Binaural Speech Enhancement Based on Deep Attention Layers

Tom Gajecki & Waldo Nogueira

Auditory Prosthetic Group, Department of Otolaryngology, Hannover Medical School, Cluster of Excellence, Hannover, Germany

Submission to the Clarity Challenge

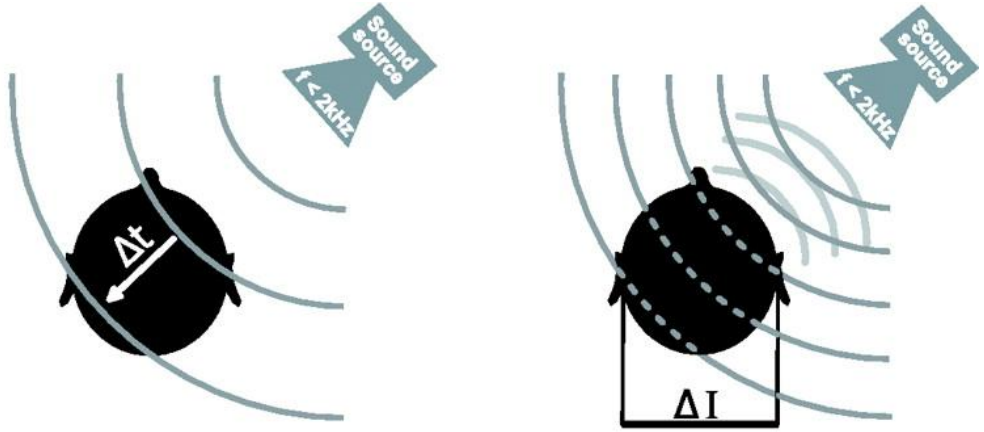
Introduction

- The speech enhancement system is a binaural end-to-end audio source separation framework that connects information between hearing aids

- We investigate the effect of applying attention mechanisms to improve binaural audio source separation

Binaural Hearing

- The ability to process the information available in pressure waves arriving at the two ears is called binaural hearing
- Binaural unmasking is the ability that the binaural hearing system has to enhance target sound sources when spatially separated from interferers [Ira J. Hirsh, 1948]
- The binaural system exploits two binaural cues:
  - Interaural time differences (ITDs)
  - Interaural level differences (ILDs)



Binaural Impairment

- Hearing loss can cause a degradation in the ability of exploiting binaural cues [S. C. Hogan, D. R. Moore, 2003]
- Binaural impairment can cause :
  - Poor speech intelligibility performance in noisy conditions when compared with a healthy hearing system
  - Poor sound localization abilities when compared to a healthy hearing system

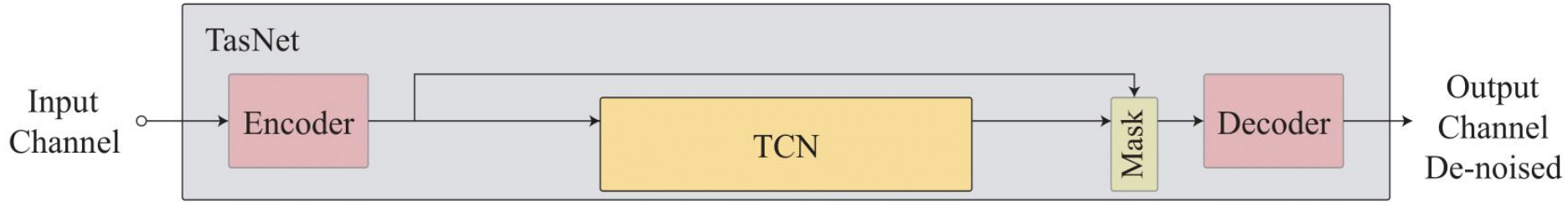
Motivation

Hearing 4all
APG Auditory Prosthetic Group
Medizinische Hochschule Hannover

- We aim at developing a speech enhancement system that uses inputs from both hearing sides
- The system should share information between sides, inspired by the binaural system, to exploit binaural cues
- We propose a deep neural network architecture which shares information between sides through intermediate layers that we will refer to as “attention layers“, inspired by other successful attention mechanisms used in other domains [A. Vaswan, 2017]
- We hypothesize that an algorithm that uses attention layers will yield better objective performance compared to a system that does not use them

