# The SGU Systems for the CHiME-7 UDASE Challenge

*Jaehoo Jang, Myoung-Wan Koo*†

Sogang University, Korea

jeahoo4128@sogang.ac.kr, mwkoo@sogang.ac.kr

## Abstract

In this work, we present a description of SGU domain-adapted speech enhancement system implementation that enhances the baseline of the CHiME-7 challenge. We introduce two significant modifications. Firstly, we replace the Sudo rm-rf [1] architecture with the Mossformer [2], which incorporates convolution-augmented joint local and global self-attention mechanisms. It performs fully-computed self-attention on local chunks and utilizes linearized low-cost self-attention over the entire sequence. As a second modification, we incorporate a speech purification technique at the baseline when conducting self-supervised learning for the student model. This technique predicts the frame-level SNR of the pseudo-target speech and utilizes them as weights for the discrepancy function between the pseudo-target speech and the student model's estimated speech. Consequently, We achieved an SI-SDR score of 12.42 on the LibriCHiME-5 dataset for both modifications. Additionally, implementing the Mossformer architecture on the CHiME-5 dataset leads to a 2.90 OVRL-MOS and 3.39 SIG-MOS. Also, the application of the purification method results in a 3.71 BAK-MOS. Finally, we demonstrate the superior performance of our approach compared to the baseline.

**Index Terms**: speech enhancement, noise suppression, domain adaptation, CHiME-7 challenge

## 1. Introduction

Speech enhancement systems that utilize supervised learning primarily rely on the methodology of extracting clean speech through a masking network [1, 3, 4, 5]. However, if only unlabeled noise mixtures are available without clean source speech, it's impossible to train such systems. Accordingly, several studies have proposed unsupervised learning methods for speech enhancement system that can employ such noise mixtures in the training process free from the constraints of clean source speech [6, 7, 8, 9].

In order to leverage the knowledge of a model trained from a different domain, the CHiME-7 challenge aims to improve the noise suppression performance on the in-domain speech by utilizing both an unlabeled in-domain CHiME-5 [10] dataset and a labeled out-of-domain(OOD) Librimix [11] dataset. RemixIT pipeline is a baseline provided by the challenge organizers. In this system, the fully-supervised teacher model is trained by using Librimix. Then, CHiME-5 data is fed into the frozen teacher model, which outputs pseudo-target speeches and noise waveforms. These are used to create noise-permuted bootstrapped mixtures, which are then provided to the student model for self-supervised learning. Additionally, the parameters of the student model can be transferred back to the teacher model for continuous refinement at the end of each epoch.

Limitation of this system stems from its dependency on a distillation-based pipeline driven by a teacher model. This dependency, coupled with domain imbalance issues, raises concerns about ensuring the quality of pseudo-target speech. Specifically, the performance of a teacher model trained on the speech from a different domain degrades notably when confronted with input from another domain. Deterioration of the performance is primarily due to the distinctive prosodic information, linguistic contextual dependencies, speaker characteristics, and other inherent attributes that are unique to the train data.

To overcome this obstacle, we propose an enhanced system with two primary modifications implemented within the RemixIT pipeline. The initial modification consists of implementing the Mossformer architecture in the enhancement system's back-end to efficiently capture the long-range direct interaction between the global intermediate feature and the local feature. The second involves the application of the speech purification technique, which focuses on utilizing the speech quality of pseudo-target speech segments in terms of SNR. This technique is used to train the student model by emphasizing high-quality segments. The former enables a more detailed feature design compared to the baseline model employing a U-net-based masking network and contributes to the fundamental enhancement of the performance of the enhancement system. The latter leverages refined prosodic details from pseudo-target speech to facilitate performance improvement.

