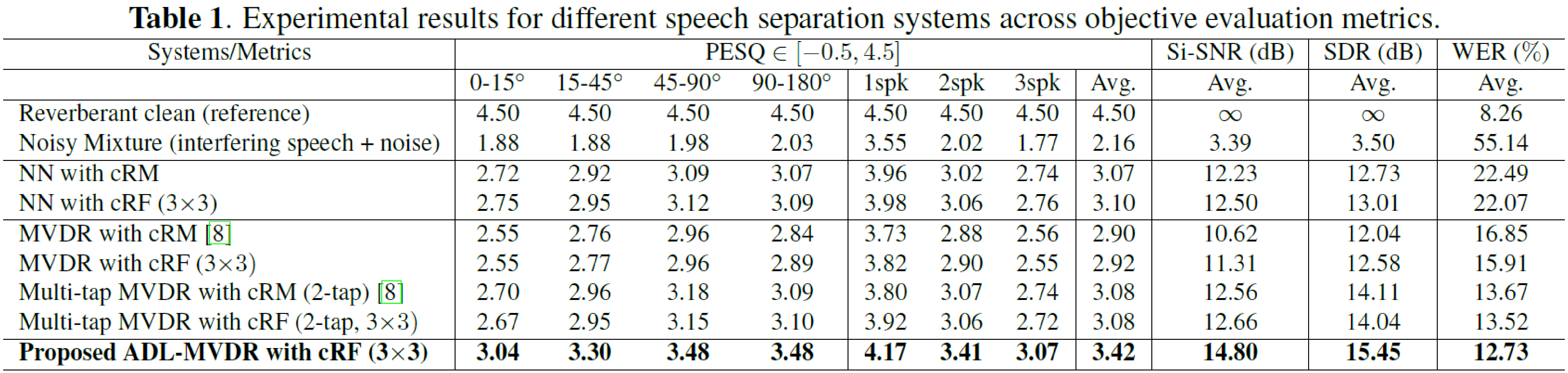


Table 1. Experimental results for different speech separation systems across objective evaluation metrics.

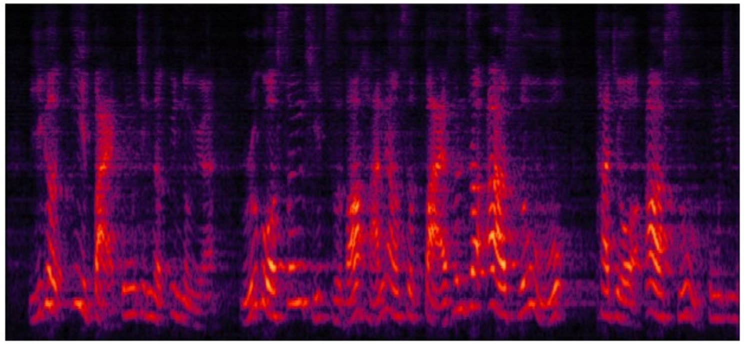
Systems/Metrics	PESQ $\in [-0.5, 4.5]$								Si-SNR (dB)	SDR (dB)	WER (%)
	0-15°	15-45°	45-90°	90-180°	1spk	2spk	3spk	Avg.	Avg.	Avg.	Avg.
Reverberant clean (reference)	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	∞	∞	8.26
Noisy Mixture (interfering speech + noise)	1.88	1.88	1.98	2.03	3.55	2.02	1.77	2.16	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**

1. Significantly better results across all metrics and ASR accuracy than purely NN-based systems.
2. Compared to purely NNs, huge improvements on ASR accuracy (i.e., WER)
3. Compared to conventional mask-based MVDRs, huge improvements on objective metrics (i.e., PESQ, SI-SNR, SDR).

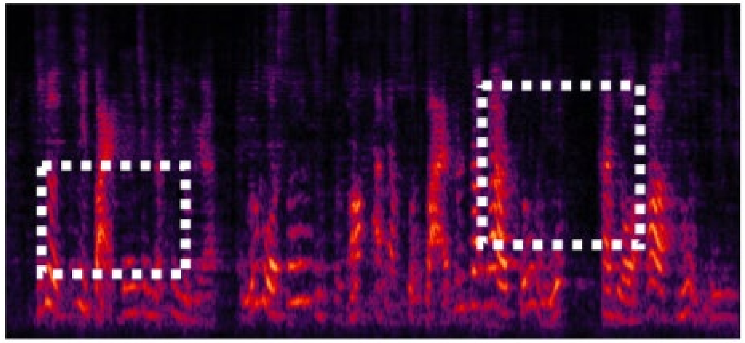
Experimental Results

3pk Mix.

