# BUT System for the First Clarity Enhancement Challenge

*Katerina Zmolikova, Jan “Honza” Černocký*

Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia

izmolikova@fit.vutbr.cz

## Abstract

This paper describes BUT's efforts in the development of the system for the first Clarity challenge, concerned with enhancing the intelligibility of speech-in-noise for hearing-impaired listeners. Our system consists of three main parts: beamforming, post-enhancement neural network, and listener-adjustment neural network. For the beamforming module, we use a Minimum Variance distortion-less response (MVDR) beamformer computed from time-frequency masks estimated by the Complex Gaussian mixture model (CGMM). For the post-enhancement neural network, we use causal ConvTasnet architecture and train it with multi-task objectives of STOI, SNR, and PMSQE measures. The third, listener adjustment neural network extends the second module by an auxiliary neural network that estimates gains based on listener audiogram. Overall algorithmic latency of our processing is 210 samples (about 4.76 milliseconds). For training of the systems, we use only the data provided with the Clarity dataset. Overall, our system achieves a mean MBSTOI of 0.671 as compared with the baseline system with a mean MBSTOI of 0.415.

**Index Terms**: speech recognition, human-computer interaction, computational paralinguistics

## 1. Introduction

Clarity challenges[1] [1] address the problem of hearing aid processing of speech-in-noise. The task in the first enhancement challenge is to improve the intelligibility of the mixture signals containing a single target and single interferer signal.

## 2. Overview of the system

As depicted in Figure 1, our system consists of three main modules:

1. **Beamforming** This module takes the signals from all channels and applies Minimum Variance Distortionless Response (MVDR) beamformer based on time-frequency masks estimated by Complex Gaussian Mixture Model (CGMM). This produces two signals corresponding to the left and right ear.
2. **Neural network post-enhancement** Neural network with causal Conv-Tasnet architecture then uses the two beamformed signals together with two channels of the original mixture to further enhance the intelligibility of the signals. Multiple loss functions considering both intelligibility and quality of the output signals are used to train the network in a multi-task fashion.
3. **Neural network listener adjustment** Finally, the audiogram is taken into account to adjust the signals to a particular listener. This is done by applying a gain to each dimension of encoded representations of the signals. The gains are estimated by an auxiliary neural network that takes the audiogram as the input. This neural network is trained to improve the binaural intelligibility after forwarding the signal through the hearing loss model. For this, both intelligibility and hearing loss models from the baseline were modified to be differentiable.

**Total algorithmic latency:** All processing blocks act in a causal manner. The algorithmic latency thus stems from the windowing of the representation that is used for the processing. Overall, this corresponds to 210 samples of look-ahead, i.e. about 4.76 milliseconds. The latency of individual blocks is specified in corresponding sections.

**Computational requirements:** The infrastructure used to run the experiments was, in the case of CPU, an Intel(R) Xeon(R) CPU 5675 @ 3.07GHz, with a total memory of 37GB, and in the case of GPU, a Tesla P100 PCIe with 16GB of memory. The inference is run on CPU and takes $2\times$ real-time for the beamforming and $10\times$ real-time for the neural network.

## 3. Beamforming

We perform online beamforming of mixed-signal from channels CH1-CH3 using both left and right ear recordings (overall 6 channels). The beamforming is done on Short-time Fourier transform (STFT) representation, with a window size of 200 samples (about 4.5 milliseconds) with 100 samples shift. The beamforming is using a model introduced in [2], where complex Gaussian mixture model (CGMM) is used to estimate time-frequency masks and subsequently, Minimum Variance Distortionless response (MVDR) beamformer is estimated from the masks and applied on the observed signals. While in [2], the algorithm is block-online, here, we use a block of one frame only, similar to frame-online processing in [3]. The CGMM uses a prior on the covariance matrices of target and interference speech, consisting of the apriori expectation of the covariance matrix $\Psi_f^{(\nu)}$ and number of apriori observations $\eta^{(\nu)}$, where $f$ denotes frequency bin, $\nu \in \{t, i\}$ is the index of the source (target or interference). In our work, we estimate $\Psi_f^{(\nu)}$ from target and interference recordings in the training set, and set $\eta^{(t)} = \eta^{(t)} = 1$ based on the performance on development set. Additionally, we fix the posterior of the mixture weights of frames corresponding to the first two seconds to the target component. The resulting posterior of mixture weights of each time-frequency bin (time-frequency mask) is used to compute the MVDR beamformer, using the frame-online update rule defined in [3]. We apply two beamformer filters, with reference microphones CH1-left and CH1-right to obtain the binaural signal. We also multiply the resulting STFT with the estimated time-frequency mask to further reduce residual interference.

