Robust Binaural Sound Localisation With Temporal Attention

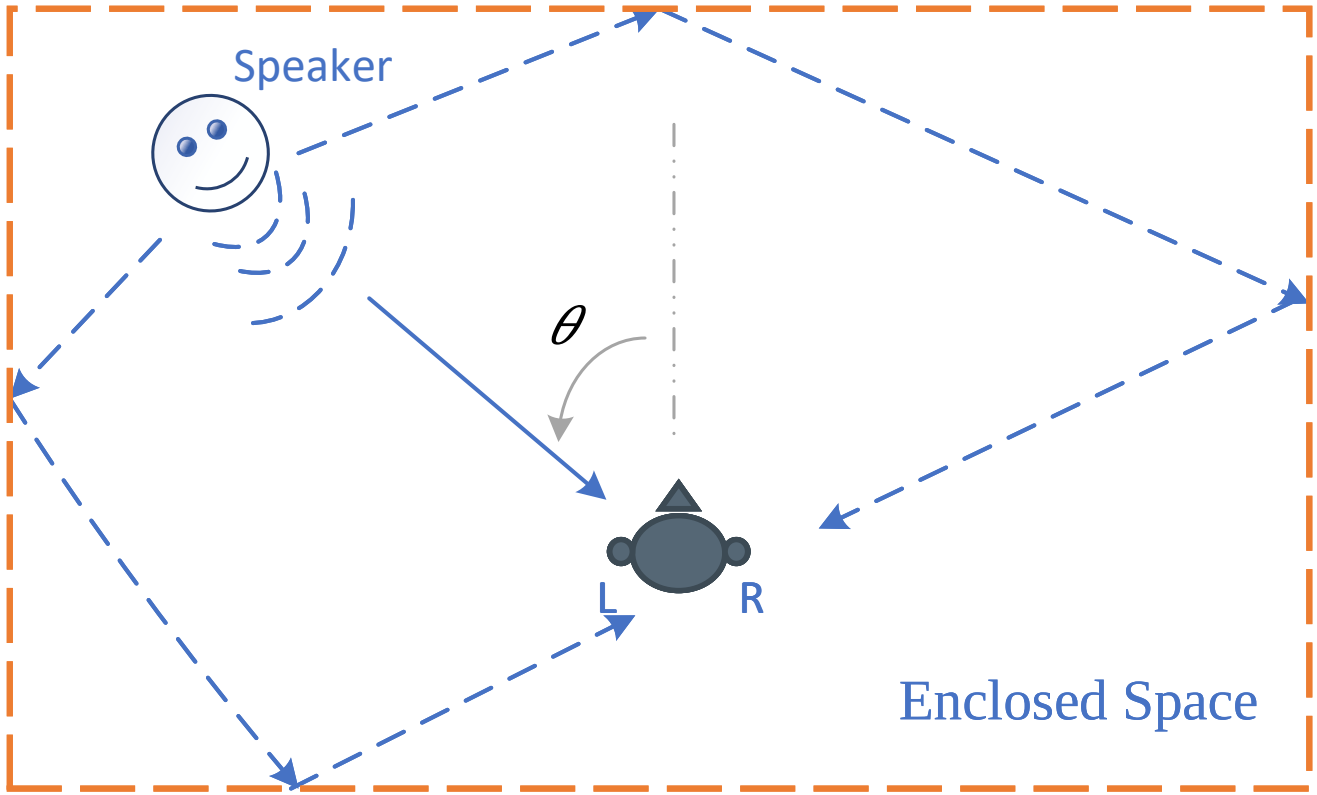
Qi Hu, Ning Ma and Guy J. Brown

Institute of Acoustics of Chinese Academy of Sciences, Beijing, CHINA
University of Sheffield, Sheffield, UK

