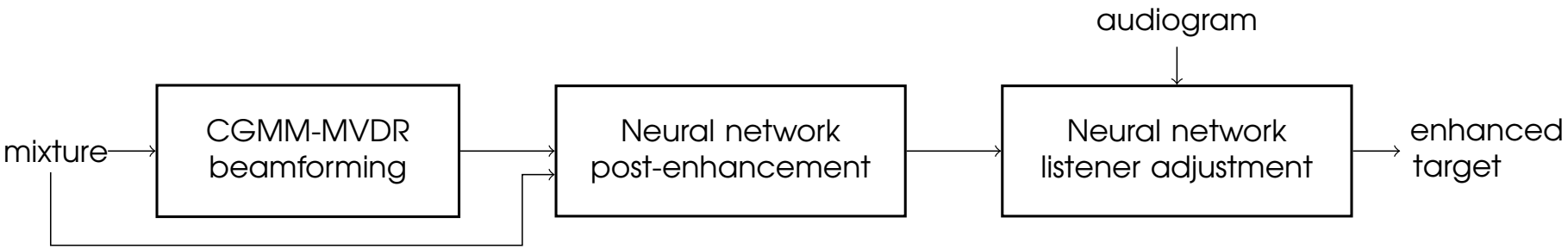
BUT system for the First Clarity Enhancement Challenge

(Submission E007)

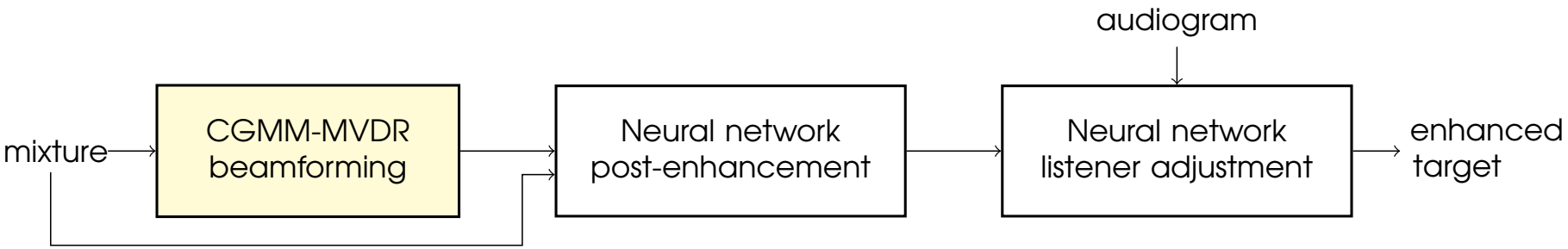
Kateřina Žmolíková, Jan "Honza" Černocký

Overall system



1. CGMM-MVDR Beamforming
   - initial reduction of interference
   - makes use of all available channels
   - makes use of prior information on position and start of speech
2. Neural network post-enhancement
   - further enhancement of the intelligibility
3. Neural network listener adjustment
   - uses audiogram to tailor the output to the listener's hearing loss

Overall system

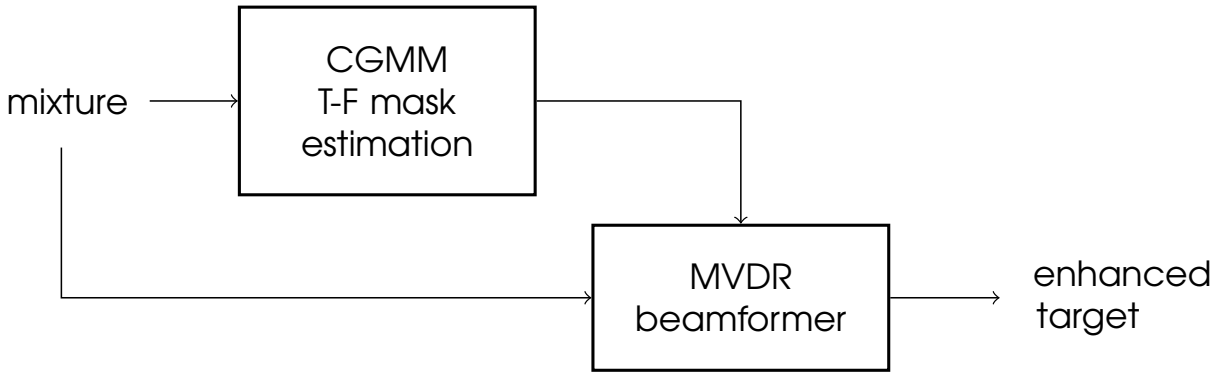


1. CGMM-MVDR Beamforming
   - initial reduction of interference
   - makes use of all available channels
   - makes use of prior information on position and start of speech