Methods: The architecture

- We propose a binaural speech enhancement system built upon a well known single-channel speech separation algorithm: TasNet [Y. Luo and N. Mesgarani, 2019]

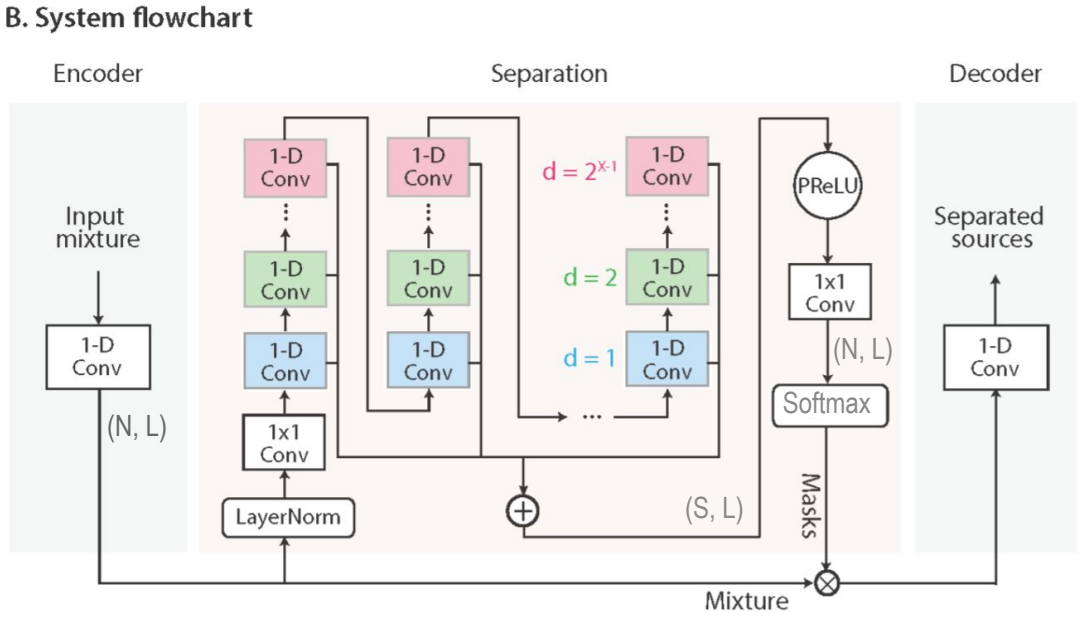


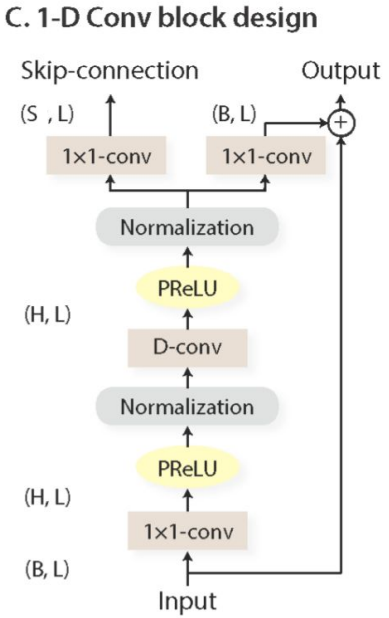
- The encoder maps a segment of the mixture waveform to a high-dimensional representation
- The temporal convolution network (TCN) calculates a multiplicative function (i.e., a mask) for the desired target source
- The decoder decoder reconstructs the source waveforms from the masked features

Methods: The architecture

- We propose a binaural speech enhancement system built upon a well known single-channel speech separation algorithm: TasNet [Y. Luo and N. Mesgarani, 2019]

