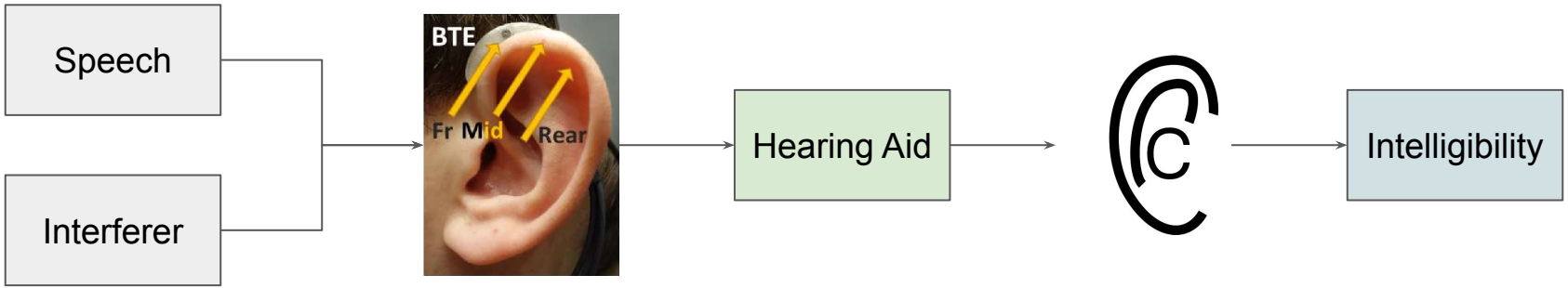
A Two-Stage End-to-End System for Speech-in-Noise Hearing Aid Processing

Zehai Tu, Jisi Zhang, Ning Ma, Jon Barker
Department of Computer Science, University of Sheffield, UK

Background

Clarity recap:

- Interferer: noise and speech
- Six channels (three for each ear)
- Hearing impaired listeners
- Target: maximising intelligibility



Background

Scene data:

- Training: 6000 scenes, 24 speakers
- Development: 2500 scenes, 10 speakers
- Evaluation: 1500 scenes, 6 speakers

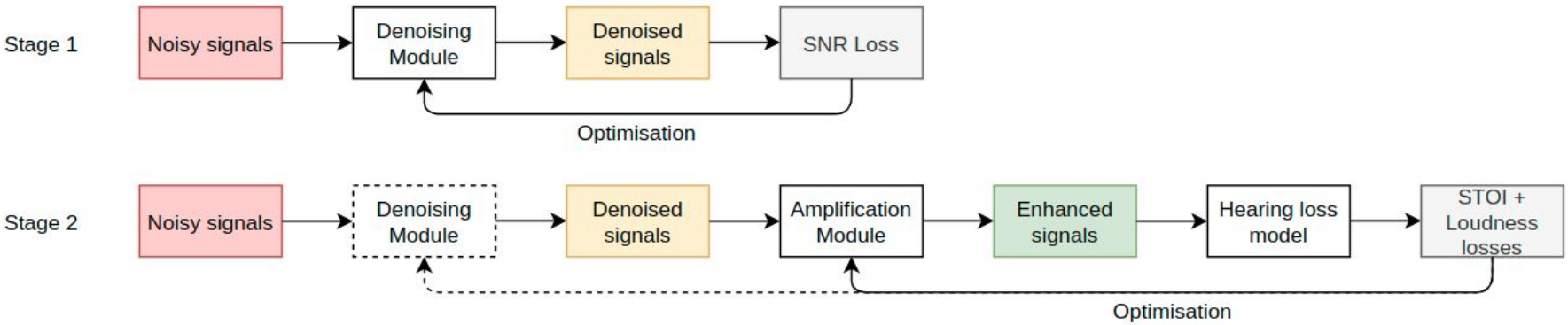
Listener data:

- Pure tone audiograms
- 100 audiograms for training and development
- 50 audiograms for evaluation

Method

Method overview:

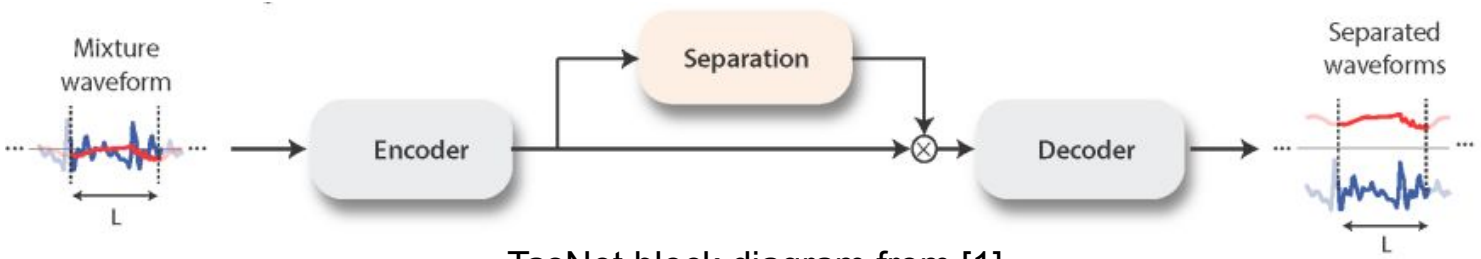
- Stage one: optimising *denoising module*
- Stage two: optimising *amplification module*