2. Neural network post-enhancement
   - further enhancement of the intelligibility

3. Neural network listener adjustment
   - uses audiogram to tailor the output to the listener's hearing loss

STFT representation, 200 samples (∼4.5 ms) with 100 samples shift



Minimum variance distortionless response (MVDR) beamformer

- (Higuchi et al. 2018)
- 6 channels in, 2 channels out

$$\mathbf{w}_{t,f} = \frac{\mathcal{Y}_{f,t}^{-1}\mathcal{R}_{f,t}^{(s)}\mathbf{e}}{\mathrm{tr}(\mathcal{Y}_{f,t}^{-1}\mathcal{R}_{f,t}^{(s)})}$$

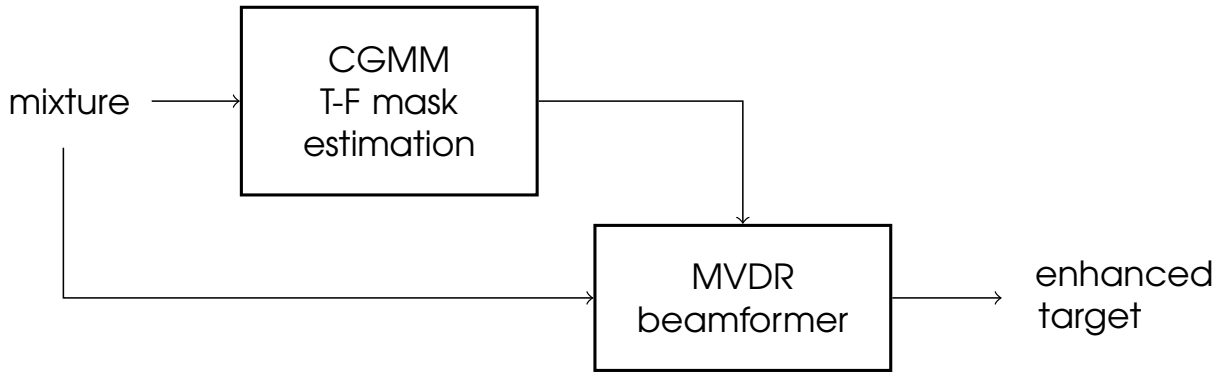
$\mathcal{Y}_{f,t}$ cross-power spectral density of mixture

$$\mathcal{Y}_{f,t} = \mathcal{Y}_{f,t-1} + \mathbf{y}_{f,t}\mathbf{y}_{f,t}^{H}$$

$\mathcal{R}_{f,t}^{(s)}$ cross-power spectral density of target

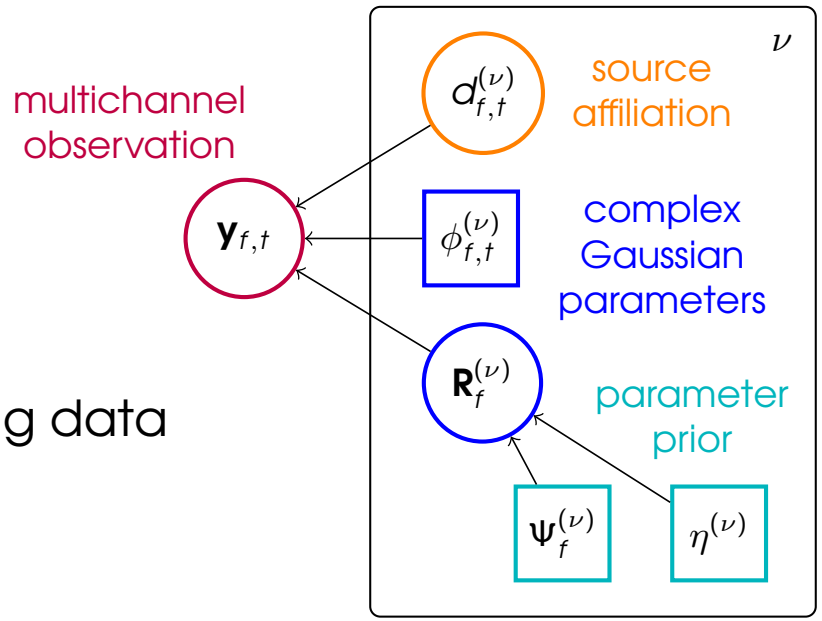
$$\mathcal{R}_{f,t}^{(s)} = \mathcal{R}_{f,t-1}^{(s)} + M_{f,t}^{(s)}\mathbf{y}_{f,t}\mathbf{y}_{f,t}^{H}$$