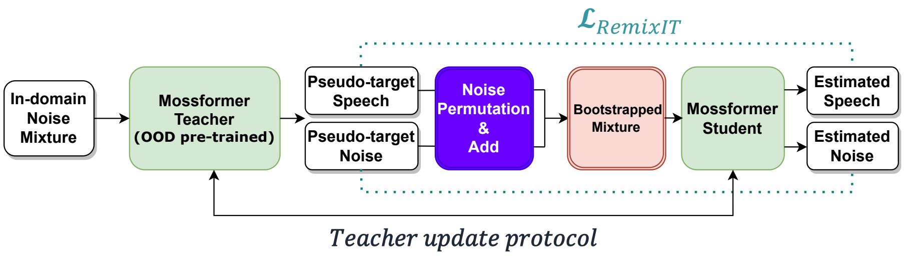
## 2. System description

The overall system architecture we proposed is illustrated in Figure 1, representing two distinct speech enhancement system pipelines operating independently. In Figure 1a, we simply replace the Sudo rm -rf with the Mossformer as the back-end of pipeline. and in Figure 1b, we applied the purification technique in the form of a discrepancy function between the pseudo-target speech and the estimated speech of the student model. In 2.1, we briefly introduce Mossformer, outlining the specific model structure, and the composition of the masking network. In 2.2, we explain the assumption of speech purification technique and its suitability within the RemixIT pipeline. Also we provide an in-depth exploration of the process of designing the discrepancy function for purification, employing the SNR predictor as a fundamental element of the technique.

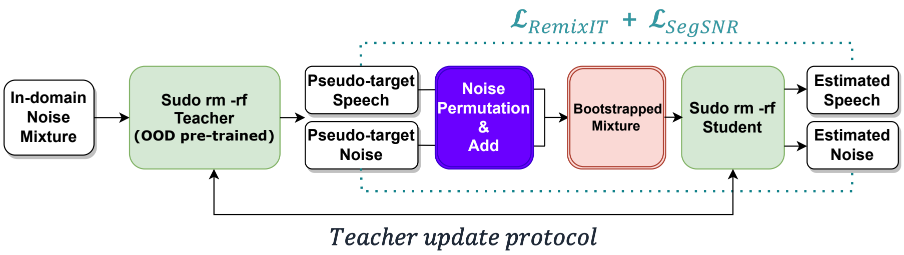
### 2.1. Mossformer Adaptation

Transformer-based speech enhancement models like Sepformer [5], have shown impressive results in the task of speech separation by intentionally designing long-range interaction among speech sequences, mainly through a multi-head self-attention mechanism. Nonetheless, employing this approach results in significant computational limitations in terms of context size. And it imposes a negative influence on the long-range feature interaction because of the temporal dependencies between distant features. To overcome this issue, the Mossformer architec-

†Corresponding Author




(a) *Mossformer-based system pipeline*




(b) *Purification method-based system pipeline*

Figure 1: *Overview of the systems we have developed.* $(a)$ *denotes the back-end adjustment utilizing Mossformer.* $(b)$ *involves the combination of* $L_{SegSNR}$ *(segmental SNR loss) and* $L_{RemixIT}$*, where The former is calculated by multiplying the segmental SNR with the weights acquired when pseudo-target speech is used as input for the SNR predictor. The more detailed explanation of segmental SNR loss can be found in 2.2.3.*

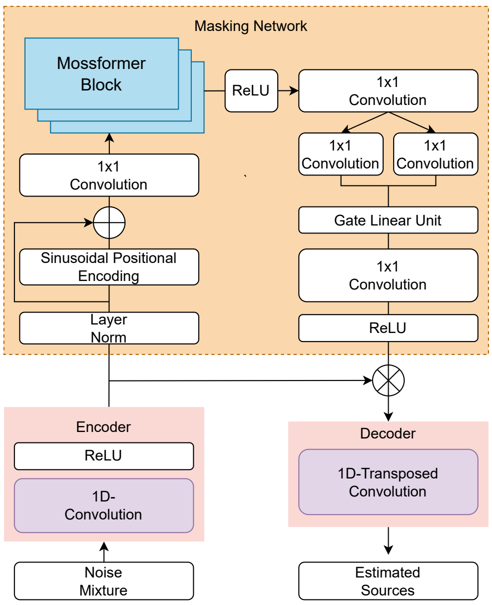



Figure 2: *The Mossformer architecture*

ture has been proposed.

