Band-wise Front-end Distortion Suppression for Robust Speech Recognition

Siyi Zhao, Wei Wang, Yanmin Qian †

Auditory Cognition and Computational Acoustic Lab
MoE Key Lab of Aritificial Intelligence, AI Institute
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai
{zsy_coding, wangwei.sjtu, yanminqian}@sjtu.edu.cn

2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP) | 979-8-3315-1682-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISCSLP63861.2024.10800235

Abstract

Advancements in deep learning techniques have significantly improved automatic speech recognition (ASR). However, improving robustness to acoustic interferences, such as background noise and reverberation, remains crucial for accurate speech recognition in noisy environments. Traditional speech enhancement (SE) modules used as pre-processing front-ends often fail to integrate well with ASR models due to intelligibility distortions introduced by the SE module. Inspired by the observation adding technique and band-wise network design of recent SE models, we propose a novel adaptive band-wise observation adding (BOA) module for front-end distortion suppression. The lightweight BOA network learns time- and frequency-variant OA factors, adaptively balancing the contributions of noisy and enhanced speech signals for diverse noisy conditions. Experiments conducted on both simulated and real-world test sets demonstrate that the BOA module produces more effective dynamic OA factors under various conditions compared to fixed OA factors, leading to improved ASR performance.

Index Terms: robust speech recognition, speech enhancement, observation adding

1. Introduction

Over the past decade, significant advancements in deep learning methodologies have driven substantial progress in automatic speech recognition (ASR). A crucial focus within ASR research has been improving the robustness of speech recognition to address various acoustic interferences, such as background noise and reverberation [1, 2, 3, 4, 5]. This area of research aims to ensure accurate and efficient speech recognition despite the presence of diverse and challenging noise conditions. Therefore, developing robust speech recognition techniques is essential for improving ASR system performance in real-world, noisy environments.

One approach to achieving robust speech recognition is to use a speech enhancement (SE) module as a pre-processing front-end to reduce noise within the speech signal. However, numerous prior studies have observed that the enhanced speech output from SE does not consistently result in optimal recognition accuracy for subsequent ASR tasks [6, 7, 8, 9]. This inconsistency is often attributed to sub-optimal intelligibility distortions introduced during the enhancement process.

To address this issue, [10, 11] proposed joint training of the SE and ASR modules using ASR objectives, thereby minimizing the loss of linguistic information caused by distortions during SE. However, this joint training approach is computationally expensive and challenging to converge. Additionally, the resulting SE and ASR modules are deeply coupled, making it difficult to adapt the SE module to different ASR models or integrate various SE modules into the ASR system.

In [12], the authors introduced a simple yet effective approach to mitigate these challenges through the observation adding (OA) technique. The OA technique adds the enhanced speech output from SE to the original noisy speech input. By heuristically searching for the optimal weight to balance the contributions of the noisy and enhanced speech signals through multiple decodings on the development set, the OA technique has demonstrated improvements in SE and ASR cascaded system performance. Moreover, it offers a more flexible and efficient method for integrating SE modules into ASR systems. Nevertheless, the optimal OA factor can vary significantly across different noisy conditions, and an inappropriate OA factor can lead to substantial performance discrepancies. This variability necessitates determining a fixed OA factor for each noisy condition, which is impractical for real-world applications.

In this paper, inspired by the observation adding (OA) technique and the band-wise network design of recent SE models [13, 14, 15, 16], we propose a band-wise observation adding (BOA) module. This module employs a very lightweight network to learn time- and frequency-variant OA factors. The BOA module is designed to adaptively adjust the contributions of the noisy and enhanced speech signals across different frequency bands and time frames. We start with off-the-shelf SE and ASR modules and train the BOA module end-to-end with these modules, freezing all parameters except for those in the BOA module. The training is conducted on a simulated noisy dataset with a wide range of signal-to-noise ratios. Due to the small number of learnable parameters, the training process can be completed in very few steps. Once trained, the BOA module dynamically optimizes the OA factor for varying conditions, improving the overall performance of the ASR system.

2. Methodology

2.1. Band-Split RNN (BSRNN)

Although the proposed method can be applied to any SE and ASR models, we focus on the BSRNN model [13], which performs interleaved sequence-level and band-level processing, making it suitable for speech enhancement.

The BSRNN model processes the input mixture by separating the complex-valued spectrogram X of the input into K predefined subbands of different bandwidths. A real-valued subband feature Z_i is then generated by concatenating the real and imaginary parts of each subband spectrogram B_i.