May 6, 2023

Outline

Introduction: Problem Formulation



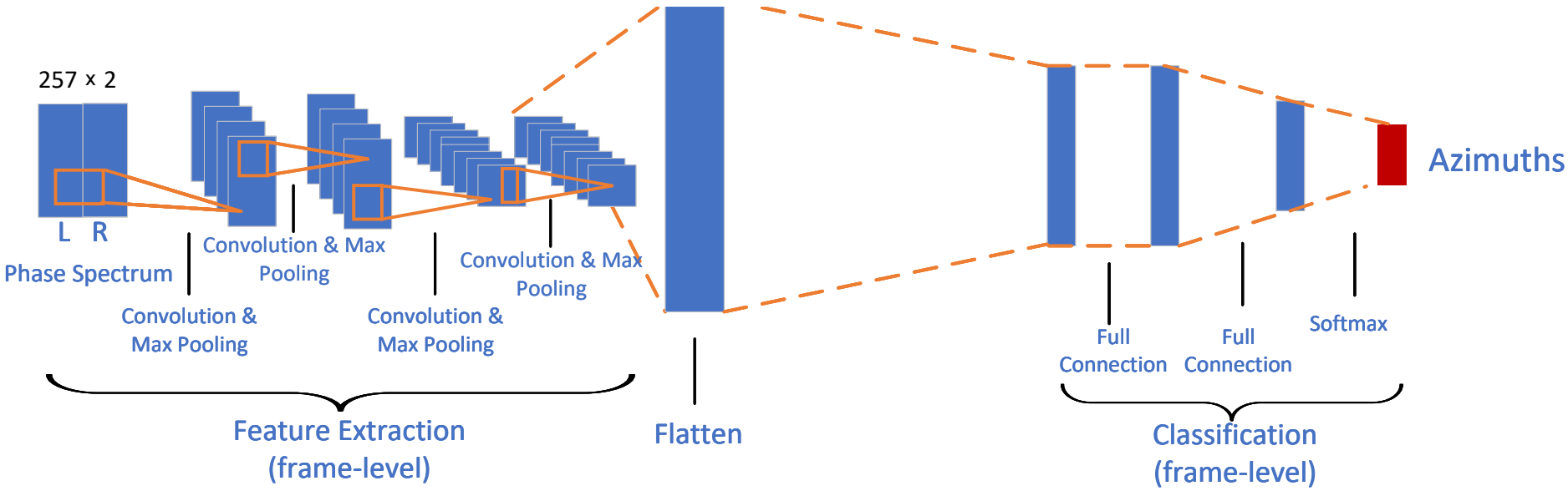
- Noise and reverberation
- Objective: estimating the DOA of the target sound source.

Introduction: Motivation

- Narrow-band antenna signal processing derived methods, being sensitive to noise and reverberations.
  - GCC-PHAT (Generalized Cross Correlation with Phase Transform)
  - SRP (Steered Response Power)
  - Subspace-based Methods
- Deep Neural Network (DNN) based methods, where the front- and back-end processes are decoupled.
  - Spatial Feature Enhancement
  - Spatial Feature Selection using Target Related T-F Masks
  - Robust Improvements on Back-end Localisation Models via Multi-conditional Training (MCT) or Headmovements