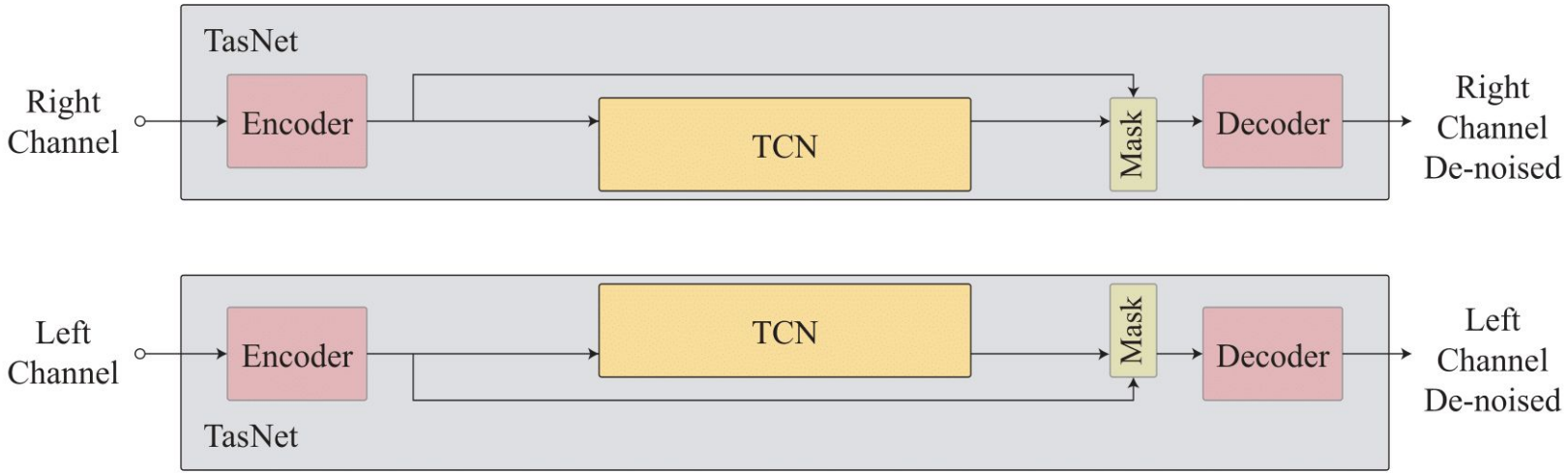
Symbol	Description
N	Number of filters in autoencoder
L	Length of the filters (in samples)
B	Number of channels in bottleneck and the residual paths' 1×1-$conv$ blocks
S	Number of channels in skip-connection paths' 1×1-$conv$ blocks
H	Number of channels in convolutional blocks
P	Kernel size in convolutional blocks
X	Number of convolutional blocks in each repeat
R	Number of repeats