The masking network of Mossformer incorporates a gated attention unit structure with joint local and global self-attention. This design creates a mask that efficiently captures long-range interaction while removing temporal dependencies. Furthermore, structuring the attention network using a single-head self-attention not only reduces computational requirements but also enhances the context capacity. Additionally, a module based on depth-wise convolution is introduced to extract the key, query, and value for attention, allowing for a fine-grained design of local patterns within the features. Due to the aforementioned benefits, Mossformer achieves state-of-the-art results on the WSJ0-2/3mix [12] and WHAM!/WHAMR! [13, 14] datasets. We simply replaced the baseline back-end with the MossFomer in our pipeline. Corresponding architecture is illustrated in Figure 2.

### 2.1.1. *Architecture*

The model is structured with the typical form of a speech separation model, consisting of an encoder-decoder and a masking network. The encoder consists of 1D-convolution layer and ReLU activation, while the decoder is composed of standard 1D-transposed convolution layers. The encoder transforms input speech data into hidden representations, which are then fed into the masking network that generates masks in the hidden dimension. Additionally, the generated masks and the encoder's hidden representations are element-wise multiplied and fed to the decoder. Subsequently, the decoder produces estimated sources. Through this architecture, the model can perform the speech enhancement task of separating noise from the mixture.

The masking net includes normalization, positional encoding, 1x1 convolutions before entering the Mossformer block. the masking net input is initially normalized and goes through the procedure of element-wise multiplication with a positional encoding vector. then it passes the 1x1 convolution followed by mossformer blocks. The output of the final Mossformer block is passed through a 1x1 convolution, then extended to the number of sources intended to be estimated. After that, each source goes through a 1x1 convolution and is subsequently treated with a shared Gated Linear Unit. Following this, a final 1x1 convolution is conducted, and the output passes the ReLU activation to produce non-negative masks for each source.

#### 2.1.2. *Mossformer block*

Mossformer block, the core part of the model, incorporates convolution modules, joint local and global single-head self-attention, and an attentive gating mechanism. The convolution modules capture local feature patterns using linear layers, SiLU activation, and 1D depth-wise convolutions.

There are three convolution modules. One of the convolution modules produces a hidden representation of each block's initial input X. Following that, this representation is subject to the application of scale and offset, along with RoPE [15], resulting in the generation of queries Q, Q′, as well as keys K, K′. Q′ and K′ are used for global attention, while Q and K are employed for local attention. The remaining convolution modules produce values U and V, on which both local and global self-attention is applied. The outcomes of attention are summed to generate new attention values, U′ and V′.

The process of joint local and global self-attention is illustrated in equation (1). $*_{global}$ signifies global attention output, and $*_{local}$ denotes local attention output. $h$ represents the index of non-overlapping chunk units in local attention. At this point, $\mathrm{U}'_{local}$ and $\mathrm{V}'_{local}$ are constructed by stacking attention outputs performed for all chunks. The scaling factors of each attention method are denoted as $\beta$ and $\gamma$.

$$
\begin{aligned}
\mathrm{V}'_{global} = \mathrm{Q}'(\beta \mathrm{K}'^{\mathrm{T}} V), \quad \mathrm{U}'_{global} = \mathrm{Q}'(\beta \mathrm{K}'^{\mathrm{T}} U) \\
\mathrm{V}'_{local,h} = \mathrm{ReLU}^2(\gamma \mathrm{Q}_h \mathrm{K}_h^T)\mathrm{V}_h \\
\mathrm{U}'_{local,h} = \mathrm{ReLU}^2(\gamma \mathrm{Q}_h \mathrm{K}_h^T)\mathrm{U}_h \\
\mathrm{V}' = \mathrm{V}'_{local} + \mathrm{V}'_{global}, \quad \mathrm{U}' = \mathrm{U}'_{local} + \mathrm{U}'_{global}
\end{aligned}
\tag{1}
$$

U, U′, and V, V′ are utilized in combination with the gating mechanism, as shown in Equation (2), to form the resulting output sequences O′ and O′′. Next, O′ and O′′ proceed element-wise multiplication and are fed to an additional convolution module. The final output sequence O of each block is shaped by a skip connection between the extra convolution module's output and the block's initial input X. This process is described in Equation (3), and further details can be found in [2].

