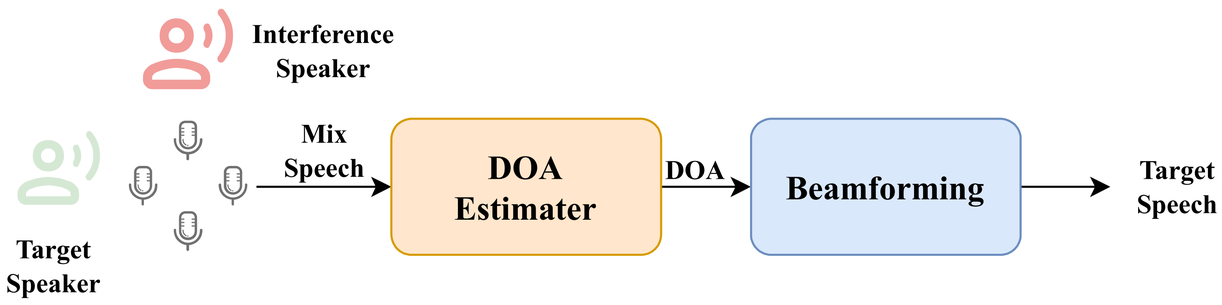
Robust Target Speaker Direction of Arrival Estimation

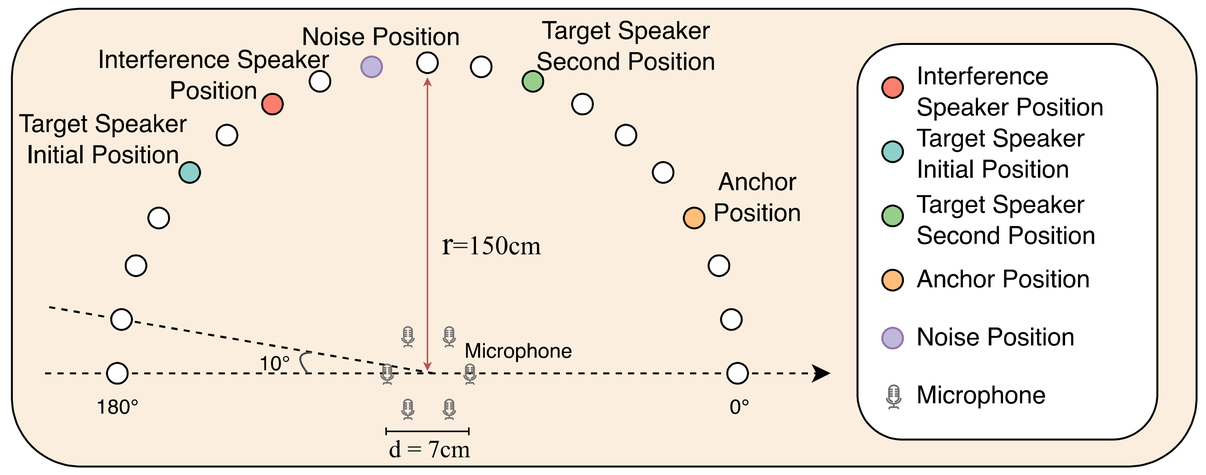
Zixuan Li, Shulin He, Xueliang Zhang
College of Computer Science, Inner Mongolia University, China