Method

Conv-TasNet [1]:

- Designed for single-channel speech separation
- Encoder: 1-D convolution
- Separation: 1-D convolution blocks, skip connections; output masks
- Decoder: 1-D convolution
- Loss: SI-SNR

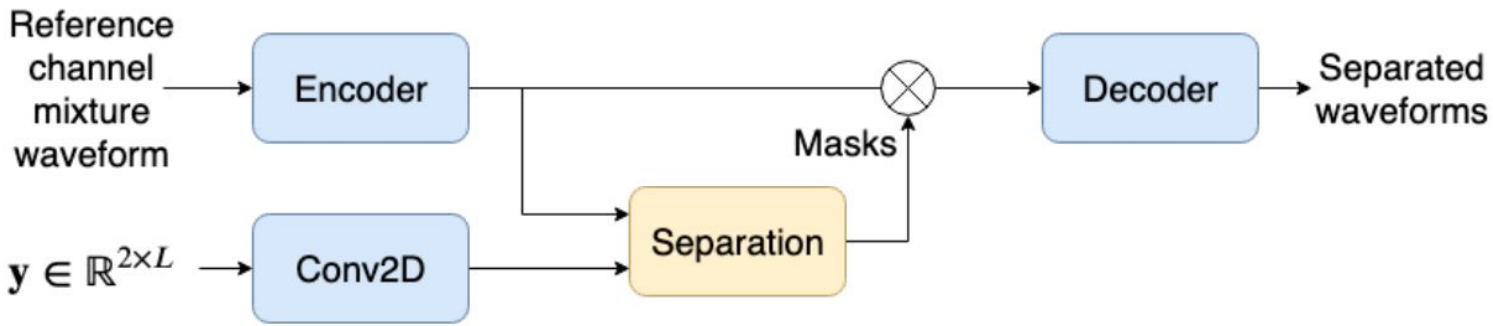


TasNet block diagram from [1]

Method

Multi-channel (MC) Conv-TasNet [2]:

- Spatial encoder (Conv2D) for spatial feature extraction
- Used as the *denoising module*
- Six channels as input to the spatial encoder

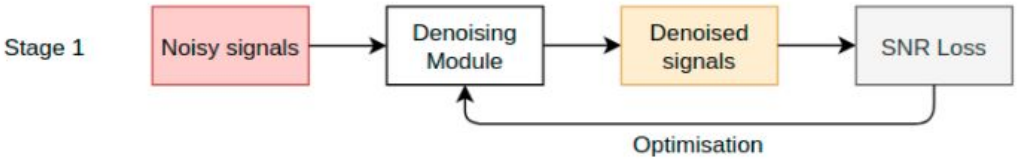


Multi-Channel Conv-TasNet block diagram from [2]

Method

Stage one:

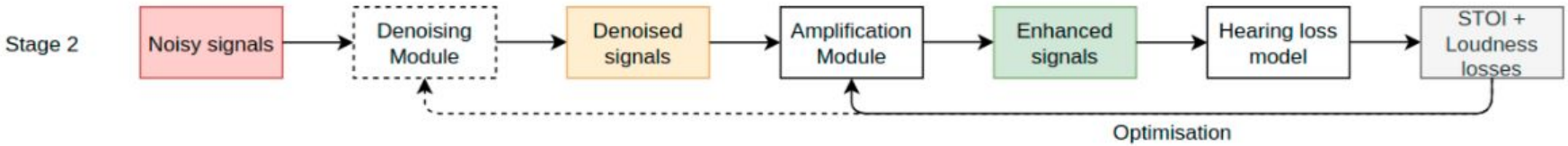
- Optimising *denoising module*: MC-Conv-TasNet (one for left, and one for right)
- Input: six channels, output: single channel (left or right)
- Target: single channel anechoic signal (left or right)
- SNR loss (SPL matters)



Method

Stage two:

- Optimising *amplification module*: Conv-TasNet (E002) or FIR filter (E009)
- Input: single channel, output: single channel (both left or right)
- Hearing loss model: differentiable approximation to MSBG model
- STOI + loudness loss
- Joint optimisation



Results

Initial evaluation (with L0001 only):

Denoising module	Amplification module	Joint optimisation	MBSTOI	DBSTOI
Baseline	Baseline	-	0.414	-
MC-Conv-TasNet	Baseline	-	0.545	0.650
MC-Conv-TasNet	Conv-TasNet	True	0.645	0.836
MC-Conv-TasNet	Conv-TasNet (E002)	False	0.651	0.827
MC-Conv-TasNet	FIR (E009)	False	0.646	0.766