STFT representation, 200 samples (∼4.5 ms) with 100 samples shift



Complex Gaussian mixture model

- (Higuchi et al. 2017)
- T-F mask := $p(d_{f,t}^{(\nu)} = t \,|\mathbf{y}_{f,t})$
- parameters estimated with E-M
- online estimation
- parameter prior estimated on training data
- first 2 seconds fixed $p(d_{f,t} = t) = 0$



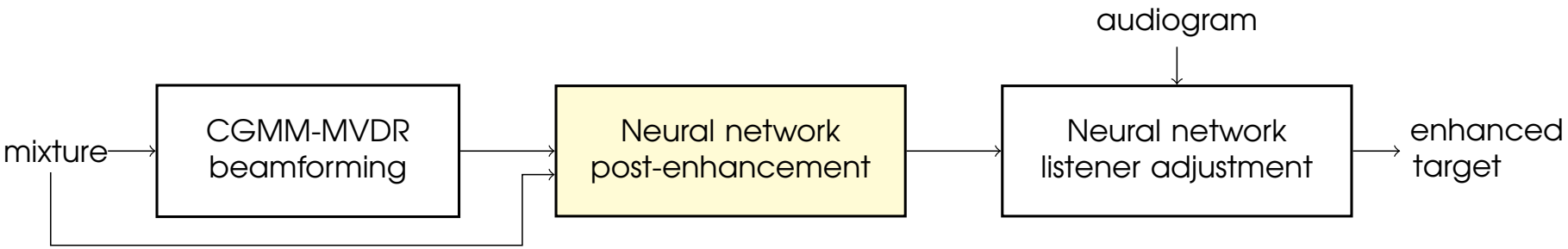
CGMM-MVDR Beamforming: Results

method	MBSTOI (w/o HL)	MBSTOI (w HL)
baseline	-	0.415
CGMM+MVDR	0.707	**0.599**

MBSTOI (w/o HL) comparing enhanced signals with target anechoic signals

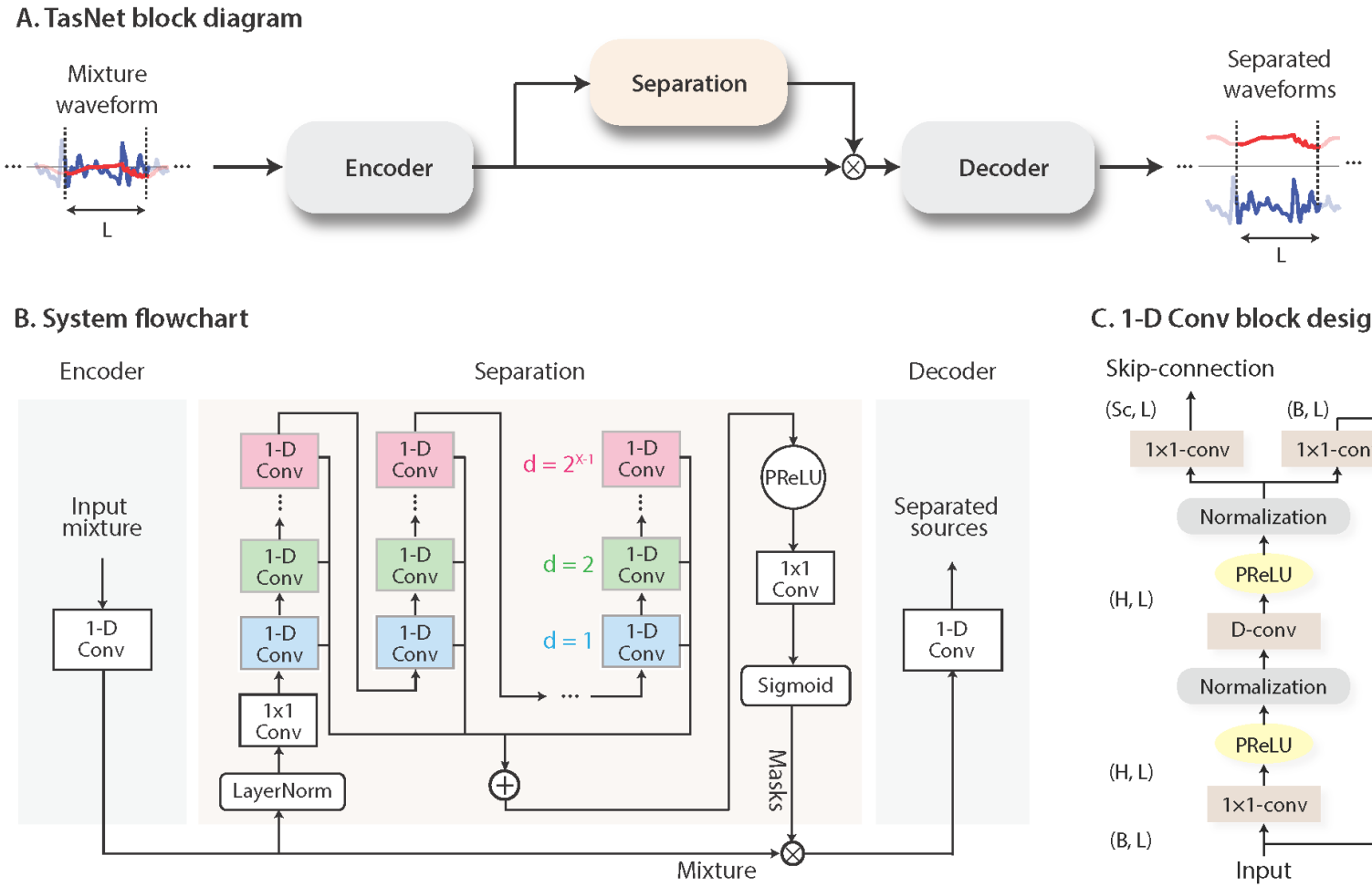
MBSTOI (w HL) comparing enhanced signals *processed by hearing loss model* with target anechoic signals