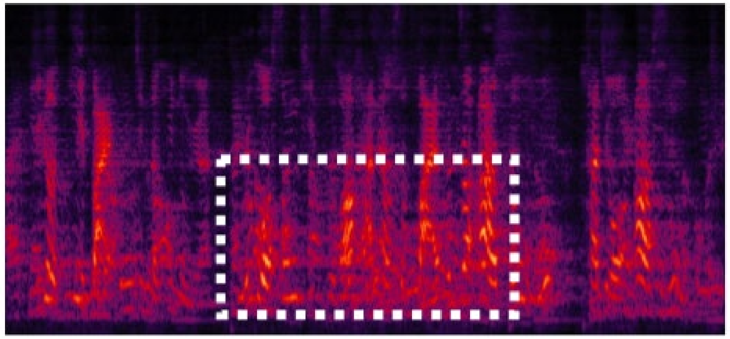
Some example spectrograms



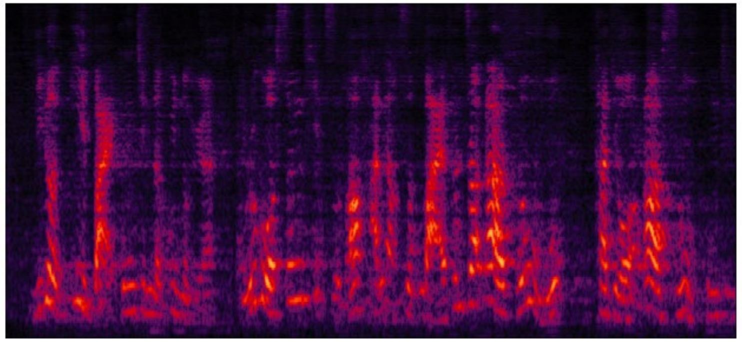
Clean reference



Purely NNs



Conventional MVDR

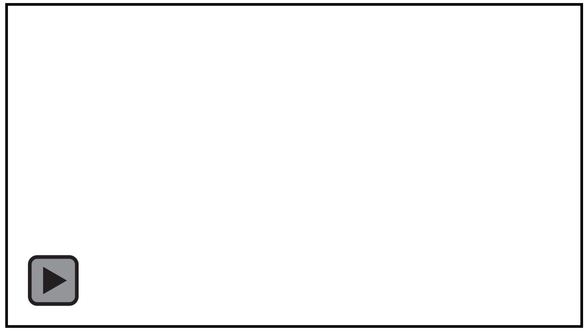


Prop. ADL-MVDR

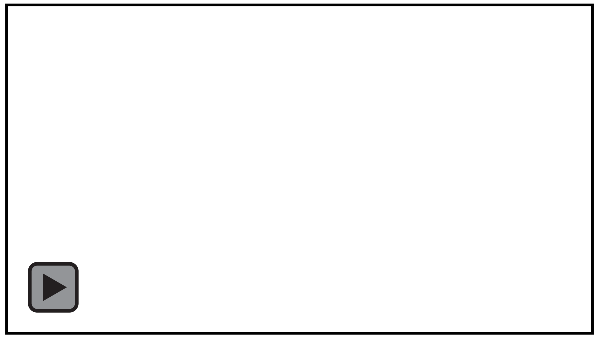
https://zzhang68.github.io/adlmvdr/

Experimental Results

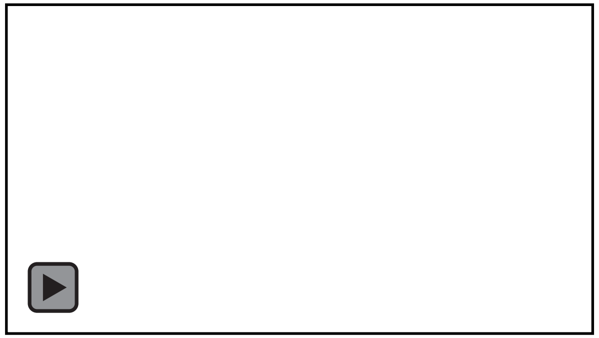
Real-world demos:



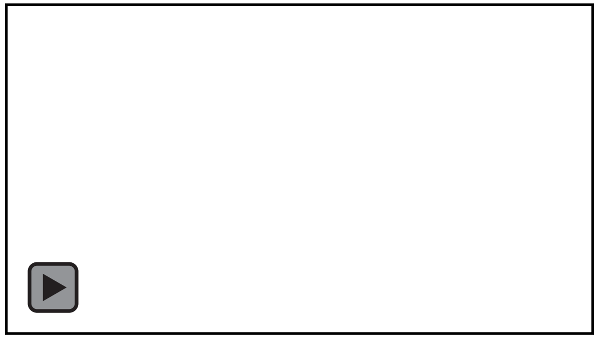
Mixture



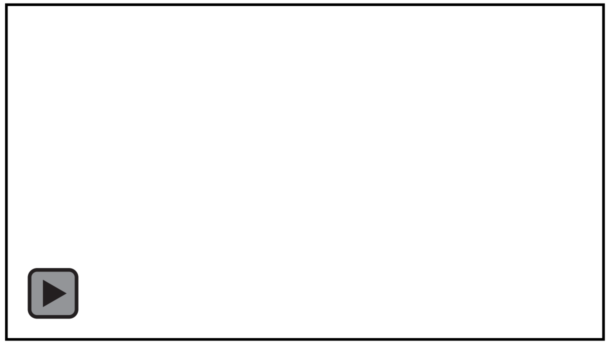
Separated **male** voice by multi-tap MVDR



Separated **male** voice by **ADL-MVDR**



Separated **female** voice by multi-tap MVDR



Separated **female** voice by **ADL-MVDR**

https://zzhang68.github.io/adlmvdr/

- Introduction
- Proposed ADL-MVDR System
- Experimental Setup
- Experimental Results
- **Conclusions and Future Work**

Conclusions:

1. A novel all deep learning MVDR method has been proposed to recursively learn the spatio-temporal filtering for multi-channel target speech separation.

2. The proposed system outperforms prior arts across several objective metrics and ASR accuracy.

3. The future of our proposed ADL-MVDR framework is promising and it could be generalized to many other speech separation systems.

Thank you for your attention!