Results

Final objective evaluation (MBSTOI):

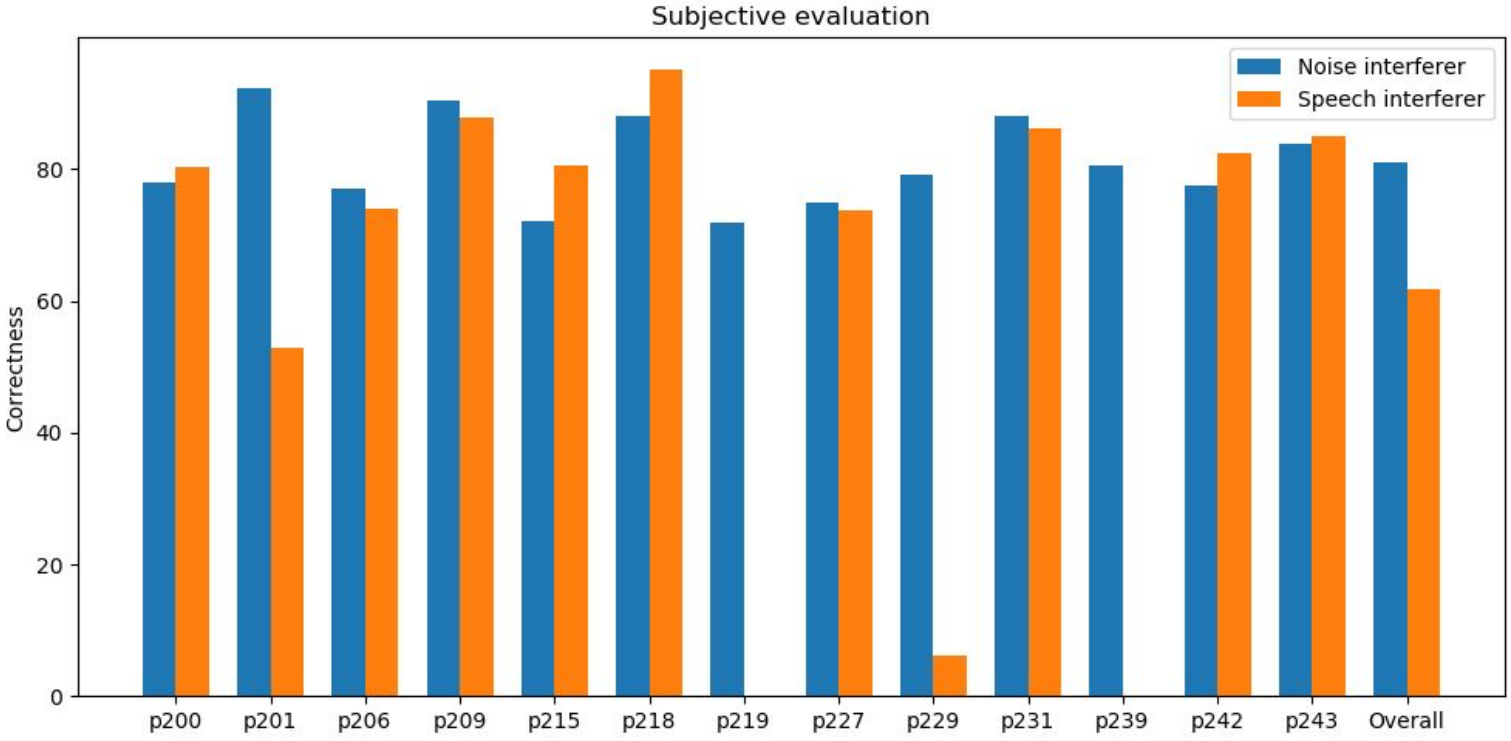
Method	Speech interferer		Noise interferer		Overall	
	Median	Mean	Median	Mean	Median	Mean
Baseline	0.33	0.34	0.28	0.29	0.31	0.31
E002	0.70	0.70	0.67	0.67	0.69	0.69
E009	0.74	0.73	0.69	0.69	0.72	0.71

Results

Final subjective evaluation:

Method	Correctness (per cent)		
	Speech interferer	Noise interferer	Overall
Baseline	43.98	30.30	37.13
E009	61.88	81.03	71.45

Results


Subjective evaluation
Noise interferer
Speech interferer
Correctness
0
20
40
60
80
p200
p201
p206
p209
p215
p218
p219
p227
p229
p231
p239
p242
p243
Overall

References

[1] Luo Y, Mesgarani N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation[J]. IEEE/ACM transactions on audio, speech, and language processing, 2019, 27(8): 1256-1266.

[2] Zhang J, Zorilă C, Doddipatla R, et al. On end-to-end multi-channel time domain speech separation in reverberant environments[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 6389-6393.