$$\mathrm{O}' = \phi(\mathrm{U} \otimes \mathrm{V}'), \quad \mathrm{O}'' = \mathrm{U}' \otimes \mathrm{V} \tag{2}$$

$$\mathrm{O} = \mathrm{X} + \mathrm{ConvM}(\mathrm{O}' \otimes \mathrm{O}'') \tag{3}$$

### 2.2. Speech purification

#### 2.2.1. *Assumptions and Suitability*

The initial application of the speech purification method in terms of a self-supervised learning scheme was proposed in the work by [16]. In previous work, it is assumed that the frames of noise mixture with high SNR are almost identical to frames of clean speech. And the target speech is assumed to potentially contain noise.

The main idea of speech purification is to design the discrepancy function in a way that is more affected by the frames in the pseudo-target speech with higher SNR values. To achieve this, in [16], a novel discrepancy function named Segmental SNR loss is proposed. This loss function efficiently incorporates the frame-wise SNR of the target speech as a weighting factor.

Since the teacher model in the RemixIT pipeline is trained by OOD data, it can't create a perfectly clear speech for the target domain. This leads to the target speech characteristics in [16] resembling the pseudo-target speech produced by the teacher model. Therefore, during the training of the student model, we utilize the Segmental SNR loss to our implemented system pipeline.

#### 2.2.2. *SNR predictor*

The frame-wise SNR-based weight that is multiplied with the Segmental SNR loss is derived from a simple regressive model called the SNR predictor, which is based on an RNN architecture. The SNR predictor is pre-trained and combined with the whole system pipeline during the student model training while being kept in a frozen state. The SNR predictor's training process, which is shown in equation (4), is based on a labeled dataset containing clean speech and noise mixture. $s$ represents clean speech, $n$ denotes noise, and $\alpha$ signifies the segmental SNR between the noisy mixture $x$ and $s$. $\hat{\alpha}$ refers to the frame-wise SNR estimated by the SNR predictor given $x$. Here, $h$ and $\mathcal{W}_h$ respectively represent the SNR predictor and its parameters. The training of the SNR predictor aims to minimize the MSE between the segmental SNR $\alpha$ and the estimated frame-wise SNR values $\hat{\alpha}$.

$$
\begin{aligned}
x &= s + n \\
\alpha &= \mathrm{SegSNR}(s, x) \\
\hat{\alpha} &= h(x; \mathcal{W}_h) \\
\mathcal{W}_h &\leftarrow \underset{\mathcal{W}_h}{\mathrm{argmin}}\, \mathrm{MSE}(\hat{\alpha}, \alpha)
\end{aligned}
\tag{4}
$$

The target segmental SNR(SegSNR) is calculated by the following equation (5). $v_i$ is denoted as the target clean speech, while $r_i$ denotes the residual between $v_i$ and the estimated speech $\hat{v}_i$. Detailed explanations regarding the symbols can be found in section 2.2.3.

$$\mathrm{SegSNR}_j(v, \hat{v}) = 10 \log_{10}\left[\frac{\sum_{i=Hj}^{Hj+N-1}(w_{i-Hj} v_i)^2}{\sum_{i=Hj}^{Hj+N-1}(w_{i-Hj} r_i)^2}\right] \tag{5}$$

When training in our pipeline, the SNR predictor takes a pseudo-target speech as input and makes individual predictions of SNR for each frame. The resulting logits from the SNR predictor, known as frame-wise SNR, are processed through a sigmoid function. Ultimately, these logits are transformed into weights within the range of 0 to 1. To be specific, frames predicted with high SNR values will yield weights closer to 1, while frames predicted with low SNR values will result in weights closer to 0. Then, we multiply the weight with the segmental SNR. Note that the frame-wise SNR weight is computed using only the pseudo-target speech by SNR predictor, while the segmental SNR is calculated between the pseudo-target speech and the estimated speech. As a result, the segmental SNR loss is obtained.