**Algorithmic latency:** The parameters of the spatial model and beamformer are updated online, frame-by-frame. The latency stems from the STFT representation with windows of 200 samples, corresponding to about 4.54 milliseconds.

[1] http://claritychallenge.org

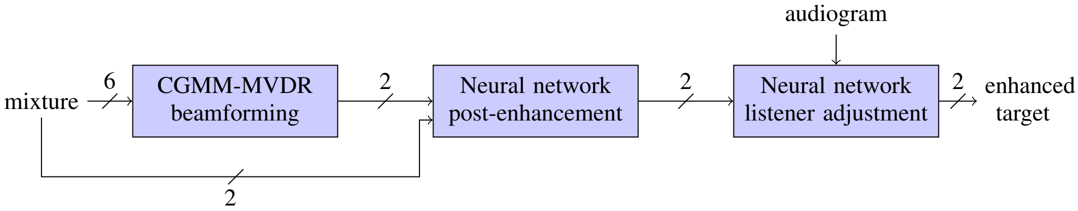
audiogram
mixture
6
CGMM-MVDR beamforming
2
Neural network post-enhancement
2
Neural network listener adjustment
2
enhanced target
2

Figure 1: *Overall scheme of the system.*

## 4. Neural network post-enhancement

The outputs of beamforming are further post-processed by a neural network. We use ConvTasnet architecture in its causal variant [4], as implemented in Asteroid toolkit [5]. We set the hyperparameters to $N = 256, B = Sc = 128, H = 512, P = 3, R = 3, X = 8$. The encoder and decoder use windows of 10 samples with 5 samples shift. We use both binaural signals on the input of the network, both are processed by the same encoder and then concatenated. In addition, we also use the unprocessed signal from CH1 on the input, to enable recovering from unsuccessful beamforming (e.g. in a case where the angle between target and interference is small). The output of the network is again a binaural signal. The network is trained with a multi-task loss function. The main component is the STOI [6] loss function, with a weight of 0.9. Next, we use SNR and PMSQE [7] loss, both with weights of 0.05 to prevent the network from overfitting to the STOI measure. The target for all parts of the loss function is the target anechoic signal.

**Algorithmic latency:** The ConvTasnet architecture is causal apart from the initial encoder, which uses 10 sample windows. This causes the algorithmic latency of 10 samples, corresponding to about 0.23 milliseconds.

## 5. Neural network listener adjustment

To adjust the processing to a particular listener, we used the audiogram in the neural network processing stage. We took the trained model described in Section 4 and extended it with an auxiliary network, consisting of 3 fully connected layers, with Leaky ReLU activations. The auxiliary network takes on input the audiogram and the representation on the output of the separator part of ConvTasnet. The output of the auxiliary network was then used as gains for the estimated representations before decoding. The auxiliary network was initialized to output approximately one, thus mimicking the processing without the auxiliary network. We fixed the parameters of the original network and trained the auxiliary network only, with the objective of MBSTOI [8] of signals processed by the MSBG hearing loss model [9] (corresponding to the evaluation measure used in the challenge). For this, we rewrote the baseline models provided by the organizers into their differentiable versions. As targets of the optimization, we again used the anechoic target signals.

**Algorithmic latency:** This part only extends the neural network introduced in the previous section. The extension processes each frame separately. Therefore, this does not introduce any algorithmic latency in addition to the one introduced in Section 4.

Table 1: *Speech intelligibility results on development set. HL+MBSTOI refers to evaluation of MBSTOI after application of hearing loss model to the signals and corresponds to the final evaluation metric in the challenge. All reported numbers are mean results over the set.*

| | method | MBSTOI | HL+MBSTOI |
|---|---|---|---|
| (1) | baseline | - | 0.415 |
| (2) | CGMM+MVDR | 0.707 | 0.599 |
| (3) | NN post-enh (SNR) | 0.767 | 0.631 |
| (4) | NN post-enh (bf-only) | 0.770 | 0.635 |
| (5) | NN post-enh | **0.795** | 0.657 |
| (6) | NN listener (random) | 0.759 | 0.662 |
| (7) | NN listener | 0.759 | **0.671** |

## 6. Results and discussion

### 6.1. Objective intelligibility results

The results of our system are shown in Table 1. We report two metrics: *MBSTOI* which is the modified binaural STOI computed directly from the enhanced signals, and *HL+MBSTOI*, which corresponds to computing MBSTOI after applying the baseline hearing loss model to the enhanced signals (this is the final evaluation metric used in the challenge).

