# CITEAR: A Two-Stage End-to-End System for Noisy-Reverberant Hearing-Aid Processing

*Chi-Chang Lee*[1,2*], *Hong-Wei Chen*[3*], *Rong Chao*[2,4], *Tsung-Te Liu*[3], *Yu Tsao*[2]

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[3]Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan
[4]Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan, Taiwan

r08922a27@csie.ntu.edu.tw, r10943116@ntu.edu.tw, f14071075@gs.ncku.edu.tw,
ttliu@ntu.edu.tw, yu.tsao@citi.sinica.edu.tw

## Abstract

In this report, we present a hybrid neural network system on the task of the 2nd Clarity Enhancement Challenge. The system, consisting of two stages, handles noisy-reverberant corruption followed by post-processing to compensate for listener-specific hearing loss. For handling noisy-reverberant corruption, an end-to-end speech enhancement model was used. For post-processing, we designed an auditory correction (AC) module formed by a rule-based filter to reduce hearing loss effects. In our experiments, we analyzed the effectiveness of model architectures, amounts of training data, the head-rotation feature, and the post-processing module. The experimental results show that our proposed system can effectively reduce noisy-reverberant corruption and gain performance improvement toward listener-specific hearing loss.

**Index Terms**: speech enhancement, dereverberation, hearing-loss compensation

## 1. Introduction

The goal of speech enhancement (SE) is to map a distorted speech into its clean version. Various deep learning (DL) models have been used to formulate a regression function for SE [1–9]. In practice, an SE unit is commonly used as a pre-processor in speech-related applications, such as automatic speech recognition [10–12], speech emotion recognition [13], and hearing aids [14,15]. However, for the listening loss (HL) situation, listeners usually suffer from huge degradation of their perception. Since the HL effect occurs within individual listeners' ears, a vanilla SE unit cannot handle the HL effect without knowing its conditional information.

Thus far, numerous attempts have been made to reduce the hearing loss effect of specific listeners. In particular, the 1st Clarity Enhancement Challenge (CEC1) [16] is a competition that aims to build an SE system and improve the corresponding prediction's intelligibility (STOI) [17, 18] under a specific HL condition. To consider specific HL conditions, most approaches pass their output of SE models into the officially provided HL simulator [19] before calculating objectives, e.g., STOI loss [20]. These SE models can be roughly categorized into beamforming and non-beamforming approaches. For beamforming approaches [21–24], they take the two-stage strategy that first processing inputs by a beamformer with their own relative transfer function (RTF) schemes and then feed their outputs into the downstream neural network module. For non-beamforming approaches [25–28], they investigate different architectural designs of the end-to-end SE modules to directly convert multi-channel inputs into the prediction. All of these existing systems achieve notable improvements under the measurement of Modified Binaural STOI (MB-STOI) [29]. However, for measuring the hearing-aids performance, another well-known metric, Hearing-Aid Speech Perception Index (HASPI) [30], has not been considered in the challenge yet.

In this paper, we describe our submission for the 2nd Clarity Enhancement Challenge (CEC2) [16]. Our designed system is based on two stages. First, we apply an end-to-end SE model utilizing head-rotation features to reduce noisy-reverberant corruption. Next, we develop a general rule-based filter as post-processing for hearing-loss compensation, which corrects the denoised results based on a given audiogram, avoiding the need for fitting the model to a particular hearing-loss simulator and an objective. More details of our system will be illustrated in the next section.

## 2. Methodology

### 2.1. System overview

Fig. 1 shows the overall flow of our proposed system. Our system has two stages that respectively conduct speech enhancement and hearing-loss compensation. First, for speech enhancement, we aim to reduce the corruption made by noisy-reverberant environments and thus transform the low-quality signals into clean ones. To train the SE module, we prepare batch pairs of noisy-reverberant signals and their clean versions, estimating the six-channel inputs into the corresponding anechoic targets. For the binaural case in CEC2 [16], we separately train two SE modules for the left and right ears. Next, for hearing-loss compensation, we post-process the prediction results based on the first stage. To enrich the perception of the serving listener with particular hearing loss, by inputting the audiograms as conditional features, we design an auditory correction (AC) algorithm to amplify listener-specific frequency bins. Finally, we use the AC algorithm to refine the prediction from the SE module and thereby produce clear and highlighted signals that are more audible and more intelligible to the target listener. In the following sections, we will further describe the details of the architecture and objective of our SE system in Section 2.2. To the end, in Section 2.3, we will introduce the implementation of the post-processing toward the given audiogram features.