Methods: The architecture

Independent model



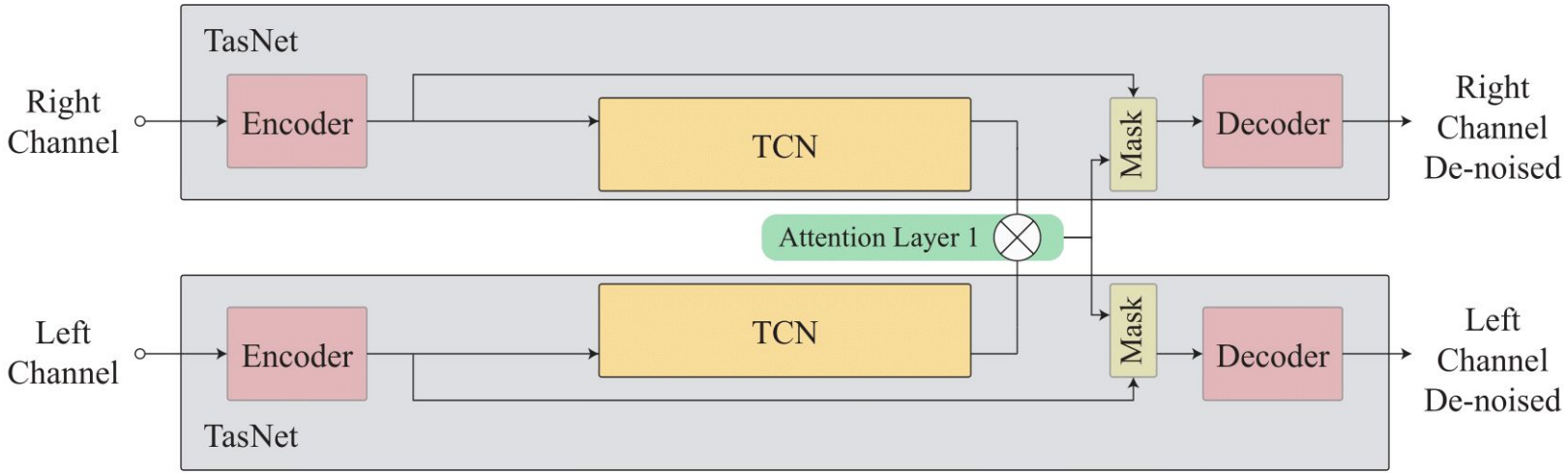
Methods: The architecture

Medizinische Hochschule Hannover

Auditory Prosthetic Group

Hearing 4all

Single Attention model



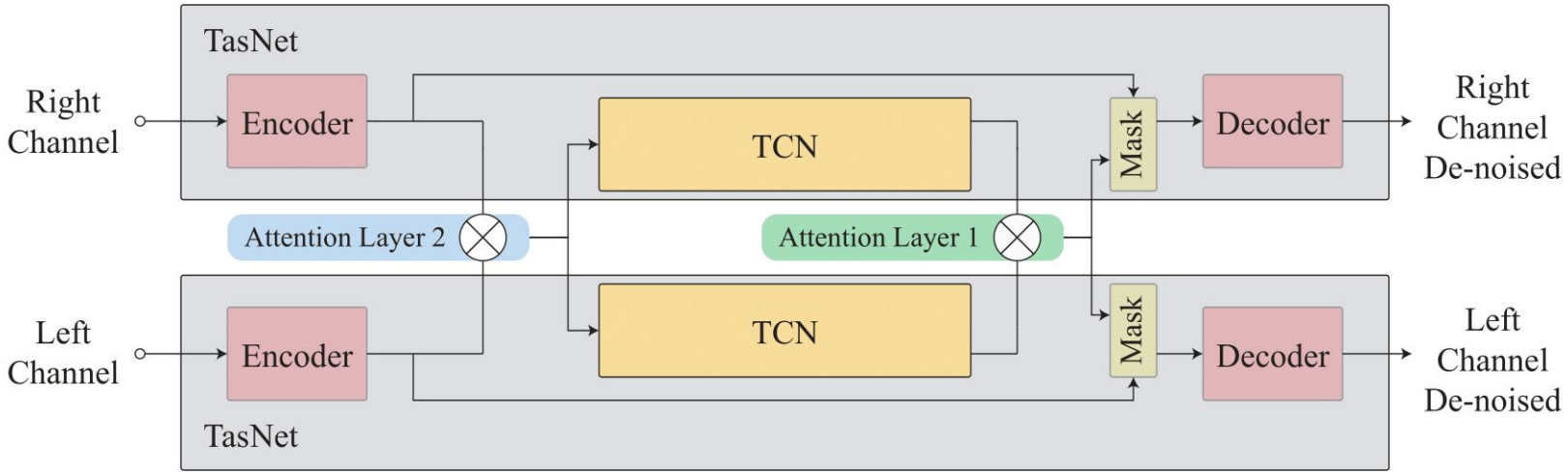
Methods: The architecture

Medizinische Hochschule Hannover

Auditory Prosthetic Group

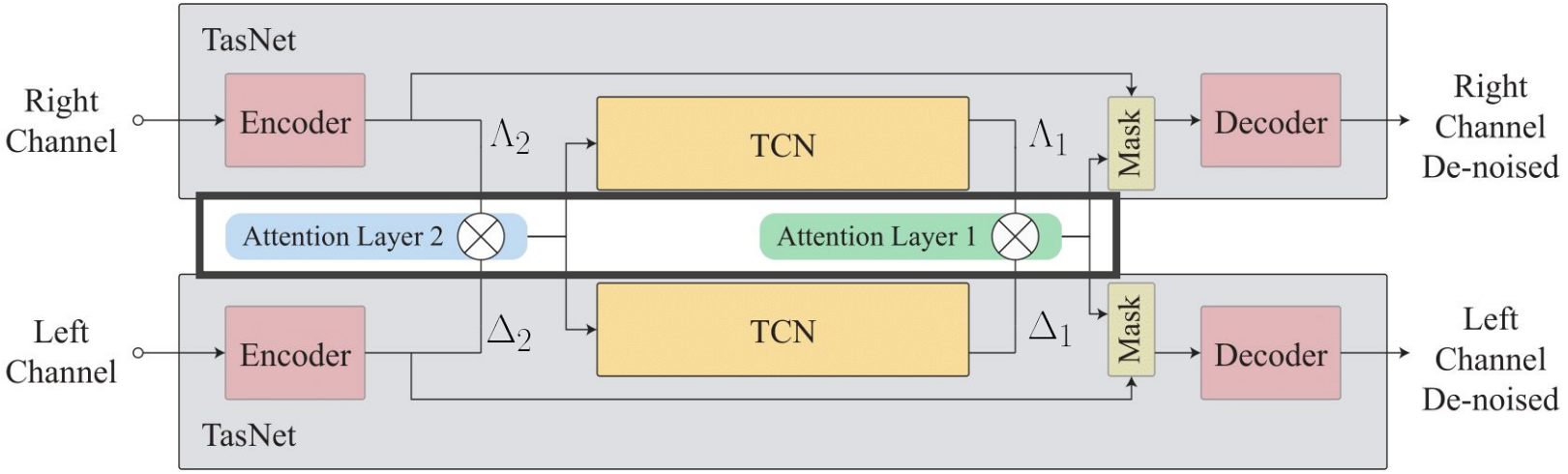
Hearing 4all

Double Attention model



Methods: Attention layers

- Intermediate layers that perform element wise dot product to the latent representations of the signals in each side: $Attention(\Lambda, \Delta) = \Lambda \otimes \Delta$
- They do not introduce trainable parameters



Source code implemented in TensorFlow [M. Abadi et.al., 2015] available online at: https://github.com/APGDHZ/BinAttSE

Methods: Hyperparameters

Description	Value
Number of filters in autoencoder	64
Length of the filters	16
Number of channels in the bottleneck blocks	64
Number of channels in the skip-connections	S
Number of channels in the convolutional blocks	64
Kernel size in convolutional blocks	128
Number of convolutional blocks in each repeat	2
Number of repeats	2

Table 1: *Hyperparameters used for training the models. The parameter that corresponds to the size of the attention layers (S) is a factor that is investigated in this work and its value is variable (refer to sections 2.2 and 3).*

Methods: Hyperparameters

Medizinische Hochschule Hannover

Hearing 4all

Auditory Prosthetic Group

Description	Value
Number of filters in autoencoder	64
Length of the filters	16
Number of channels in the bottleneck blocks	64
Number of channels in the skip-connections	S
Number of channels in the convolutional blocks	64
Kernel size in convolutional blocks	128
Number of convolutional blocks in each repeat	2
Number of repeats	2

Table 1: *Hyperparameters used for training the models. The parameter that corresponds to the size of the attention layers (S) is a factor that is investigated in this work and its value is variable (refer to sections 2.2 and 3).*

The input filter size of 16 samples causes an algorithmic latency of 2ms for an input signal sampled at 8kHz