Overall system



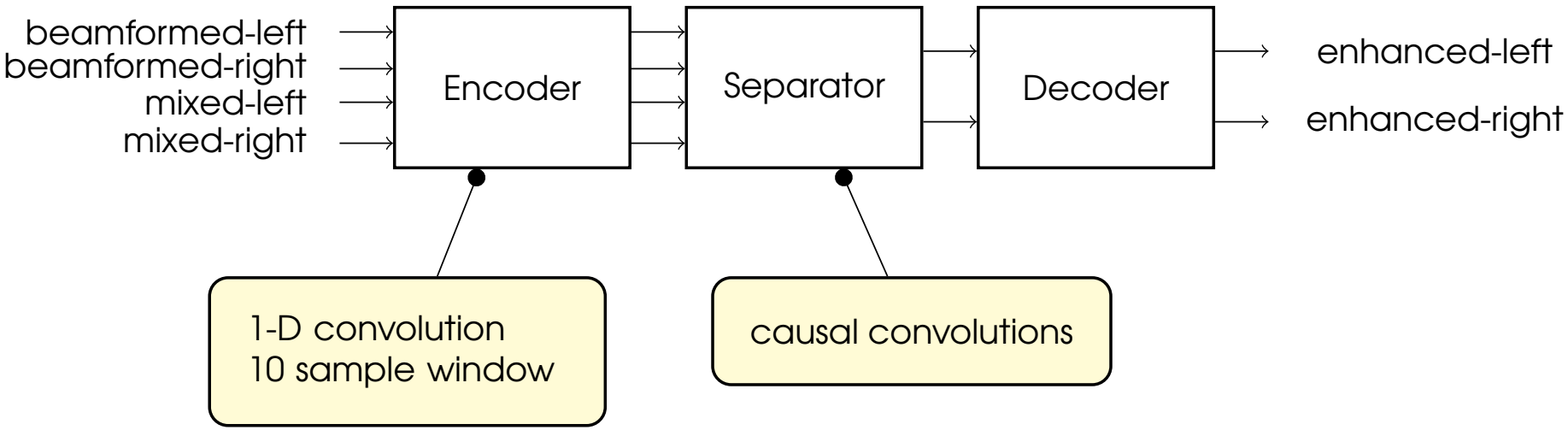
1. CGMM-MVDR Beamforming
   - initial reduction of interference
   - makes use of all available channels
   - makes use of prior information on position and start of speech
2. Neural network post-enhancement
   - further enhancement of the intelligibility
3. Neural network listener adjustment
   - uses audiogram to tailor the output to the listener's hearing loss

ConvTasNet (Luo et al. 2019)