Given a fullband feature Z merged with these subband features, BSRNN employs a dual-path time-frequency RNN architecture to capture both inter-band and intra-band features. The subsequent steps involve sequence modeling along the time dimension T and band modeling along the band dimen-

†:Corresponding Author

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 09,2026 at 05:18:20 UTC from IEEE Xplore. Restrictions apply.

sion K, utilizing RNNs and repeated L times. Each of the K bands is then processed by band-specific MLPs and concatenated to form the predicted complex-valued time-frequency mask $M \in \mathbb{C}^{F\times T}$ and the complex-valued residual spectrogram $R \in \mathbb{C}^{F\times T}$. The output speech spectrogram S is obtained by taking the complex-valued Hadamard product of M and X, and combining it with the residual spectrogram R.

2.2. Band-Wise Observation Adding (BOA)

Artifact errors introduced by single-channel SE front-ends are a potential cause of ASR performance degradation in noisy conditions. To address this, a simple yet effective method called observation adding (OA) has been proposed [17].

OA involves adding a scaled version of the observed signal back to the enhanced signal:

$$S_{OA} = \omega_{OA} \cdot S_{noisy} + S_{enh} \tag{1}$$

where S_{OA} is the modified enhanced signal, S_{noisy} is the observed signal, and S_{enh} is the enhanced signal. The scaling factor is denoted as ω_{OA}. This approach can also be applied to the intermediate feature representations of these three types of speech.

To improve ASR performance while maintaining model efficiency, we propose the band-wise observation adding (BOA) module, as shown in the grey block in Figure 1. This lightweight module can be easily integrated into different SE models.

The BOA module takes the output of the final layer of the band and sequence modeling modules and feeds it into a bidirectional long short-term memory (BiLSTM) layer followed by a fully connected (FC) layer. The sigmoid output layer generates $OA \in \mathbb{R}^{K\times T}$, which represents the time- and frequency-variant OA factor corresponding to the target spectrogram S:

$$OA = \text{Sigmoid}(\text{FC}(\text{BiLSTM}(Q))) \tag{2}$$

Figure 1 illustrates the structure of the BOA module, the calculation process of time- and frequency-variant OA factors, and the application of OA to the target spectrogram $S \in \mathbb{C}^{F\times T}$.

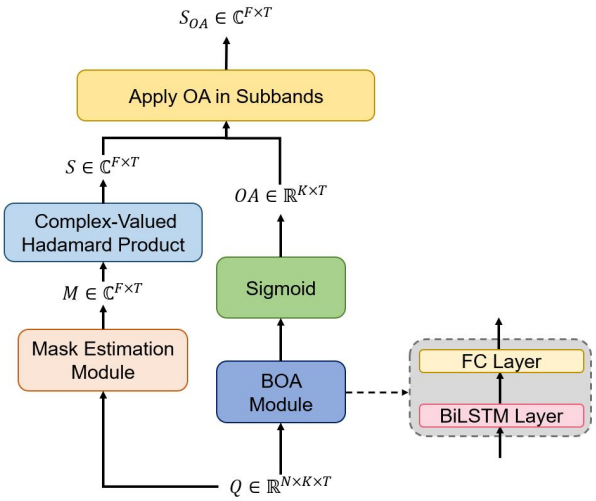


Figure 1: *Process of calculating time- and frequency-variant OA factors through the BOA module for enhanced speech.*

In the original BSRNN model, the fullband feature tensor Z passes through stacked sequence-level and band-level RNNs, resulting in tensor $Q \in \mathbb{R}^{N\times K\times T}$, which serves as the input to the mask estimation module.

In the proposed architecture, the output of the stacked time and frequency RNN modules is fed into the BOA module, which then outputs the corresponding OA factor $OA \in \mathbb{R}^{K\times T}$.

For extracting the target source, the mask estimation module computes a fullband complex-valued time-frequency (T-F) mask and multiplies it with the input spectrogram X:

$$\bar{S} = M \odot X + R \tag{3}$$

where R represents the residual part.

The OA mix module, shown in Figure 1, combines the original and enhanced input tensors using the OA factor to generate mixed speech across the defined subbands.

Denote the subband frequency ranges at time t as $\{(f^{i,t}_{\text{start}}, f^{i,t}_{\text{end}})\}_{i=1}^{K}$, where K is the number of subbands. For each subband $i \in \{1, 2, ..., K\}$, the output tensor X_{oa} is calculated by combining the noisy and enhanced inputs weighted by the OA factor. The procedure is formally demonstrated in Algorithm 1.