### 2.2. Head-rotation-aware speech enhancement

For speech enhancement, we focus on developing a denoising module to suppress interferences from both noise and speech sources made by reflection. We choose the work, SUccessive DOwnsampling and Resampling of Multi-Resolution Features (SuDoRM-RF) [31,32], as our architecture which has been widely performed in end-to-end audio source separation area. It is based

[*]These authors contributed equally to this work.

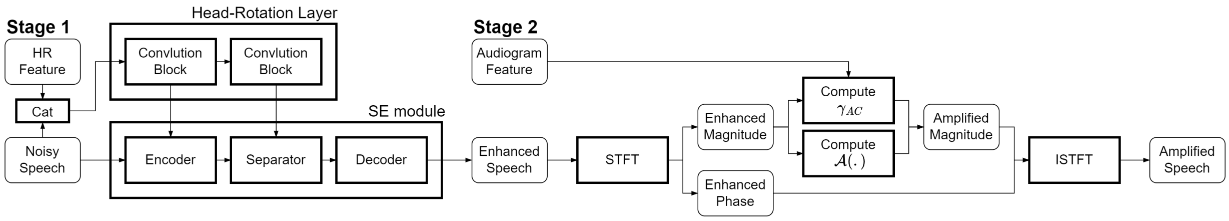

Figure 1: *Flowchart of our two-stage system.*

on an encoder-separator-decoder framework with specialized 1-D convolutional layers and layer normalization operations. The connection design is formed by the U-Net infrastructure [33]. Also, SuDoRM-RF [31, 32] directly processes inputs in the time domain, which usually requires a smaller size of kernels for operations. Following the 5 mini-second latency limitation announced by officials, the maximum kernel size in the 44.1kHz sampling rate should be lower than 220 sample points. In contrast, our maximum windowing is set to only 21 sample points of look-ahead. Thus, in our model, all processing blocks satisfy the computation requirements and thus act in a casual manner. In addition, we construct external 1-D convolutional layers to process the head-rotation feature. The six-channel input will be concatenated with the head-rotation feature to feed the external layers and then respectively add the output to the hidden maps in the front and middle levels of the main SE module. The combination flow of the two modules is also shown in Fig. 1.

Since the magnitude of signals will hugely affect the measurement of HASPI scores, we prefer using an objective that can preserve the consistency of signal-level information. Thus, referring to the previous work [25], we accordingly choose the negative value of signal-to-noise ratio (SNR) as our objective, namely the SNR loss:

$$\mathcal{L}(y, \hat{y}) = -10 \log_{10} \frac{||y||^2}{||y - \hat{y}||^2 + \tau ||y||^2} \tag{1}$$

where $\hat{y}$ is the estimated signal, and $y$ is its reference. $\tau = 10^{-\text{SNR}_{\text{max}}/10}$ is a soft threshold to prevent the issue of gradients dominating within a training batch [34]. Note that, according to [34], we set $\text{SNR}_{\text{max}}$ to 30dB.

### 2.3. Hearing-loss compensation

In our previous study, we proposed a perceptual contrast stretching (PCS) post-processing approach [35] to further improve the SE performance. The PCS approach is designed based on the critical importance function with the aim to sharpen the structures of enhanced speech and suppress residual noise. With the same idea, we propose an auditory correction (AC) algorithm as a post-processing method that combines hearing-loss compensation and stretching together. Based on each given audiogram, the specific hearing-loss pattern, the AC algorithm is designed to compensate for hearing loss by amplifying specific amplitude spectra through a gamma correction, thus further boosting the enhanced speech signals.

First, we focus on illustrating the mechanism of our amplification method. Given the prediction waveform made by denoising, we first process it via short-time Fourier transform (STFT) and take its magnitude as the feature, namely $X_{t,f}$. Then, referring to the previous work, PCS [35], we design the whole amplifying process as:

$$Y_{t,f} = \mathcal{A}(X_{t,f}) \cdot X_{t,f}^{\gamma} \tag{2}$$

where $Y_{t,f}$ denotes the modified magnitude at the $t$-th frame and $f$-th frequency bin. The value of input feature $X_{t,f}$ ranges from $[0, M]$. $\mathcal{A}$ and $\gamma$ are a scaling function and a gamma value, respectively; the scaling function $\mathcal{A}$ is defined as $\mathcal{A}(X_{t,f}) = (1 + 1/X_{t,f})^{\gamma} - (1/X_{t,f})^{\gamma}$, dynamically determined by $X_{t,f}$; the value of $\gamma$ ranges from $[0, 1]$.