Neural network post-enhancement

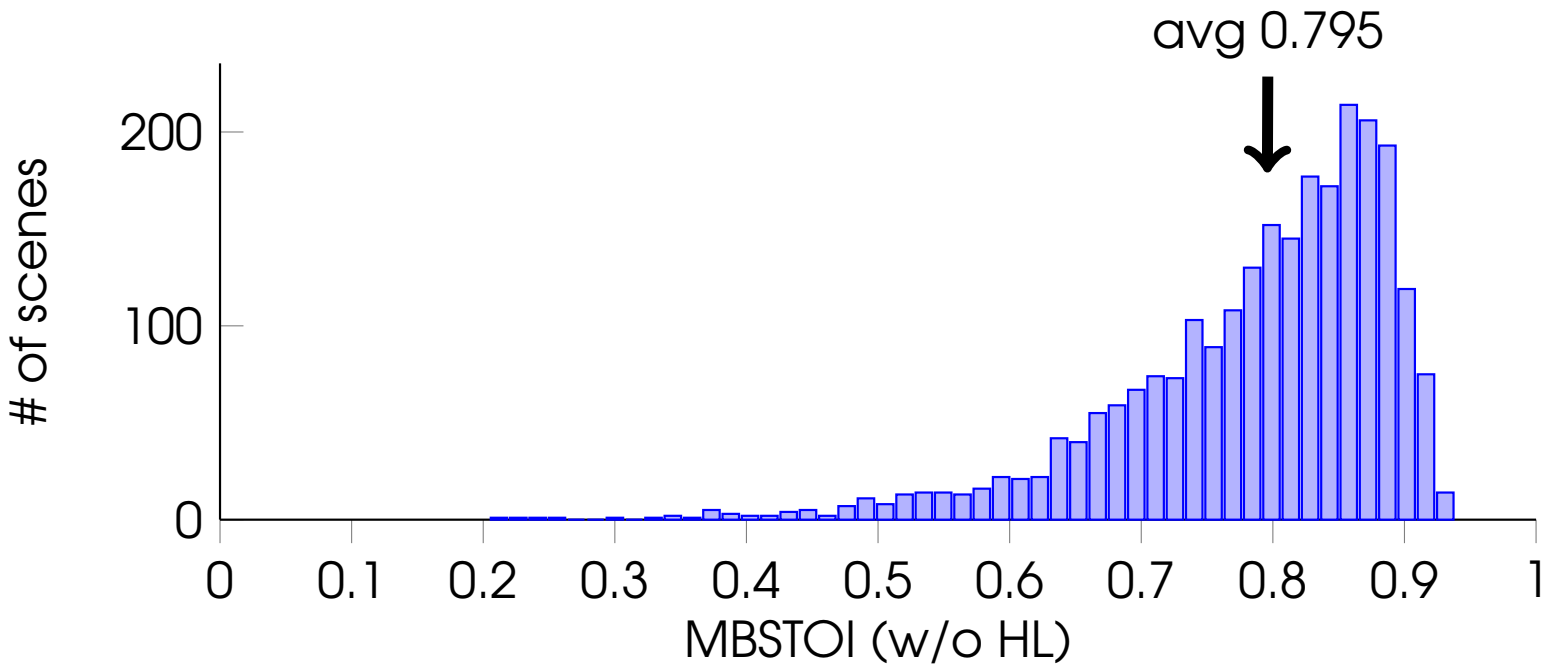
- ConvTasnet architecture (Luo et al. 2019) (Pariente et al. 2020)



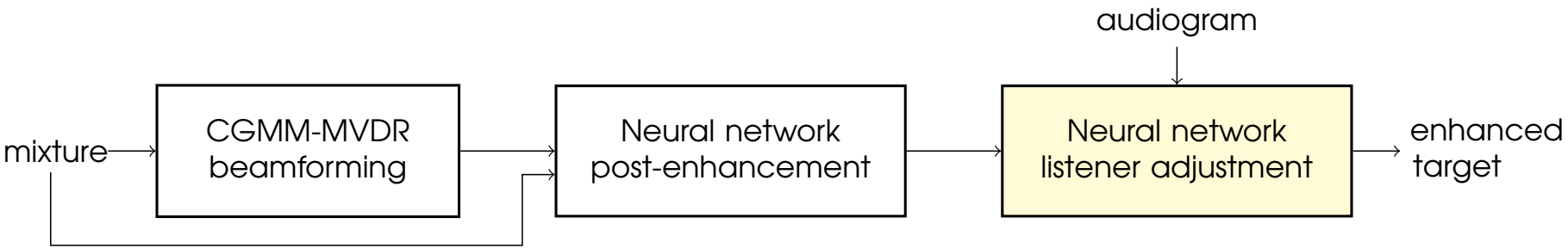
- Inputs: beamformed signals, original mixed signals
- Targets: target anechoic signals
- Objective: 0.9 STOI + 0.05 SNR + 0.05 PMSQE