Algorithm 1 Applying Observation Adding (OA) in Subbands

Input: $X_{ori} \in \mathbb{R}^{B\times T\times F}$
Input: $X_{enh} \in \mathbb{R}^{B\times T\times F}$
Input: $\text{OA} \in \mathbb{R}^{B\times T\times K}$
Output: $X_{OA} \in \mathbb{R}^{B\times T\times F}$

for t **in** $1 \ldots T$ **do**
  for i **in** $1 \ldots K$ **do**
    for $(f^{i,t}_{start}, f^{i,t}_{end})$ **in** freqency bands **do**
      $\text{OA}_{band} \leftarrow \text{OA}[f^{i,t}_{start} : f^{i,t}_{start}]$
      $X_{\text{OA}}[f^{i,t}_{start} : f^{i,t}_{start}] \leftarrow \text{OA}_{band} \cdot X_{ori}[f^{i,t}_{start} : f^{i,t}_{start}] + (1 - \text{OA}_{band}) \cdot X_{enh}[f^{i,t}_{start} : f^{i,t}_{start}]$
    end for
  end for
end for
return X_{OA}

3. Experiment

3.1. Datasets

We compare the effects of the fixed OA technique and the proposed method using both simulated and real-world noisy data. Three widely adopted corpora in the noise-robust ASR field are used: LIBRISPEECH [18], WHAM! [19], and the 1-channel track of CHIME-4 [20].

The simulated dataset includes train, dev, and test sets, synthesized by mixing the `train-clean-360`, `dev-clean`, and `test-clean` subsets of LIBRISPEECH with WHAM! noise. The total duration is 470 hours, with noise randomly selected at Signal-to-Noise Ratios (SNRs) between -10 to 10 dB.

For testing on simulated noisy data, the test set is partitioned based on SNR ranges, resulting in four test sets corresponding to the ranges: [-10, 0) dB, [0, 5) dB, [5, 10) dB, and [10, 20) dB.

For testing on real-world noisy data, we use the official CHIME-4 1-channel test sets, `dt_05_real` 1ch track and `et_05_real` 1ch track. These are recombined according to different noise environments (bus, cafe, pedestrian area, and

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 09,2026 at 05:18:20 UTC from IEEE Xplore. Restrictions apply.

street junction), with each test subset containing 760 utterances. During the training phase, we use the default single-channel synthesized speech training and validation sets from CHiME-4.

3.2. Experimental Setup

3.2.1. Training

All experiments were conducted using the ESPnet toolkit [21]. For BSRNN, we followed the setup in [13] for band splitting. Frequency bands below 16 kHz were split by a 1 kHz bandwidth, bands between 16 kHz and 20 kHz were split by a 2 kHz bandwidth, and the remaining frequencies were treated as one subband, resulting in a total of 19 subbands.

For both datasets, SE front-ends were trained for 100 epochs. Models were trained using the Adam optimizer [22] with an initial learning rate of 10^{-3} and a weight decay factor of 10^{-5}. The 16 kHz speech signal was transformed from the time domain to the time-frequency domain using Short-Time Fourier Transform (STFT) with a window size of 320 and a hop length of 160. This produced a spectrogram subsequently processed by the BSRNN. Multi-resolution STFT magnitude loss [23] and SI-SNR loss were used as the criteria for SE front-end training.

3.2.2. ASR Backend

The ASR backends used for both simulated and real-world data were publicly available Attention-Based Encoder-Decoder (AED) models using ESPnet's default configurations. The model trained on the CHiME-4 dataset[1] was used for real-world data, and the model trained on the LibriSpeech dataset[2] was used for simulated data.

3.2.3. Fixed OA Factor Search

For experiments involving fixed OA factors, we searched for the optimal OA factor on the development set within the range [0, 1] with a step size of 0.1. In practice, we found that OA factors larger than 0.3 consistently led to more severe degradation and, therefore, are not included in the experiments.

4. Results and Analysis

4.1. Results on Synthesized Noisy Speech

Table1 and Figure 2 present the performance of SE front-ends and OA related methods on the simulated noisy test set using the same ASR backend.