#### 2.2.3. *Segmental SNR loss*

The segmental SNR loss is detailed in equation (6). The $J$, $H$, and $N$ respectively denote the number of frames, hop size, and frame size, while $j$ refers to the index of a specific frame. $w_i$ represents the Hann window function of length $N$, $\overline{s}$ denotes the pseudo-target speech (not the bootstrapped mixture), and $\overline{r}$ is the residual vector between $\overline{s}$ and the estimated speech. $p_j$ denotes the weight for the j-th frame. Ultimately, we combine the

| Model | Type | LibriCHiME-5 SI-SDR | CHiME-5 OVRL-MOS | CHiME-5 BAK-MOS | CHiME-5 SIG-MOS |
|---|---|---|---|---|---|
| Sudo rm-rf (baseline) | Supervised | 9.39 | 2.81 | 3.54 | 3.23 |
| | RemixIT | 11.70 | 2.86 | 3.65 | 3.28 |
| | RemixIT$_{\text{vad}}$ | 11.57 | 2.85 | 3.66 | 3.27 |
| Sudo rm-rf$_p$ | RemixIT$_{\text{vad}}$ | 12.42 | 2.88 | **3.71** | 3.33 |
| Mossformer | Supervised | 10.63 | 2.88 | 3.52 | 3.39 |
| | RemixIT | 12.42 | **2.90** | 3.60 | **3.39** |
| | RemixIT$_{\text{vad}}$ | **12.58** | 2.84 | 3.48 | 3.35 |

Table 1: *Overall experiment results of our implemented system pipeline. The baseline system is Sudo rm-rf. Improved version with speech purification is Sudo rm-rf$_p$. The remaining is the Mossformer implementation system.*

segmental SNR loss with the SI-SDR, which is used as the loss function in the original RemixIT pipeline, with equal weights as shown in equation (7). We proceeded with training by assigning the same weights to the SI-SDR loss for speech, the SI-SDR loss for noise, and the segmental SNR loss ($\lambda_1$=$\lambda_2$=$\lambda_3$).

$$\mathcal{L}_{SegSNR} = -\frac{1}{J}\sum_{j=0}^{J-1} p_j [10\log_{10}\frac{\sum_{i=Hj}^{Hj+N-1}(w_{i-Hj}\overline{s}_i)^2}{\sum_{i=Hj}^{Hj+N-1}(w_{i-Hj}\overline{r}_i)^2}] \tag{6}$$

$$\begin{aligned} \mathcal{L}_{RemixIT} &= \mathcal{L}_{speech} + \mathcal{L}_{noise} \\ \mathcal{L}_{total} &= \lambda_1\mathcal{L}_{speech} + \lambda_2\mathcal{L}_{noise} + \lambda_3\mathcal{L}_{segSNR} \\ \lambda_1 + \lambda_2 + \lambda_3 &= 1 \end{aligned} \tag{7}$$

# 3. Experimental setup

To address the CHiME-7 UDASE challenge, we follow the guidelines and utilize three different datasets: CHiME-5 (unlabeled in-domain dataset), Librimix (labeled out-of-domain dataset), and LibriCHiME-5 (labeled dataset resembling the in-domain data). We extract subsets for training, development, and evaluation from each dataset using an official toolkit from the CHiME-7 challenge's github[1]. The model is trained using these subsets in the original format provided by the toolkit.

To implement the system pipelines described in the Figure 1, we initially used the provided baseline implementation to assess its performance in our experimental setup. And we leveraged two external tools. The first tool we employed is the Mossformer architecture implementation[2], as described in [2]. During experiments, a large version of Mossformer with 42.1 million parameters was utilized. The second tool integrates an SNR predictor and the segmental SNR loss implementation[3]. Additionally, we used publicly available pre-trained weights of the SNR predictor from the same github[3] and froze them during training. The pre-trained weights were trained using a mixture of utterances from Librispeech[17] and noises from MUSAN[18]. More detailed information about the training scheme of the SNR predictor is described in [16].