Neural network post-enhancement

method	MBSTOI (w/o HL)	MBSTOI (w HL)
baseline	-	0.415
CGMM+MVDR	0.707	0.599
NN post-enh (SNR)	0.767	0.631
NN post-enh (bf-only)	0.770	0.635
NN post-enh	**0.795**	**0.657**



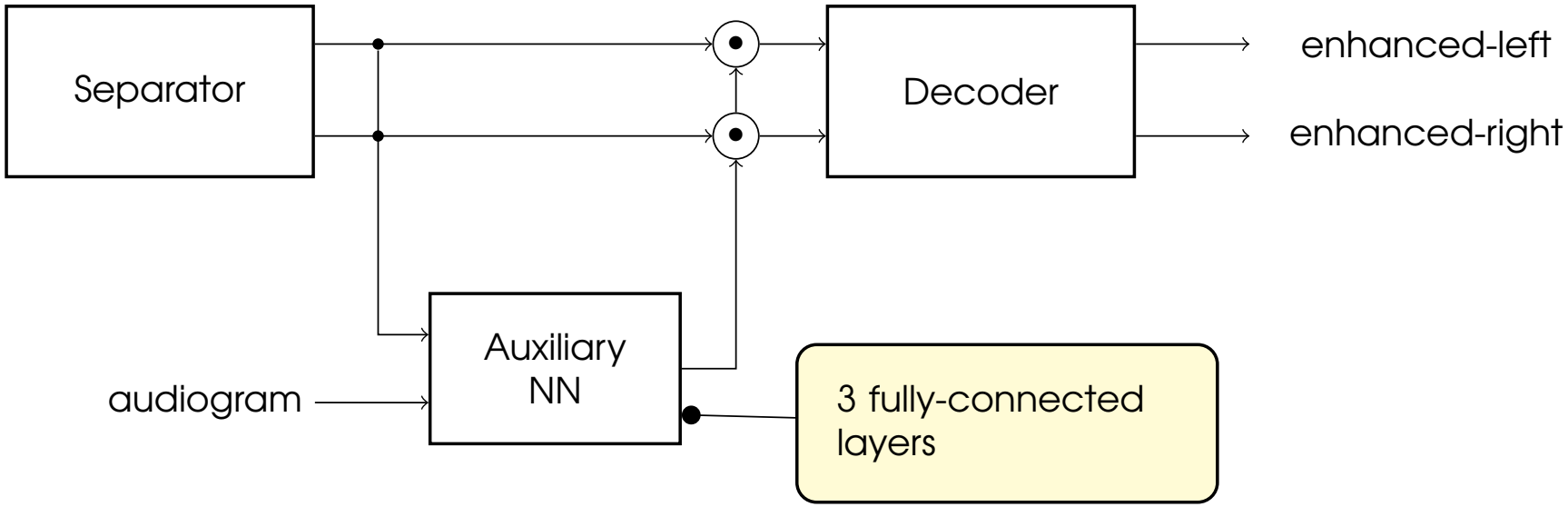
Overall system



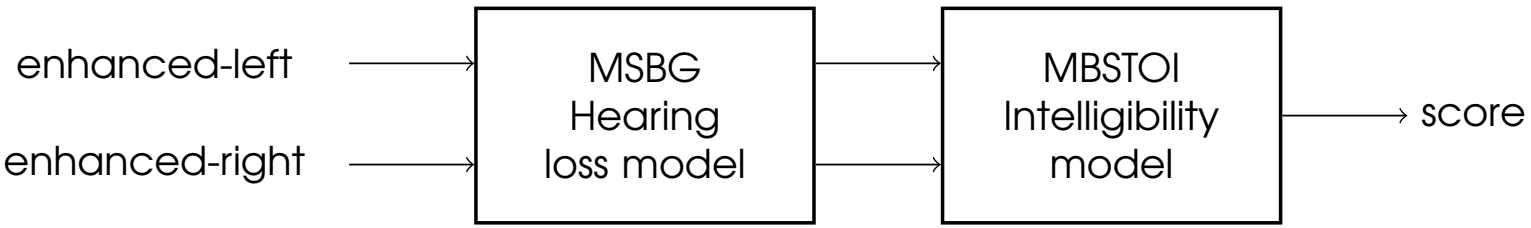
1. CGMM-MVDR Beamforming
   - initial reduction of interference
   - makes use of all available channels
   - makes use of prior information on position and start of speech
2. Neural network post-enhancement
   - further enhancement of the intelligibility
3. Neural network listener adjustment
   - uses audiogram to tailor the output to the listener's hearing loss