Table 1: *WER[%] results on* LibriSpeech *and* WHAM! *simulated noisy test datasets.*

Method	$-10 \sim 0$ dB	$0 \sim 5$ dB	$5 \sim 10$ dB	$10 \sim 20$ dB
1. Noisy	42.9	21.8	9.3	8.0
2. BSRNN	10.4	6.4	3.9	3.2
3a. + OA = 0.1	13.3	7.4	4.1	3.9
3b. + OA = 0.2	10.8	6.3	3.7	3.1
3c. + OA = 0.3	18.4	10.3	5.4	4.6
4. + BOA	**10.3**	**6.2**	**3.6**	**2.8**

[1]https://huggingface.co/pyf98/chime4_conformer_e12_linear2048

[2]https://huggingface.co/espnet/Shinji_Watanabe_librispeech_asr_train_asr_transformer_e18_raw_bpe_sp_valid.acc.best

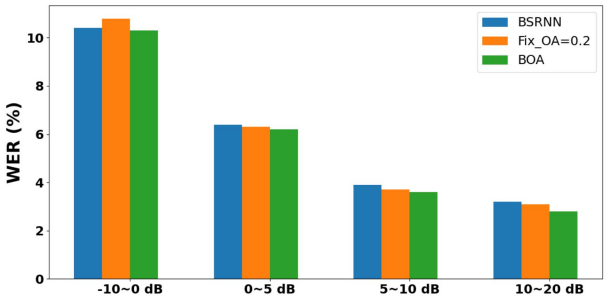


Figure 2: *Comparison of ASR performance using different methods on simulated test set at four SNR ranges.*

As shown in Table 1, comparing results 3a, 3b, and 3c, different fixed OA values can significantly impact the ASR performance. Moreover, although the optimal OA value searched on the development set is 0.1 (not shown in the table), the best performance on the test set is achieved with an OA value of 0.2. This discrepancy highlights the sensitivity of the OA factor across different datasets. In contrast, the BOA method fine-tunes its OA computation module on the training set and achieves better results than the best fixed OA value of 0.2 on the test sets. Comparing the results on the test sets of different SNRs, BOA shows more significant improvement at higher SNRs, likely because distortions introduced by the BSRNN SE frontend have a more pronounced negative impact when the original speech is relatively clean.

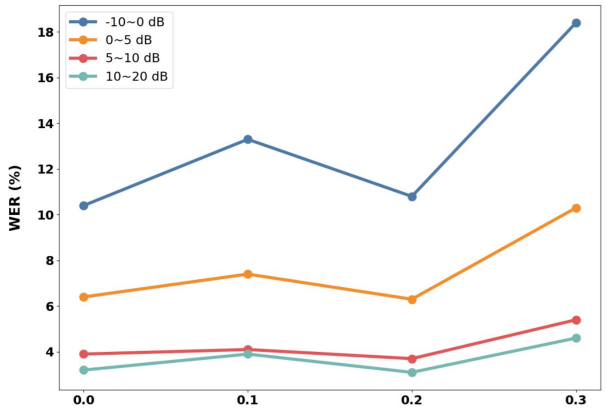


Figure 3: *Comparison of ASR performance using different fixed OA factors on simulated test sets at four SNR ranges.*

4.2. Results on Real-world Noisy Speech

As presented in Table 2, we carry out experiments on 1-channel real-world noisy speech from the CHiME-4 challenge, which is split into four subsets (bus, caf, ped, and str) according to acoustic scenes.

As shown in Table 2, in real-world scenarios, the distortion introduced by the SE front-end is more significant. The BSRNN-enhanced speech consistently shows worse performance. Applying a fixed OA factor mitigates the degradation but still results in worse ASR performance compared to the original speech. By introducing the BOA module, performance consistently improves compared to the fixed OA results. In all scenarios except the 'bus' scenario, the negative impact intro-

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 09,2026 at 05:18:20 UTC from IEEE Xplore. Restrictions apply.

Table 2: *WER[%] results on* CHiME-4 *real noisy test datasets.*

Method	bus	caf	ped	str	avg
Noisy	**23.8**	20.5	18.4	15.9	19.7
BSRNN	37.7	25.7	22.1	19.1	26.2
+ OA = 0.3	29.3	21.9	18.8	17.7	21.9
+ BOA	24.0	**18.8**	**17.0**	**15.9**	**18.9**

duced by SE distortion is completely eliminated.

