Binaural Speech Enhancement Based on Deep Attention Layers

Tom Gajecki & Waldo Nogueira

Department of Otolaryngology, Medical University Hannover and Cluster of Excellence Hearing4all, Hannover, 30625, Germany.

gajecki.tomas@mh-hannover.de

Abstract

Here we describe our submission for the 1st Clarity Enhancement Challenge. The algorithm that we present is based on two conv-TasNets and combines the information contained on each of the listening sides to provide the model with potential binaural cues. This information is combined through intermediate layers that we will refer to as "attention layers", inspired by the classical attention layers used in sequence to sequence modeling. The implemented model is fed with stereo signals and outputs its de-noised version with 2 ms latency. Results show that attention layers can improve the signal-to-distortion ratio, and could further improve speech intelligibility scores.

Index Terms: binaural speech enhancement, deep neural networks, attention layers

1. Introduction

This short report describes a submission for the 1st Clarity Enhancement Challenge [1]. The designed system is based on two conv-TasNets [2] and combines binaural information through intermediate layers which will be referred to as "attention layers". We present an analysis to assess the actual effect these attention layers have on the models' performance and the reason why we selected the submitted model for final evaluation. Figure 1 gives an overview of the complete system. The following section gives more details on the present submission.

2. Methods

2.1. End-to-end Speech Enhancement

The main speech enhancement algorithm is based on two conv-TasNets [2] and consists of three processing stages, as shown in Figure 1: an encoder, a causal dilated 1D temporal convolution network (TCN), and a decoder. The encoder creates a latent representation of the input audio signal, used to estimate a mask for each time step. The TCN then acts as a separator and the de-noised audio is resynthesized by the decoder module. The model was implemented in Tensorflow 2.0 [3] and the code for training and evaluating it can be found online[1].

2.2. Attention Layers

The main aspect we aim at investigating in this study is the effect that sharing information between listening sides has on the models' performance. We propose to share this information by means of attention layers, inspired by the classical attention layers used in sequence to sequence modeling [4]. These layers apply dot-product attention to each channel of the latent representation at specific stages of the processing, as shown in Figure 1. Specifically, let Λ and $\Delta \in \mathbb{R}^{C\times T\times S}$ be the left and right latent representations (on each of the listening sides) at a given processing stage, respectively. Here, C is the number of channels to be enhanced (i.e. one per hearing side), T is the number of time steps of the encoded signal, and S is the number of channels in the latent representation. We compute the attention operation as follows:

$$Attention(\Lambda, \Delta) = \Lambda \otimes \Delta. \tag{1}$$

To investigate how the attention operation affects the models' performance, we tested three configurations; one with no attention layers ("Independent"), one with only one attention layer, after the TCN (i.e., attention layer 1 in Figure 1; "Single attention"), and another one that uses two attention layers, one after the coding stage and another one right after the TCN ("Double attention"). Furthermore, we investigate the effect that increasing the number of filters used in the skip-connections has on the performance of the model. Specifically, we tested $S = \{4, 8, 16, 32, 128, 256, 512, 1024\}$. It is important to point out that because the first attention layer is the attention operation between the left and right coded inputs, only the second attention layer size is variable; see Figure 1. We trained each configuration and attention layer size (or the number of filters in the skip-connections for the "Independent" condition) 5 times to allow statistical inference.

2.3. Hyperparameters

Hyperparameters of the implemented models are shown in Table 1. For a detailed description of these hyperparameters refer to [2].

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).*

2.4. Dataset

The audio dataset was provided by the 1st Clarity Enhancement Challenge [1]. The training data consist of 6,000 scenes including 24 different speakers. The development dataset, used to monitor the models' performance, consists of 2,500 scenes including 10 target speakers. Each scene corresponds to a unique

[1]https://github.com/APGDHZ/BinAttSE

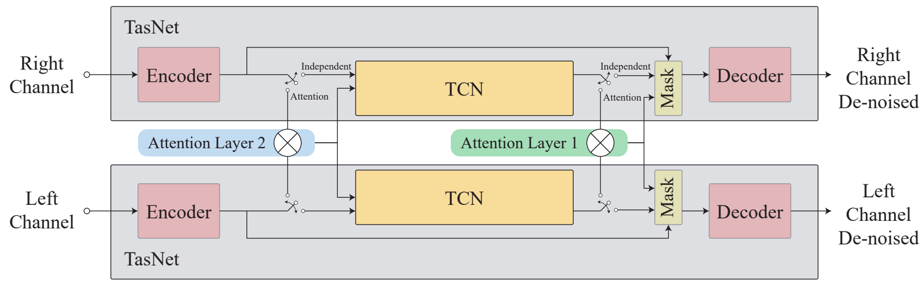

Figure 1: *Block diagram of the evaluated algorithms. "Independent" model bypasses both attention layers. "Single attention" model uses attention layer 1 and bypasses attention layer 2. "Double attention" model uses both attention layer blocks.*