Subsequently, we incorporated the aforementioned methods into the RemixIT baseline system separately. In one approach, we simply replaced Sudo rm-rf with Mossformer. While in the other, we applied the purification method with SNR predictor. For the former, we followed the configuration for the large model as described in [2], while the latter maintained the same hyperparameters as the baseline setting without setting 200 epochs. During training, we used a learning rate of 1.5e-4 for 100 epochs with the Adam optimizer. After 100 epochs, if the loss did not improve for 10 consecutive epochs, we reduced the learning rate by a factor of 3. In both approaches, we maintained the same learning rate throughout the training process. All experiments were conducted on six NVIDIA A100 GPUs with 80 GB of memory.

[1]https://github.com/UDASE-CHiME2023/baseline
[2]https://github.com/modelscope/modelscope
[3]https://github.com/IU-SAIGE/pse

# 4. Result

## 4.1. Performance of the proposed systems

Table 1 shows our experiment results. We used self-supervised learning with two subsets: unlabeled-10$s$ (RemixIT setting) and vad-10$s$ (RemixIT$_{vad}$ setting) from CHiME-5. The baseline Sudo rm-rf experiment yielded an SI-SDR score of 11.57 using the vad-10$s$ subset. This was achieved by training the models from scratch without altering the provided code by the challenge organizers. As incorporating purification techniques, the SI-SDR score improved to 12.42. Additionally, the corresponding systems achieved the highest BAK-MOS score of 3.71.

Mossformer outperforms Sudo rm-rf in SI-SDR with an impressive score of 12.58 in RemixIT$_{vad}$ setting. In RemixIT setting, Mossformer also achieved the highest scores, recording 3.39 for SIG-MOS and 2.90 for OVRL-MOS, respectively. Despite having significantly more parameters and slower training speeds compared to Sudo rm-rf, latency is not a constraint in this challenge, so we proposed both system pipelines.

Consequently, we submitted two systems for the challenge. ISDS1 utilized the Mossformer model in the RemixIT setting, trained on unlabeled-10$s$ data. For ISDS2, we employed the Sudo rm-rf model with the purification method in the RemixIT$_{vad}$ setting.

## 4.2. Results of the challenge

The evaluation of systems developed for the CHiME-7 UDASE challenge involves two main stages: objective and subjective evaluation. In the first stage, objective evaluation was performed for all submissions, assessing the SI-SDR for the LibriCHiME-5 eval set and the DNS-MOS performance on the subset of the CHiME-5 eval set. The results are explained in section 4.2.1. The second stage involved a listening test for the output of the top four systems selected from the first stage,

| | LibriCHiME-5 | CHiME-5 | | |
|---|---|---|---|---|
| System | SI-SDR (dB) | OVRL | BAK | SIG |
| N&B | 13.0 | 3.07 | 3.93 | 3.39 |
| ISDS1 | **12.4** | **2.90** | 3.60 | **3.39** |
| ISDS2 | **12.4** | 2.88 | **3.70** | 3.32 |
| RemixIT-VAD | 10.1 | 2.84 | 3.62 | 3.28 |
| RemixIT | 9.4 | 2.82 | 3.64 | 3.26 |
| CMGAN-base | 7.8 | 3.40 | 3.97 | 3.76 |
| OOD teacher | 7.8 | 2.88 | 3.59 | 3.33 |
| Input | 6.6 | 2.84 | 2.92 | 3.48 |
| CMGAN-FT | 4.7 | 3.55 | 3.93 | 3.92 |

Table 2: *Objective evaluation results for all submissions (sorted by SI-SDR scores).*

corresponding to other subsets of the CHiME-5 eval set. Ultimately, the ranking of systems was determined based on the DNS-MOS score obtained by each system.

Except to the baseline systems (Input, OOD teacher, RemixIT, and RemixIT-VAD) and our proposed systems (ISDS1 and ISDS2), there are three different submissions. The N&B system integrated the MetricGAN [19] discriminator and Uformer [20] as the back-end enhancement model. This system included a UNA-GAN [21] application with the CHiME-5 in-domain noise extracted by VAD to generate an in-domain noise mixture. Furthermore, perceptual contrast stretching (PCS) [22] was employed as a pre-and post-processing method. The CMGAN-base system, similar to N&B, used MetricGAN but employed Conformer [23] as the back-end enhancement model. CMGAN-FT is the fine-tuned version of CMGAN-base using the LibriCHiME-5 dev set.