Outline

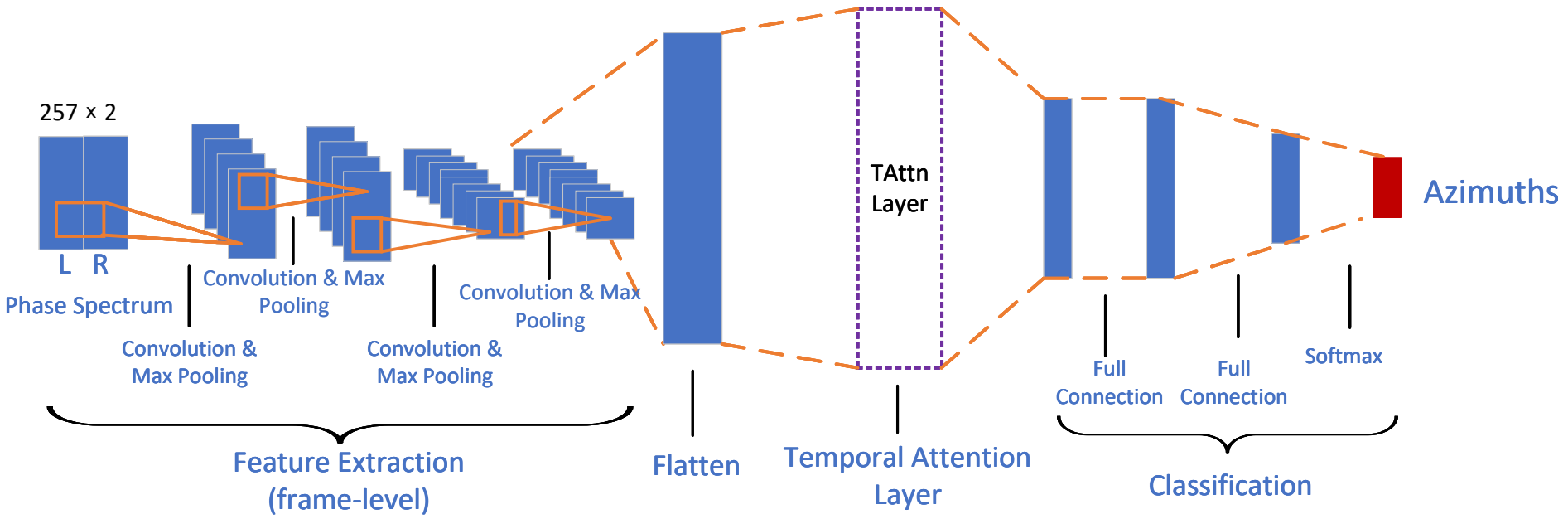
Baseline: Architecture



- Denoting the left and right channels by 'L' and 'R', respectively.
- Using four convolutional layers to extract IPD-like features for sound localisation.
- Estimating the azimuth using a classifier.
- Integrating the frame-level output probabilities of the azimuth estimates through averaging or temporal attention.

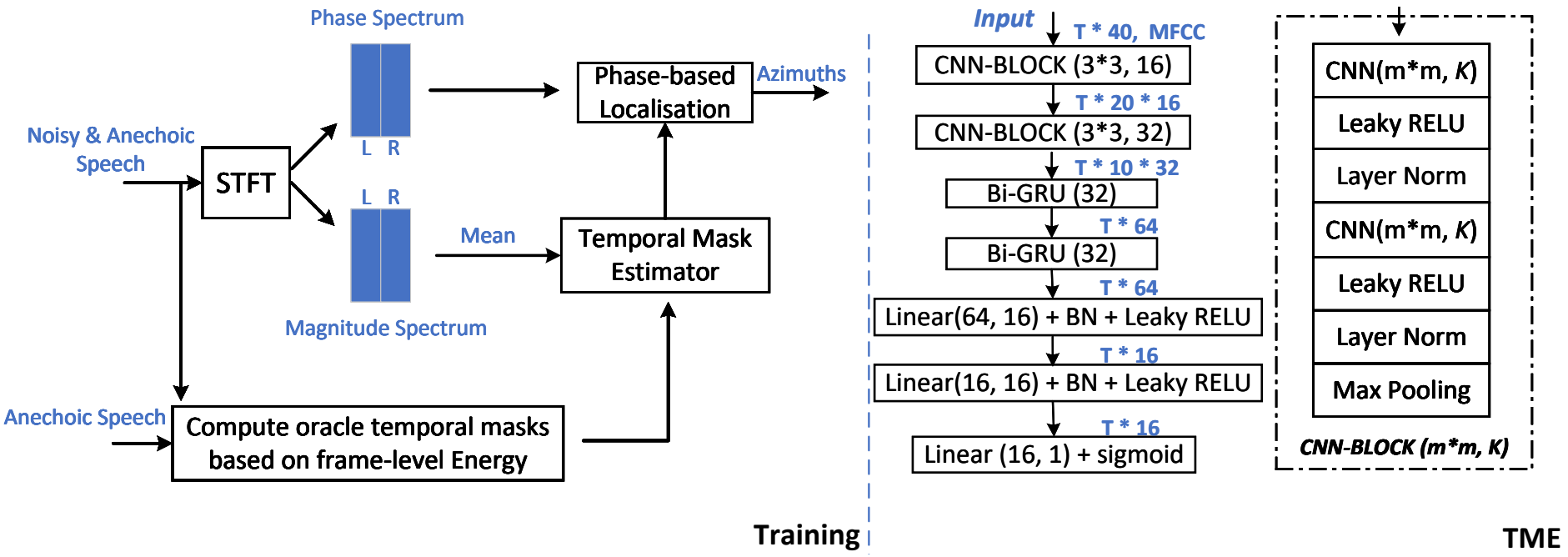
S. Chakrabarty et al., Multi-Speaker DOA Estimation Using Deep Convolutional Networks Trained with Noise Signals. IEEE Journal Of Selected Topics in Signal Processing, 2019.