Compared with conventional PCS [35], we schedule the gamma coefficients by considering the audiogram features. More specifically, given the hearing loss pattern at particular frequencies in an audiogram, we define our gamma value as $\gamma_{AC}[f]$, where $f$ represents the corresponding frequency bin. Based on the input format provided by CEC2 officials [16], the audiogram configuration is listed as [$F_0$=250, $F_1$=500, $F_2$=1000, $F_3$=2000, $F_4$=3000, $F_5$=4000, $F_6$=6000, $F_7$=8000], where the index ranges from 0 to 7. In the following, the determination of $\gamma_{AC}[f]$ can be divided into three steps.

In the first step, we adopt a rule function to mitigate hearing level in the audiogram as follows:

$$\hat{L}_i = \begin{cases} L_i - 15, & \text{if } L_i > 2L_{cut}, \\ L_i - 5, & \text{if } L_i > L_{cut}, \\ L_i, & \text{otherwise.} \end{cases} \tag{3}$$

where $\hat{L}_i$ and $L_i$ are modified and original hearing level of the frequency $F_i$, and $L_{cut}$ represents a cut value of hearing level. In this study, we set $L_{cut} = 30$dB.

In the second step, we scale the value of $\hat{L}_i$:

$$\bar{L}_i = \alpha \cdot \hat{L}_i + \delta \tag{4}$$

where $\bar{L}_i$ denotes the scaled result of the frequency $f_i$, and $\alpha$ and $\delta$ represents the scaling factor and bias value. In this study, we set $\alpha = 1/75$ and $\delta = 1$.

In the third step, we use the previously derived result $\bar{L}_i$ as features and thus design $\gamma_{AC}$ as:

$$\gamma_{AC}[f] = \begin{cases} \bar{L}_0, & \text{if } f \leq F_0, \\ \frac{(f - F_{i_l})}{(F_{i_u} - F_{i_l})}(\bar{L}_{i_u} - \bar{L}_{i_l}) + \bar{L}_{i_l}, & \text{if } f > F_0 \end{cases} \tag{5}$$

where $f$ denotes the frequency bin, and $i_l$, $i_u$ are the indexes of the audiogram configuration such that $i_l = \underset{i:f<F_i}{\text{argmax}}\{F_i\}$ and $i_u = \underset{i:f\geq F_i}{\text{argmin}}\{F_i\}$, respectively.

Finally, to prevent the values of $Y_{t,f}$ from exceeding the signal range, we first apply the Hardtanh function [36] as a clipping function on $Y_{t,f}$. After that, we take the clipped result of $Y_{t,f}$ and the corresponding phase component of $X_{t,f}$ to synthesize the eventual time-domain signal by the inverse STFT (ISTFT) operation.

## 3. Experiments

### 3.1. Experimental setup and implementation details

The datasets used in the experiments include the originally provided scene signals and the extended data made by the official mixing tool, consisting of 6,000 and 36,000, respectively. During training, all the data loading processes would conduct shifting and truncation operations for augmentation and thus enrich sample diversities.

The head-rotation-aware SE system has three training stages. First, we trained our main SE module with the original scene data. Next, we froze the main SE module and externally trained the head-rotation layer with the original scene data. Finally, we jointly trained the main SE module and the external head-rotation layer with the combination of original and extended scene data. To train the main SE module in the initial stage, we choosed the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a learning rate of $5 \times 10^{-4}$, and a batch size of 8 with 500,000 steps. To train the external head-rotation layer, we choosed the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a learning rate of $5 \times 10^{-6}$, and a batch size of 8 with 100,000 steps. To jointly train the SE module and the head-rotation layer in the final stage, we choosed the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a learning rate of $5 \times 10^{-6}$, and a batch size of 8 with 300,000 steps.