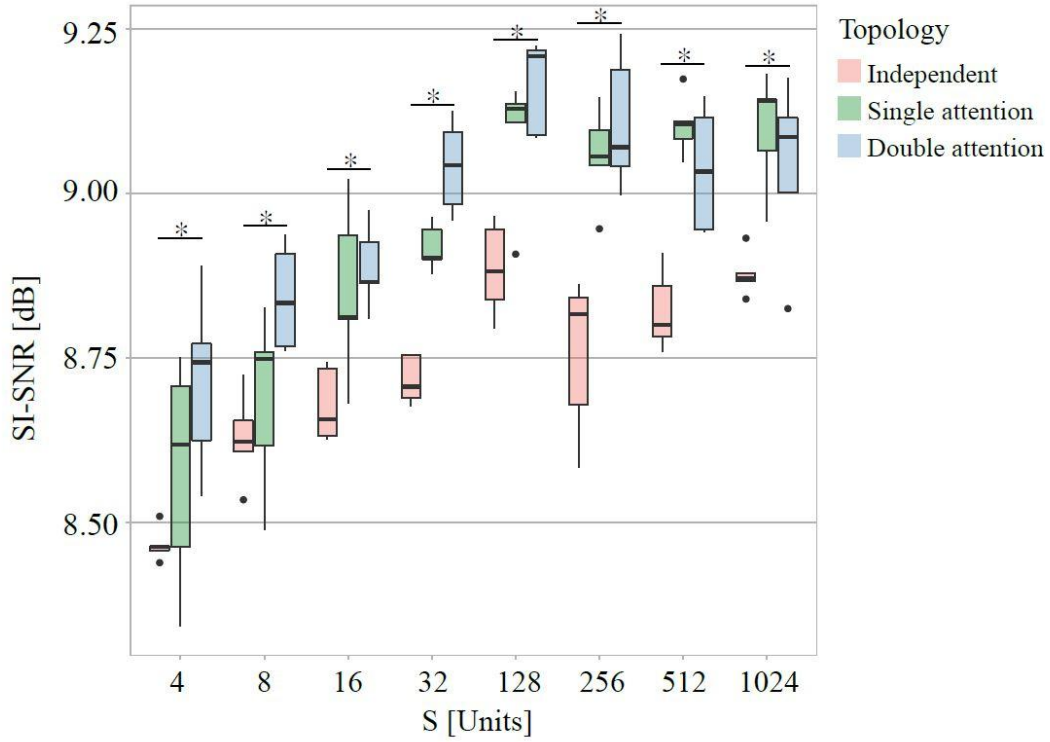
Methods: Training

Hearing 4all
APG Auditory Prosthetic Group
Medizinische Hochschule Hannover

- 8 Model sizes were selected for training (S = {4,8,16,32,128,256,512,1024}) to investigate the effect of the attention/skip connection size
- Each model was trained 5 times to account for variance related to random initialization
- The models were trained for a maximum of 100 epochs using early stopping with a patient of 5 epochs looking at the validation set
- 4-second long audio sections corresponding to the front mic, sampled at 8kHz were used for training
- The learning rate was halved if the accuracy of the validation set did not improve during 3 consecutive epochs
- For the optimization, Adam [D. P. Kingma et.al., 2015] was used to maximize the scale invariant source-to-noise-ratio (SI-SNR) [J. L. Roux et.al., 2019]

Results: SI-SDR

- SI-SNR as a function of the number of attention layers and attention size



Results: Validation MBSTOI

- MBSTOI [A. H. Andersen et.al., 2018] scores as a function of the number of attention layers and attention size for the validation dataset

S	Validation MBSTOI			
	Baseline	Ind.	Single att.	Double att.
-	0.41	-	-	-
4	-	0.70	0.61	0.71
8	-	0.63	0.65	0.71
16	-	0.61	0.65	0.68
32	-	0.67	0.62	0.61
128	-	0.61	0.57	0.62
256	-	0.64	0.65	0.65
512	-	0.64	0.66	**0.77**
1024	-	0.65	0.65	0.69

Table 2: *Maximum validation MBSTOI for all of the tested algorithms. Bold value indicates the best performing algorithm configuration.*

Results: Evaluation MBSTOI

- MBSTOI [A. H. Andersen et.al., 2018] scores as a function of the number of attention layers and attention size for the evaluation dataset, for different interferer types [I. Demirsahin et.al., 2020]

Interferer	Evaluation MBSTOI	
	Baseline	Submitted Algorithm
Speech	0.34	**0.55**
Noise	0.29	**0.48**

Table 3: *Evaluation MBSTOI for the baseline system and the submitted algorithm, for different interferer types.*

Results: Speech Tests

- Due to a suboptimal individualized hearing loss compensation our algorithm obtained worse performance than the baseline provided

Conclusions

- In this work we present a binaural speech enhancement method based on deep binaural attention layers
- Noise reduction amount seems to be proportional to the number of attention layers, being significantly higher using a double attention system when compared to a non-attentive one
- We observed a threshold above which larger attention size yields poorer performance
- Future work should investigate other attention mechanisms such a additive attention to reduce computational cost

Medizinische Hochschule Hannover

Hearing 4all

Auditory Prosthetic Group

Questions?

Ira J. Hirsh , "The Influence of Interaural Phase on Interaural Summation and Inhibition", The Journal of the Acoustical Society of America 20, 536-544, 1948.

S. C. Hogan, D. R. Moore. Impaired binaural hearing in children produced by a threshold level of middle ear disease. J Assoc Res Otolaryngol. 2003.

Y. Luo and N. Mesgarani, “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, pp. 1256–1266, 2019.

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Man´e, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015, software available from tensorflow.org.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H.Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

I. Demirsahin, O. Kjartansson, A. Gutkin, and C. Rivera, “Open-source Multi-speaker Corpora of the English Accents in the British Isles,” in Proceedings of The 12th Language Resources and Evaluation Conference (LREC). Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 6532– 6541.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, Eds., 2015.

J. L. Roux, S.Wisdom, H. Erdogan, and J. R. Hershey, “Sdr – half-baked or well done?” in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 626–630.

A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, “Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions,” Speech Communication, vol. 102, no. C, pp. 1–13, 2018.