target utterance and a unique segment of noise from an interferer [5], mixed at source-to-noise ratios (SNRs) ranging from -6 to 6 dB. The three sets are balanced for the target speaker's gender. Binaural Room Impulse Responses (BRIRs) were used to model a listener in a realistic acoustic environment. The audio signals for the scenes are generated by convolving source signals with the BRIRs and summing. BRIRs were generated for hearing aids located in each listening side, providing with 3 channels each (front, mid, rear). For this study, only the front channel was used at an 8 kHz sampling rate.

2.5. Training

The models were trained for a maximum of 100 epochs on batches of two 4-second long audio segments. The initial learning rate was set to 1e-3. The learning rate was halved if the accuracy of the validation set did not improve during 3 consecutive epochs, early stopping with a patience of 5 epochs was applied as a regularization method, and only the best performing model was saved. For the model optimization, Adam [6] was used to maximize the scale-invariant source-to-noise ratio (SI-SNR) [7]. The models were trained and evaluated using a PC with an Intel(R) Xeon(R) W-2145 CPU @ 3.70GHz, 256 GB of RAM, and an NVIDIA TITAN RTX as the accelerated processing unit.

3. Results

Figure 2 shows box-plots of the mean left/right SI-SNR as a function of S. Here it can be seen that the models that use attention layers perform numerically better than the "Independent" model. Specifically, it can be seen that a double attention layer yields significantly better performance than the model that uses none. It can also be seen from Figure 2 that there is a positive correlation between attention size and SI-SNR, however, this trend seems to happen until about an attention size of 128 units, above which the performance stops improving. Note that the configurations with the same S contain the same number of trainable weights.

Table 2 shows the maximum MBSTOI [8] scores achieved for all the different tested models (5 for each configuration) and the provided baseline [1]. It can be seen, that in general, the models that use attention layers yield the highest scores. Based on these results we decided to select the best performing "double attention layer" model with attention layers of size 512 (see bold condition in Table 2). This model contains 479,872 trainable parameters and with our computing system, performed audio speech enhancement in $21\mu s/sample$.

Table 3 shows the evaluation MBSTOI [8] scores achieved by the submitted algorithm and the provided baseline [1].

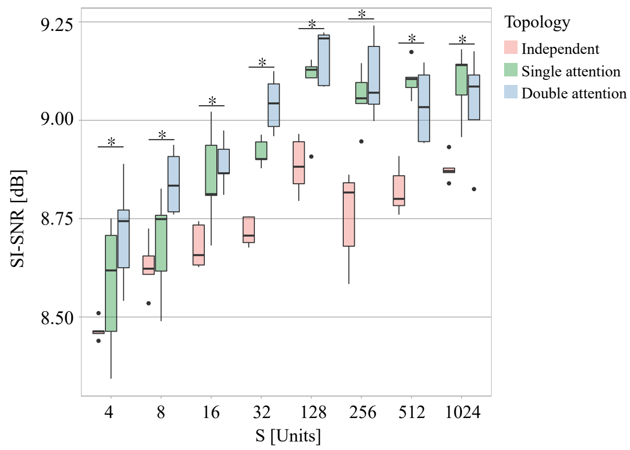

Figure 2: *Boxplots of the mean left/right SI-SNR for each of the tested configurations. The black horizontal bars within each of the boxes represent the median for each condition.*

S	Validation MBSTOI			
	Baseline [1]	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.*

Interferer	Evaluation MBSTOI	
	Baseline [1]	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.*

4. Discussion & Conclusions

In this short report, we described our methods to test different potential submissions for the 1st Clarity Enhancement Challenge. Based on the results we decided to submit a model which performs well both on, SI-SNR and MBSTOI measures, specifically, a binaural speech enhancement method based on two conv-TasNets and containing two attention layers of size 512. This model obtained a validation MBSTOI score of 0.77 and a mean left/right validation SI-SNR of 9.17 dB. The score obtained in the evaluation dataset, however, showed a drop in MBSTOI score of about 0.25.

5. References

[1] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, G. N. J. F. Culling, E. Porter, and R. V. Muñoz, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in proceedings of the annual conference of the international speech communication association," in *INTERSPEECH 2021*, Brno, Czech Republic, 2021.

[2] 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.

[3] 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. Mané, 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. Viégas, 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. [Online]. Available: https://www.tensorflow.org/

[4] 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. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[5] 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. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.804

[6] 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. [Online]. Available: http://arxiv.org/abs/1412.6980

[7] 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.

[8] 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.