#### 4.2.1. Objective evaluation

Table 2 represents the evaluation results of objective metrics for all submissions. From the perspective of SI-SDR, N&B achieved the highest score of 13.0, while in terms of DNS-MOS, CMGAN-FT performed best with OVRL-MOS and SIG-MOS scores of 3.55 and 3.92, respectively. In the case of BAK-MOS, CMGAN-base outperformed others with a score of 3.97. Our proposed system, ISDS1, ranked second in SI-SDR and fourth in OVRL-MOS scores.

Note that our proposed systems did not proceed with data augmentation as the other three submissions did for the generalization of the system. Specifically, ISDS1 solely utilized the attention structure within Mossformer, while ISDS2 ensured generalization ability through the use of auxiliary purification SNR loss. This differs from the approaches used by N&B, which conducted in-domain noise mixture generation, and CMGAN-base and FT, which generated enhanced spectrograms using magnitude masks. Hence, we can expect our proposed systems to perform better in terms of objective metrics with additional data augmentation. Furthermore, by integrating the modifications made to ISDS1 and ISDS2, we can anticipate further performance improvements.

Following the first-stage evaluation results, our proposed ISDS1 system, which incorporates Mossformer adaptation, has been chosen for a listening test. However, since the ISDS2 system also demonstrated identical SI-SDR scores and very similar DNS-MOS scores to ISDS1, we believe there is a need to conduct a listening test for ISDS2 as well.

| Ranking | System | Mean | 95% CI | Median |
|---|---|---|---|---|
| 1 | N&B | 4.30 | 0.01 | 4.38 |
| 2 | ISDS1 | **3.08** | 0.01 | 3.00 |
| 3 | RemixIT-VAD | 2.97 | 0.01 | 2.88 |
| 4 | CMGAN-FT | 2.75 | 0.01 | 2.63 |
| 5 | Input | 2.20 | 0.01 | 2.19 |

Table 3: *Evaluation result of BAK-MOS*

| Ranking | System | Mean | 95% CI | Median |
|---|---|---|---|---|
| 1 | Input | 3.97 | 0.01 | 4.00 |
| 2 | ISDS1 | **3.43** | 0.01 | 3.56 |
| 3 | N&B | 3.41 | 0.01 | 3.63 |
| 4 | RemixIT-VAD | 3.02 | 0.02 | 3.25 |
| 5 | CMGAN-FT | 2.63 | 0.01 | 2.63 |

Table 4: *Evaluation result of SIG-MOS*

| Ranking | System | Mean | 95% CI | Median |
|---|---|---|---|---|
| 1 | N&B | 3.11 | 0.01 | 3.25 |
| 2 | ISDS1 | **2.75** | 0.01 | 2.75 |
| 3 | Input | 2.68 | 0.01 | 2.75 |
| 4 | RemixIT-VAD | 2.45 | 0.01 | 2.50 |
| 5 | CMGAN-FT | 2.14 | 0.01 | 2.13 |

Table 5: *Evaluation result of OVLR-MOS*

#### 4.2.2. Listening test

The subject evaluation involved 32 participants split into 4 panels, assessing 128 audio samples under 5 experimental conditions. Participants sat in a listening booth wearing headphones and listened to short 4-5 second speech samples.

As a result of the conducted listening test, we confirmed that our system secured the second position in all sub-evaluation metrics of DNS-MOS, as demonstrated in Tables 3, 4, and 5. In the case of BAK-MOS, the N&B system, which utilized in-domain noise extracted from the CHiME-5 train set for mixture generation, outperformed others significantly. However, for SIG-MOS, the ISDS1 system slightly edged ahead of the competition. This underscores the superiority of Mossformer's attentive gating mechanism module in capturing attributes of speech signals compared to the other competing submission systems. Finally, through the evaluation results, our proposed system achieved second place in this challenge.

## 5. Conclusions

Our speech enhancement system showed considerable performance improvement and surpassed the baseline system through two key modifications: the integration of the Mossformer architecture and the employment of the speech purification method.