For the architectural detail of the SE module (SuDoRM-RF), the separation module has 512 input channels, 256 output channels, up-sampling depth of 5, and 16 blocks; the encoder module has 6 input channels, 512 output channels, and a kernel size of 21; the decoder module has 512 input channels, 1 output channels, and a kernel size of 21. For the architectural detail of the head-rotation layer, the number of input channels is 7, the output channel for the front level is 512, and the output channel for the middle level is 256.

### 3.2. Experimental results

This section describes the evaluation in the development set of different previous works and various setups. Here, we use the HASPI score as the metric measurement, ranging from 0 to 1.

#### *3.2.1. Comparison with other models*

Table 1 shows the HASPI score with respect to different other models. **BSLN** represents the official baseline system provided by CEC2 [16]. **CTASN** represents the Conv-Tasnet [37] model trained with the original scene data. **SUDO** represents the SuDoRM-RF [31, 32] model trained with the original scene data.

Table 1: *HASPI results of the comparison with other models.*

| method | BSLN | CTASN | SUDO |
|---|---|---|---|
| HASPI | 0.2493 | 0.4344 | 0.4756 |

From Table 1, we can see that **SUDO** achieved the best result. Moreover, in their original paper [31], they also listed a complete study of the computation requirement and showed the notable efficiency improvement with respect to **CTASN**. Therefore, we choose **SUDO** as the architecture of our main SE module and accordingly combine it with other setups.

#### *3.2.2. Ablation study*

Table 2 shows the HASPI score with respect to different setups introduced in Section 2. The postfix term denotes the corresponding setup. **-hr** represents taking the head-rotation feature and combining the SE module with the external head-rotation layer. **-ext** represents training our model with the combination of original and extended scene data. **-ac** represents post-processing the denoising results by the AC algorithm. As we see in Table 2, all the additional setups can bring improvements. In particular, extending training data achieved a difference of 0.0389, representing that expanding the amounts of data can effectively reduce the over-fitting effect. On the other hand, the head-rotation feature is highly related to the spatial relationship of the recording environment, which is critical information and thus stably raises the score. Eventually, we utilize the audiogram feature to amplify the energy in the corresponding frequency bin, successfully reducing the degradation of HASPI from the HL effect. To sum up, it is clear that the integration in our system can obviously increase the HASPI score, confirming the effectiveness of each component.

Table 2: *HASPI results of the ablation study.*

| method | HASPI |
|---|---|
| SUDO | 0.4756 |
| SUDO-hr | 0.4823 |
| SUDO-ext | 0.5145 |
| SUDO-hr-ext | 0.5299 |
| SUDO-ext-ac | 0.5352 |
| SUDO-hr-ext-ac | 0.5467 |

## 4. Concluding Remarks

In this short report, we described each component in our hybrid system submitted to the 2nd Clarity Enhancement Challenge [16]. Our system can be mainly divided into the head-rotation-aware speech enhancement system and the hearing-loss compensation. In Section 2.2, we detailedly explained the design of our denoising module, effectively reducing the degradation made by interference. In Section 2.3, we clearly expressed each step of the developed AC algorithm, successfully amplifying the energy in the listener-specific frequency bin. In addition, the effectiveness of each component has been verified in Section 3.2, especially our AC algorithm, which stably boosted the performance without fitting our model to a particular hearing-loss simulator and an objective. To sum up, our system notably brings better perception under HL situations, being more clear, more audible, and more intelligible to the target listener.

## 5. References

[1] S. Wang, W. Li, S. M. Siniscalchi, and C. Lee, "A cross-task transfer learning approach to adapting deep speech enhancement models to unseen background noise using paired senone classifiers," in *Proc. of ICASSP*, 2020.

[2] Y.-C. Lin, Y.-T. Hsu, S.-W. Fu, Y. Tsao, and T.-W. Kuo, "IA-NET: Acceleration and compression of speech enhancement using integer-adder deep neural network," in *Proc. of Interspeech*, 2019.

[3] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Proc. of Interspeech*, 2013.

[4] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM TASLP*, vol. 23, no. 1, pp. 7–19, 2014.

[5] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Proc. of Interspeech*, 2014.

[6] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. of ICASSP*, 2015.

[7] A. Kumar and D. Florencio, "Speech enhancement in multiple-noise conditions using deep neural networks," in *Proc. of Interspeech*, 2016.

[8] C.-C. Lee, Y.-C. Lin, H.-T. Lin, H.-M. Wang, and Y. Tsao, "SERIL: noise adaptive speech enhancement using regularization-based incremental learning," in *Proc. of Interspeech*, 2020.