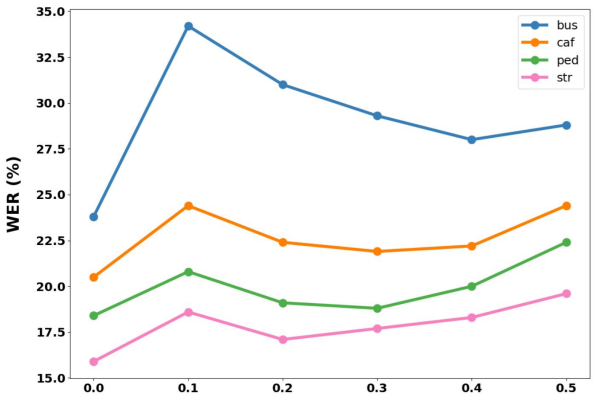


Figure 4: *Comparison of ASR performance using different fixed OA factors on real-word test set diveded by acoustic scenes.*

5. Conclusion

In this paper, we aimed to improve the noise robustness of ASR systems by using SE modules as front-ends while mitigating the distortions introduced by the SE modules. We proposed a band-wise observation adding (BOA) module inspired by the band-wise network design of recent SE models and the observation adding technique. The BOA module employs a lightweight network to learn time- and frequency-variant OA factors, adaptively balancing noisy and enhanced speech signals. Trained end-to-end with off-the-shelf SE and ASR modules, the BOA module improves ASR performance under various noisy conditions. Experiments on simulated LIBRISPEECH and real-world CHiME4 test sets demonstrate that the BOA module outperforms fixed OA factor approaches, leading to improved ASR performance.

6. Acknowledgements

This work was supported in part by the Key Research and Development Program of Jiangsu Province, China (Grant No. BE2022059-4), in part by China NSFC projects under Grants 62122050 and 62071288, and in part by Shanghai Municipal Science and Technology Commission Project under Grant 2021SHZDZX0102.

7. References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[2] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 9, no. 5, pp. 1–28, 2018.

[3] Y. Hu, N. Hou, C. Chen, and E. S. Chng, "Interactive feature fusion for end-to-end noise-robust speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6292–6296.

[4] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-task self-supervised learning for robust speech recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6989–6993.

[5] W. Wang and Y. Qian, "Hubert-agg: Aggregated representation distillation of hidden-unit bert for robust speech recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] K. Iwamoto, T. Ochiai, M. Delcroix, R. Ikeshita, H. Sato, S. Araki, and S. Katagiri, "How bad are artifacts?: Analyzing the impact of speech enhancement errors on asr," 2022.

[7] P. Wang, K. Tan *et al.*, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 39–48, 2019.

[8] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 1, pp. 47–56, 2010.

[9] D. Yang, W. Wang, and Y. Qian, "Fat-hubert: Front-end adaptive training of hidden-unit bert for distortion-invariant robust speech recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.

[10] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, 2016.

[11] B. Liu, S. Nie, S. Liang, W. Liu, M. Yu, L. Chen, S. Peng, C. Li *et al.*, "Jointly adversarial enhancement training for robust end-to-end speech recognition." in *Interspeech*, 2019, pp. 491–495.

[12] T. Ochiai, K. Iwamoto, M. Delcroix, R. Ikeshita, H. Sato, S. Araki, and S. Katagiri, "Rethinking processing distortions: Disentangling the impact of speech enhancement errors on speech recognition performance," *arXiv preprint arXiv:2404.14860*, 2024.

[13] J. Yu and Y. Luo, "Efficient monaural speech enhancement with universal sample rate band-split rnn," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[14] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6633–6637.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 09,2026 at 05:18:20 UTC from IEEE Xplore. Restrictions apply.

[15] J. Chen, W. Rao, Z. Wang, J. Lin, Z. Wu, Y. Wang, S. Shang, and H. Meng, "Inter-subnet: Speech enhancement with subband interaction," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] X. Hao and X. Li, "Fast fullsubnet: Accelerate full-band and sub-band fusion model for single-channel speech enhancement," *arXiv preprint arXiv:2212.09019*, 2022.

[17] K. Iwamoto, T. Ochiai, M. Delcroix, R. Ikeshita, H. Sato, S. Araki, and S. Katagiri, "How bad are artifacts?: Analyzing the impact of speech enhancement errors on asr," *arXiv preprint arXiv:2201.06685*, 2022.

[18] 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, pp. 5206–5210.

[19] 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," *arXiv preprint arXiv:1907.01160*, 2019.

[20] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[21] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] Y.-J. Lu, S. Cornell, X. Chang, W. Zhang, C. Li, Z. Ni, Z.-Q. Wang, and S. Watanabe, "Towards low-distortion multi-channel speech enhancement: The espnet-se submission to the l3das22 challenge," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9201–9205.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 09,2026 at 05:18:20 UTC from IEEE Xplore. Restrictions apply.