Through the UDASE challenge, we were able to evaluate the performance of integrating the Mossformer model into the RemixIT pipeline. Moreover, we demonstrated that performance improvement can be achieved not through commonly used data augmentation techniques but rather by adding an auxiliary loss function associated with the SNR of each segment of speech during system training, implicitly enhancing the generalization performance of the baseline system. As part of future work, we intend to implement and assess the performance of a system that combines the two modifications we proposed.

# 6. References


[1] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo RM -rf: Efficient networks for universal audio source separation," in *30th IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2020, Espoo, Finland, September 21-24, 2020*. IEEE, 2020, pp. 1–6. [Online]. Available: https://doi.org/10.1109/MLSP49062.2020.9231900

[2] S. Zhao and B. Ma, "Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions," *CoRR*, vol. abs/2302.11824, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2302.11824

[3] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019. [Online]. Available: https://doi.org/10.1109/TASLP.2019.2915167

[4] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2020, pp. 46–50. [Online]. Available: https://doi.org/10.1109/ICASSP40776.2020.9054266

[5] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021, pp. 21–25. [Online]. Available: https://doi.org/10.1109/ICASSP39728.2021.9413901

[6] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 2993–3007, 2022. [Online]. Available: https://doi.org/10.1109/TASLP.2022.3207349

[7] S. Fu, C. Yu, K. Hung, M. Ravanelli, and Y. Tsao, "Metricgan-u: Unsupervised speech enhancement/ dereverberation based only on noisy/ reverberated speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. IEEE, 2022, pp. 7412–7416. [Online]. Available: https://doi.org/10.1109/ICASSP43922.2022.9747180

[8] Y. Xiang and C. Bao, "A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1826–1838, 2020. [Online]. Available: https://doi.org/10.1109/TASLP.2020.2997118

[9] G. Yu, Y. Wang, C. Zheng, H. Wang, and Q. Zhang, "Cyclegan-based non-parallel speech enhancement with an adaptive attention-in-attention mechanism," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2021, Tokyo, Japan, December 14-17, 2021*. IEEE, 2021, pp. 523–529. [Online]. Available: https://ieeexplore.ieee.org/document/9689669

[10] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, B. Yegnanarayana, Ed. ISCA, 2018, pp. 1561–1565. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-1768

[11] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[12] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*. IEEE, 2016, pp. 31–35. [Online]. Available: https://doi.org/10.1109/ICASSP.2016.7471631

[13] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 1368–1372. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-2821

[14] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "Whamr!: Noisy and reverberant single-channel speech separation," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2020, pp. 696–700. [Online]. Available: https://doi.org/10.1109/ICASSP40776.2020.9053327

[15] J. Su, Y. Lu, S. Pan, B. Wen, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *CoRR*, vol. abs/2104.09864, 2021. [Online]. Available: https://arxiv.org/abs/2104.09864

[16] A. Sivaraman and M. Kim, "Efficient personalized speech enhancement through self-supervised learning," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1342–1356, 2022. [Online]. Available: https://doi.org/10.1109/JSTSP.2022.3181782

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*. IEEE, 2015, pp. 5206–5210. [Online]. Available: https://doi.org/10.1109/ICASSP.2015.7178964

[18] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, 2015. [Online]. Available: http://arxiv.org/abs/1510.08484

[19] S. Fu, C. Liao, Y. Tsao, and S. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 2031–2041. [Online]. Available: http://proceedings.mlr.press/v97/fu19b.html

[20] Y. Fu, Y. Liu, J. Li, D. Luo, S. Lv, Y. Jv, and L. Xie, "Uformer: A unet based dilated complex & real dual-path conformer network for simultaneous speech enhancement and dereverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. IEEE, 2022, pp. 7417–7421. [Online]. Available: https://doi.org/10.1109/ICASSP43922.2022.9746020

[21] C. Chen, Y. Hu, H. Zou, L. Sun, and E. S. Chng, "Unsupervised noise adaptation using data simulation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[22] R. Chao, C. Yu, S. Fu, X. Lu, and Y. Tsao, "Perceptual contrast stretching on target feature for speech enhancement," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 5448–5452. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-10478

[23] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: conformer-based metric GAN for speech enhancement," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 936–940. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-517