In rows (1) and (2) in Table 1, we observe significant improvement in HL+MBSTOI when using CGMM+MVDR beamforming as compared with the baseline processing. The neural network-based post-enhancement, as described in Section 4, further improves the intelligibility, as shown in row (5). We conducted additional experiments to further study different aspects of the model and training. First, we performed the training with SNR loss only, as opposed to the multi-task loss. Second, we trained a neural network with the beamformed input only, without using also the unprocessed mixtures at the input. The results shown on rows (3) and (4) show that both of these aspects slightly hurt the performance.

Finally, the result of using listener adjustment described in Section 5 is shown on row (7) of Table 1. The results show improvement in the HL+MBSTOI metric and degradation on MBSTOI without the hearing loss model. This is expected as this processing should learn to compensate for the hearing loss. To test whether the neural network truly uses the audiogram to adjust the output to the listeners, we used audiograms of random listeners during the evaluation in an experiment shown on row (6). While this result is still better than *NN post-enh*, it stays behind the correct listener adjustment on row (7). We can thus conclude that the gains introduced with the auxiliary network partially optimize the HL+MBSTOI loss in general, and partially adjust to the listener using the audiogram.

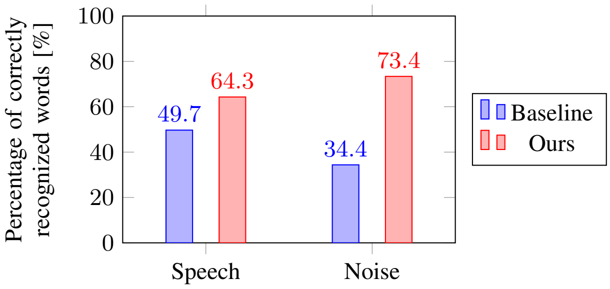

Figure 2: *Results of subjective listening tests comparing baseline and ours system. The results show percentage of words correctly recognized by the listeners in case of speech or noise as the interference.*

### 6.2. Subjective intelligibility results

Figure 2 show the results of subjective listening tests performed by 27 hearing-impaired listeners. The system clearly leads to more intelligible speech than the baseline, especially in the case of noise as interference. Note that in the case of speech as interference, the listeners might have had problems with identifying the target speaker when the interference is too suppressed, as discussed during the Clarity workshop [10]. For some listeners, the accuracy in the speech-interference condition was exactly or close to 0%, which skews the results.

## 7. Conclusions

The presented system for the Clarity challenge significantly outperforms the baseline, with the use of CGMM-MVDR beamforming and neural network-based intelligibility enhancement. In the Clarity challenge, the system obtained 2nd best result in terms of objective MBSTOI and 5th-6th place in listening tests. This shows that perhaps the gains obtained by direct optimization of the HL+MBSTOI metric might not be always reflected in the subjective intelligibility.

## 8. References

[1] S. Graetzer, J. Barker, T. Cox, M. Akeroyd, J. Culling, G. Naylor, E. Porter, and R. Viveros Muñoz, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2021, Brno, Czech Republic*, 2021.

[2] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online mvdr beamformer based on complex gaussian mixture model with spatial prior for noise robust asr," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 780–793, 2017.

[3] T. Higuchi, K. Kinoshita, N. Ito, S. Karita, and T. Nakatani, "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, 2018, pp. 531–535.

[4] 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, no. 8, pp. 1256–1266, 2019.

[5] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas *et al.*, "Asteroid: the pytorch-based audio source separation toolkit for researchers," *arXiv preprint arXiv:2005.04132*, 2020.

[6] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[7] J. M. Martin-Donas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal processing letters*, vol. 25, no. 11, pp. 1680–1684, 2018.

[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, pp. 1–13, 2018.

[9] Y. Nejime and B. C. Moore, "Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 603–615, 1997.

[10] S. Graetzer and J. Barker. Clarity-2021 workshop talk: Clarity enhancement challenge overview and results. Youtube. [Online]. Available: https://www.youtube.com/watch?v=Kjd0TID6wD8