[9] C.-C. Lee, C.-H. Hu, Y.-C. Lin, C.-S. Chen, H.-M. Wang, and Y. Tsao, "Nastar: Noise adaptive speech enhancement with target-conditional resampling," *arXiv preprint arXiv:2206.09058*, 2022.

[10] A. Nicolson and K. K. Paliwal, "Deep Xi as a front-end for robust automatic speech recognition," in *Proc. of CSDE*, 2020.

[11] Z. Meng, S. Watanabe, J. R. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *Proc. of ICASSP*, 2017.

[12] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *Latent Variable Analysis and Signal Separation*, E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský, Eds. Springer International Publishing, 2015, pp. 91–99.

[13] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. W. Schuller, "Towards robust speech emotion recognition using deep residual networks for speech enhancement," in *Proc. of Interspeech*, 2019.

[14] K. Tan, X. Zhang, and D. Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in *Proc. of ICASSP*, 2019.

[15] I. Fedorov, M. Stamenovic, C. Jensen, L. Yang, A. Mandell, Y. Gan, M. Mattina, and P. N. Whatmough, "Tinylstms: Efficient neural speech enhancement for hearing aids," in *Proc. of Interspeech*, 2020.

[16] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, R. Viveros Munoz *et al.*, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Proc. of Interspeech*, 2021.

[17] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. of ICASSP*, 2010.

[18] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM TASLP*, vol. 24, pp. 2009–2022, 2016.

[19] Z. Tu, N. Ma, and J. Barker, "Optimising hearing aid fittings for speech in noise with a differentiable hearing loss model," in *Proc. of Interspeech*, 2021.

[20] S. Fu, T. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM TASLP*, vol. 26, no. 9, pp. 1570–1584, 2018.

[21] A. H. Moore, S. Hafezi, R. Vos, M. Brookes, P. A. Naylor, M. Huckvale, S. Rosen, T. Green, and G. Hilkhuysen, "Elo-spheres consortium system description," *Proc. of Clarity*, 2021.

[22] K. Zmolikova and J. Cernock, "But system for the first clarity enhancement challenge," *Proc. of Clarity*, 2021.

[23] S. J. Yang, S. Wisdom, C. Gnegy, R. F. Lyon, and S. Savla, "Listening with googlears: Low-latency neural multiframe beamforming and equalization for hearing aids," *Proc. of Clarity*, 2021.

[24] M. Tammen, H. Gode, H. Kayser, E. J. Nustede, N. L. Westhausen, J. Anemüller, and S. Doclo, "Combining binaural lcmp beamforming and deep multi-frame filtering for joint dereverberation and interferer reduction in the clarity-2021 challenge," *Proc. of Clarity*, 2021.

[25] Z. Tu, J. Zhang, N. Ma, and J. Barker, "A two-stage end-to-end system for speech-in-noise hearing aid processing," *Proc. of Clarity*, 2021.

[26] X. Chen, Y. Shi, W. Xiao, M. Wang, T. Wu, S. Shang, N. Zheng, and Q. Meng, "A cascaded speech enhancement for hearing aids in noisy-reverberant conditions," *Proc. of Clarity*, 2021.

[27] T. Gajecki and W. Nogueira, "Binaural speech enhancement based on deep attention layers," *Proc. of Clarity*, 2021.

[28] P. Kendrick and M. Tribe, "Hearing aid speech enhancement using u-net convolutional neural networks," *Proc. of Clarity*, 2021.

[29] 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*, 2018.

[30] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (haspi)," *Speech Communication*, 2014.

[31] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm-rf: Efficient networks for universal audio source separation," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2020.

[32] E. Tzinis, Z. Wang, X. Jiang, and P. Smaragdis, "Compute and memory efficient universal sound source separation," *Journal of Signal Processing Systems*, 2022.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015.

[34] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, and J. R. Hershey, "Unsupervised sound separation using mixture invariant training," in *Proc. of NeurIPS*, 2020.

[35] R. Chao, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Perceptual contrast stretching on target feature for speech enhancement," *Proc. of INTERSPEECH*, 2022.

[36] "Hardtanh — PyTorch 1.12 documentation, howpublished = https://pytorch.org/docs/stable/generated/torch.nn.hardtanh.html,."

[37] 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*, 2019.