Neural network listener adjustment

- gain on estimated encoded representations
- estimated by auxiliary neural network from audiogram
- only auxiliary network trained



Objective function



Neural network listener adjustment

method	MBSTOI (w/o HL)	MBSTOI (w HL)
baseline	-	0.415
CGMM+MVDR	0.707	0.599
NN post-enh	**0.795**	0.657
NN listener (random)	0.759	0.662
NN listener	0.759	**0.671**

?

- Can we learn something better than simple compression?
- Dynamic compression based on the input content?
- Adjustment to better hearing loss models?

Conclusions

- overall improvement over the baseline 0.415 $\rightarrow$ 0.671 MBSTOI
- preliminary results of listening tests 37.14% $\rightarrow$ 64.54%

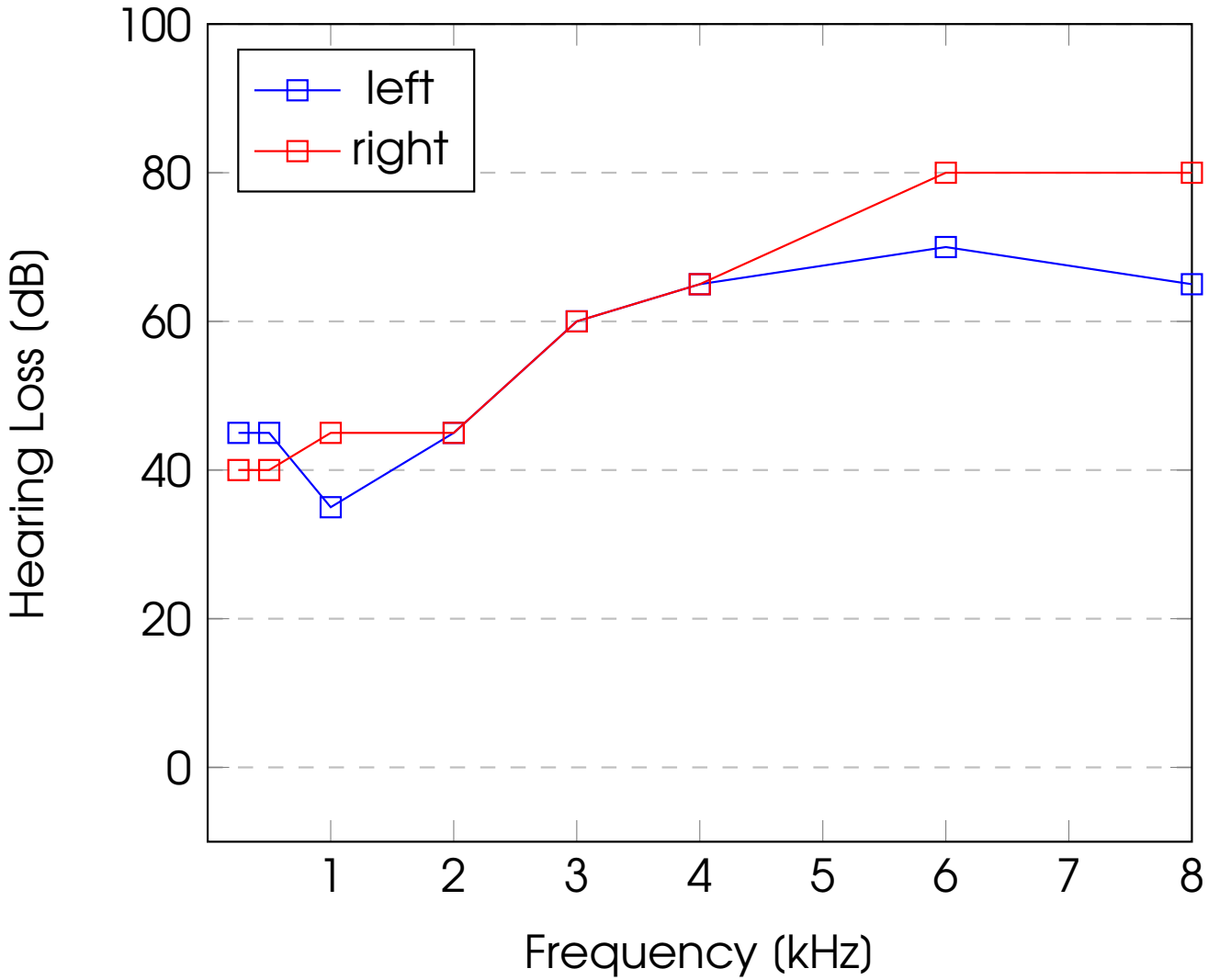
- CGMM-MVDR & NN-post-enh suppress interference well
- more things to explore about adjustment to listeners
  - other hearing loss/intelligibility models to better examine the real impact
  - more analysis of what was learned
  - improving hearing loss model to be better "optimizable"