Introduction



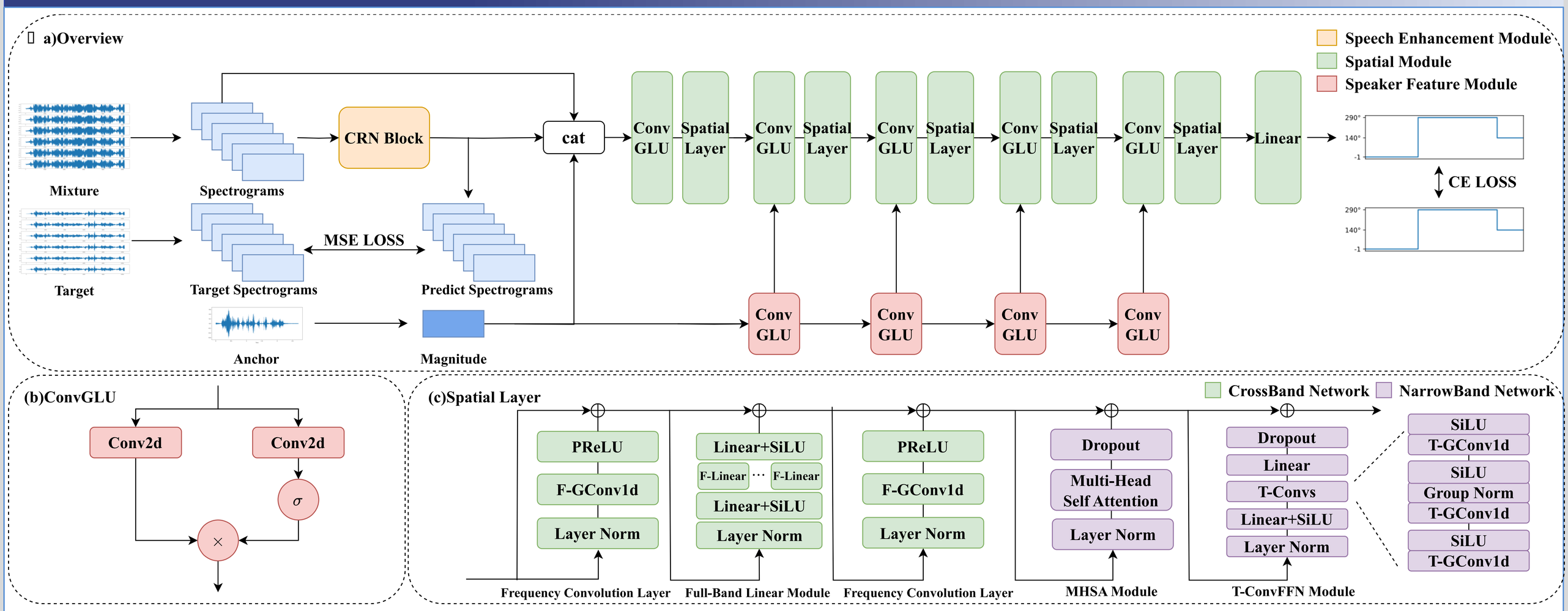
- Accurate DOA estimation is beneficial for extracting target speech in multi-speaker scenarios
- Estimating the DOA of the target speaker in multi-speaker scenarios is challenging

System Overview



- We regard DOA estimation as a classification problem, and ±10 degrees is an acceptable error.

Method



- Speech Enhancement Module reducing the adverse effects of noise and reverberation.
- Spatial Module extracts spatial information using full-sub band modeling while downsampling to reduce computation.
- Speaker Feature Module utilizes anchor to extract speaker information for speaker recognition.

Result

ID	SIR	-5		-2		0		2		5		AVG	
	Metrics	VDE ↓	AR ↑	VDE ↓	AR ↑	VDE ↓	AR ↑	VDE ↓	AR ↑	VDE ↓	AR ↑	VDE ↓	AR ↑
0	TS-DOA	0.1657	0.5015	0.1369	0.5497	0.1464	0.5923	0.0981	0.6481	0.1007	0.6624	0.1285	0.6030
1	w/o Enhancement	0.1643	0.6867	0.1503	0.7015	0.1457	0.7582	0.1364	0.7210	0.1285	0.7881	0.1497	0.7117
2	w/o Local	0.1782	0.6384	0.1717	0.7225	0.1487	0.7102	0.1430	0.7443	0.1230	0.7907	0.1487	0.7093
3	w global	0.1558	0.7723	0.1232	0.7838	0.1174	0.8111	0.1113	0.8566	0.0999	0.8730	0.1131	0.8220
4	RTS-DOA(Mag)	0.1899	0.5328	0.1689	0.0273	0.1723	0.6790	0.1422	0.6926	0.1321	0.7476	0.1559	0.6642
5	**RTS-DOA**	**0.1236**	**0.8940**	**0.1019**	**0.8837**	**0.0843**	**0.9166**	**0.0830**	**0.9136**	**0.0650**	**0.9364**	**0.0896**	**0.8940**
6	**RTS-DOA(Large)**	**0.1089**	**0.8632**	**0.0830**	**0.9061**	**0.0758**	**0.9394**	**0.0698**	**0.9355**	**0.0632**	**0.9486**	**0.0769**	**0.9167**

Metrics	Params(M)	MACs(G/s)
TS-DOA	0.164	0.147
RTS-DOA	0.121	0.187
RTS-DOA(Large)	1.334	0.738

Metrics	VDE ↓	AR ↑	Params(M)	MACs(G/s)
TS-DOA(Small)	0.1167	0.6973	0.037	0.038
TS-DOA	0.0769	0.8171	0.164	0.147
TS-DOA(Huge)	0.0649	0.8555	4.609	0.730
w/o Enhancement	**0.0558**	**0.9394**	**0.031**	**0.065**

✔ By integrating speech enhancement, speaker recognition, and spatial information modeling, the accurate DOA estimation of the target speaker in multi-speaker scenarios is successfully achieved.

✔ Compared to the baseline, the accuracy of DOA prediction improved by about 30%, while maintaining a compact model with only 0.12M parameters.

Contract: cslzx@mail.imu.edu.cn