Deep Learning Based Speech Beamforming

Kaizhi Qian[1], Yang Zhang[1], Shiyu Chang[2], Xuesong Yang[1], Dinei Florencio[3], Mark Hasegawa-Johnson[1]
[1]University of Illinois at Urbana-Champaign, USA [2]IBM Watson Research Center, USA [3]Microsoft Research, USA

Problem

- Ad-hoc microphone array where the number, the locations and the types of microphones are unknown
- Intended for human consumptions instead of speech recognition

Motivation

Deep learning methods and traditional beamforming methods mutually compensate.

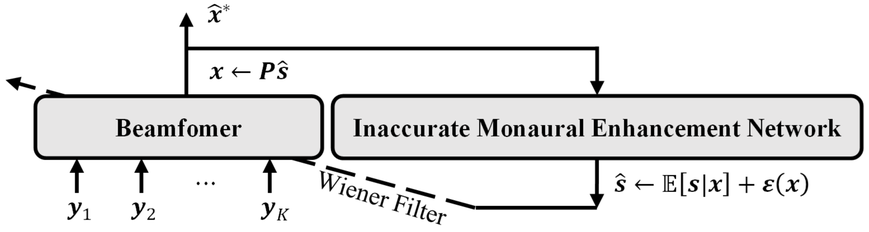
Deep Learning Methods

Pros	Cons
▪ Good at removing noise and reverberation	▪ Produce artifacts ▪ Generalize poorly to unseen noise

Traditional Beamforming Methods

Pros	Cons
▪ Produce natural-sounding speech without artifacts	▪ Require speaker location and interference characteristic ▪ Oversimplified prior knowledge

Algorithms



Beamformer

Goal: $\min_{x=Yh} \mathbb{E}[\|x - s\|_W^2 | y]$ (Beamformer Output; Clean Speech; Noisy Input)

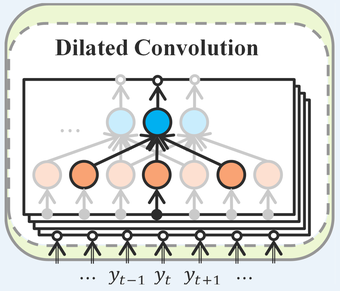
Solution:

$$x^* = P\mathbb{E}[x|y]$$
$$P = Y(Y^T W Y)^{-1} Y^T W$$

The beamformer projects the noisy speech to the beamformer output space, removing the artifacts generated by the enhancement network.

Enhancement Network

Structure: Non-causal WaveNet



The enhancement network computes the posterior expectation:

$$\mathbb{E}[x|y]$$

which removes noise and reverberation.

Experiments

Simulated Test

- S1: Seen speaker, Seen noise
- S2: Seen speaker, Unseen noise
- S3: Unseen speaker, Seen noise
- S4: Unseen speaker, Unseen noise
- GRAB: Glottal Residual Assisted Beamforming
- MVDR: Minimum Variance Distortionless Response
- IVA: Independent Vector Analysis

Speech Data
Seen: VCTK Unseen: TIMIT
Noise Data
Seen: Hu100 Unseen: FreeSFX

Er =		-10	0	10	20
SNR (dB)	DeepBeam S1	18.5	22.0	26.5	28.4
	DeepBeam S2	17.1	20.3	25.9	27.4
	DeepBeam S3	15.3	19.5	24.1	27.6
	DeepBeam S4	14.1	19.0	23.1	28.5
	GRAB	2.48	12.5	21.6	25.4
	CLOSEST	-5.13	3.38	14.9	24.8
	MVDR	8.41	12.9	22.6	26.7
	IVA	10.3	13.3	16.8	19.2
DRR (dB)	DeepBeam S1	3.45	8.97	11.2	11.5
	DeepBeam S2	7.38	11.9	12.6	11.5
	DeepBeam S3	5.60	4.85	8.43	9.78
	DeepBeam S4	2.11	6.68	7.10	9.31
	GRAB	-0.83	1.70	3.63	3.68
	CLOSEST	8.56	7.32	7.67	8.44
	MVDR	-2.17	-3.47	-3.42	-4.13
	IVA	-8.92	-8.77	-8.81	-8.99

Real-World Test

- N1: Cell phone
- N2: CombBind
- N3: Paper shuffle
- N4: Door slide
- N5: Footsteps

Audio Demo

Trained on simulated data
Test on real-world data

Noise Type		N1	N2	N3	N4	N5
SNR (dB)	DeepBeam	**20.1**	**20.0**	**16.9**	**19.6**	**18.7**
	GRAB	18.9	17.4	12.4	18.5	17.4
	CLOSEST	10.0	10.0	10.0	10.0	10.0
	MVDR	10.8	16.5	7.72	14.0	13.4
	IVA	11.7	9.74	6.83	12.4	15.9
MOS	DeepBeam	**3.83**	**3.72**	**3.63**	**4.09**	**4.20**
	GRAB	3.10	3.06	2.93	3.71	3.45
	CLOSEST	2.74	2.68	3.02	3.55	3.50
	MVDR	2.05	2.40	2.28	2.71	2.62
	IVA	1.73	2.03	1.75	1.78	2.08

Convergence Analysis