References I

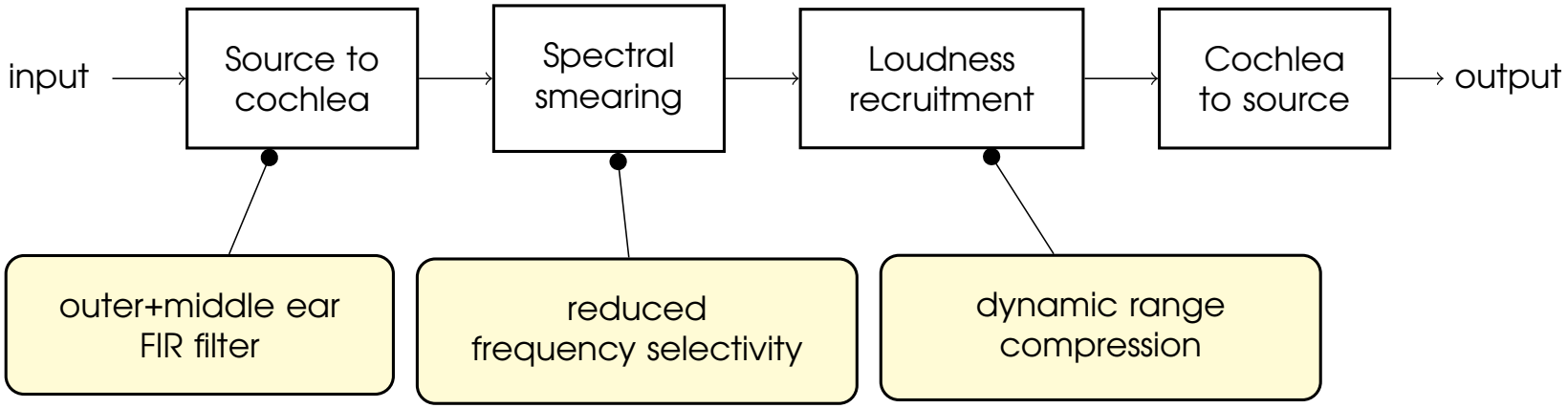
- Andersen, Asger Heidemann et al. (2018). "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions". In: Speech Communication 102, pp. 1–13.
- Baer, Thomas et al. (1993). "Effects of spectral smearing on the intelligibility of sentences in noise". In: The Journal of the Acoustical Society of America 94.3, pp. 1229–1241.
- – (1994). "Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech". In: The Journal of the Acoustical Society of America 95.4, pp. 2277–2280.
- Higuchi, Takuya et al. (2017). "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR". In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 25.4, pp. 780–793.
- Higuchi, Takuya et al. (2018). "Frame-by-frame closed-form update for mask-based adaptive MVDR beamforming". In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 531–535.

- Luo, Yi et al. (2019). "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation". In: IEEE/ACM transactions on audio, speech, and language processing 27.8, pp. 1256–1266.
- Moore, Brian CJ et al. (1993). "Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech". In: The Journal of the Acoustical Society of America 94.4, pp. 2050–2062.
- Pariente, Manuel et al. (2020). "Asteroid: the pytorch-based audio source separation toolkit for researchers". In: arXiv preprint arXiv:2005.04132.
- Stone, Michael A et al. (1999). "Tolerable hearing aid delays. I. Estimation of limits imposed by the auditory path alone using simulated hearing losses". In: Ear and Hearing 20.3, pp. 182–192.

Extra: Audiogram



(Baer et al. 1993; Baer et al. 1994; Moore et al. 1993; Stone et al. 1999)



(Andersen et al. 2018)