Proposed Method: Architecture



- Being similar to the baseline.
- Using a 'TAttn layer' to obtain utterance-level features by combining learned features.
- Regarding the azimuth estimation as a classification task.

Proposed Method: Framework



- Estimating the speech dominance within each frame by using a temporal mask estimation (TME) module.
- Training the TME module in a supervised way by using oracle masks as targets and noisy speech as inputs.
- Using Multi-tasking learning to encourage the TME to output an optimal mask for the source localisation task.

Outline

Dataset

The University Of Sheffield.

- Evaluation for a single (speech) source only.
- Monaural speech from the TIMIT, and Binaural signals were created using HRIRs.
- 37 azimuths, across the full 180° azimuth range in steps of 5°.
- 30 random sentences for each of the 37 azimuth locations for creating the training and test sets, respectively.
- For Training, the KEMAR anechoic HRIRs were used to simulate the free field condition; For testing, the Surrey reverberant HRIRs were adpoted.
- *randomly* selected signal-to-noise ratios (SNRs) within $[0, 20]$ dB for training, *fixed* SNRs (i.e., 0, 5, 10, and 20 dB) for testing.

Outline

Results

The University Of Sheffield.

Table: Localisation RMSE results (Lower is better) in degree

	Room A				Room B				Room C				Room D				Avg.
SNR (dB)	20	10	5	0	20	10	5	0	20	10	5	0	20	10	5	0	
GCC-PHAT	4.9	36.1	56.3	60.1	15.4	45.7	55.7	57.7	10.8	40.5	55.4	60.4	15.8	45.0	57.9	64.3	42.6
+ MCT	2.0	5.9	7.0	9.2	1.6	5.4	8.7	13.3	3.2	5.9	7.1	20.3	2.6	5.1	6.3	13.3	7.3
Shallow	3.3	6.1	8.2	13.6	2.7	4.6	7.4	16.1	2.9	4.9	7.2	19.9	3.3	5.4	8.0	19.6	8.3
TAttn–E	**1.6**	**1.8**	5.5	15.3	**1.0**	5.2	**4.8**	15.2	2.2	2.2	3.2	**9.0**	**1.8**	**2.1**	5.1	19.0	5.9
TAttn–J	**1.6**	**1.8**	**2.9**	**7.9**	1.1	**1.6**	5.1	**12.7**	**2.1**	**2.1**	**2.9**	11.8	1.9	**2.1**	**3.8**	**9.0**	**4.4**
TAttn–O	1.6	1.8	2.5	13.0	1.0	1.4	3.2	10.9	2.2	2.2	2.7	6.0	1.8	2.0	2.8	13.3	4.3

Table: Localisation Accuracy (%, Higher is better)

	Room A				Room B				Room C				Room D				Avg.
SNR (dB)	20	10	5	0	20	10	5	0	20	10	5	0	20	10	5	0	
GCC-PHAT	99.4	74.3	41.1	20.6	96.3	59.4	32.7	19.0	97.2	62.9	34.2	17.9	96.0	64.5	35.6	19.9	54.4
+ MCT	99.8	97.8	93.6	85.3	99.5	95.7	92.8	83.3	99.8	97.8	92.2	80.9	99.6	94.9	91.5	81.6	92.9
Shallow	99.7	96.9	90.8	80.4	99.8	96.0	90.3	78.6	99.9	98.3	94.2	75.3	99.8	97.6	90.7	72.1	91.3
TAttn–E	**100**	99.8	97.9	86.4	100	99.8	96.5	82.4	**100**	99.5	97.8	83.2	100	99.5	97.7	78.0	94.9
TAttn–J	**100**	**100**	**98.6**	**89.3**	**100**	**99.9**	**97.7**	**88.7**	**100**	**99.7**	**97.9**	**90.7**	**100**	**99.6**	**98.4**	**90.4**	**96.9**
TAttn–O	100	99.9	98.9	91.3	100	99.9	98.0	88.7	100	99.8	98.5	90.6	100	99.6	98.7	87.0	96.9

Outline

Conclusion

The University Of Sheffield.

- A novel binaural machine hearing system with temporal attention is proposed for robust sound localisation.
- The temporal attention layer integrates frame-level deep features within the localisation DNN by incorporating outputs of an TME module.
- Multi-task learning is adopted to jointly optimise the localisation and the TME module, which improves the system performance, especially in challenging scenarios.

Thanks for your attention!