ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746981

# MULTI-STAGE AND MULTI-LOSS TRAINING FOR FULLBAND NON-PERSONALIZED AND PERSONALIZED SPEECH ENHANCEMENT

*Lianwu Chen, Chenglin Xu, Xu Zhang, Xinlei Ren, Xiguang Zheng, Chen Zhang, Liang Guo, Bing Yu*

Kuaishou Technology, Beijing, China

## ABSTRACT

Deep learning-based wideband (16kHz) speech enhancement approaches have surpassed traditional methods. This work further extends the existing wideband systems to enable fullband (48kHz) speech enhancement while simultaneously ensuring automatic speech recognition compatibility and optionally, personalized speech enhancement. As shown in the evaluation results, this is achieved by employing a multi-stage and multi-loss training architecture that incorporates the recently proposed two-step structure, ASR loss produced by a back-end ASR encoder, and the speaker extraction network.

***Index Terms***— speech enhancement, noise suppression

## 1. INTRODUCTION

In recent years, deep neural network (DNN) based speech enhancement (SE) methods have achieved significant improvement over the traditional signal processing based methods. While recent systems [1, 2, 3] are performed well for real-time wideband (16kHz) SE tasks, a multi-stage and multi-loss based SE system is proposed in this work to simultaneously support fullband (48kHz) SE, automatic speech recognition (ASR) compatibility, and optionally, personalized SE.

The existing fullband SE systems generally employ psychoacoustic motivated models [4, 5] to reduce the frequency-wise feature dimensions. Compared to the existing fullband SE approaches, this work employs and extends our recently proposed multi-stage fullband SE structure [6], where a computationally efficient highband (16-48kHz) system is built on top of a wideband (0-16kHz) system. For the highband system, in addition to the highband noisy input signal, the highband system is also informed by the output of the wideband system and thus achieved better fullband SE performance compared to the existing systems. For the wideband system, it extends from [6] by adding a parallel and separately trained temporal convolutional neural network (TCNN) based SE network [7] alongside the convolution recurrent network (CRN) based SE network [8] as employed in [6]. The output of these two wideband systems is jointly processed by a fusion network (FN) to obtain the final wideband SE output. The idea of model fusion has been widely used for music source separation [9] and audio scene classification [10]. This work also demonstrates its advantage for the SE task.

Employing the multi-stage training structure also makes it easier to incorporate the ASR compatibility with the SE system in that the common ASR systems also take the wideband signals as the inputs. Existing approaches have concluded that the ASR performance can degrade when blindly feeding the output of the SE system to the ASR system, while it can be improved when employing a pre-trained back-end ASR system during the SE model training stage [11, 12, 13]. For the wideband SE model, we employ a multi-loss strategy formed by an SE loss and an ASR loss to simultaneously ensure good SE and ASR performance. The SE loss consists by an Ideal Amplitude Mask weighted Mean Absolute Logarithmic Error (IAM-MALE) loss [3] for noise suppression, a Scale-Invariant Signal-to-Distortion Ratio (SISDR) loss for speech preservation, and a wav2vec loss [14] using a pre-trained wav2vec network [15]. The ASR loss is formed by an ASR feature embedding loss obtained using the encoder portion of the WeNet system [16].

The output of the wideband SE subsystem can optionally be further enhanced by a speaker extraction network given a corresponding target speaker embedding vector for personalized SE. The speaker extraction network mimics a human's selective auditory attention that only passes through the speech that matches the reference voice of the target speaker. Both the speaker extraction network and the speaker embedding network adopt CRN structure. Besides the cross-entropy (CE) loss for speaker classification in speaker embedding network, the personalized SE also takes the joint SE loss and ASR loss same as that in the non-personalized SE.

The proposed system is evaluated using wideband and narrowband PESQ [17] for SE performance and word error rate (WER) for ASR performance. The proposed system is ranked the 11th place for non-personalized SE and the 4th place for personalized SE of the ICASSP2022 DNS Challenge [18].

## 2. SYSTEM OVERVIEW

The proposed multi-stage and multi-loss based fullband SE system is shown in Figure 1. The input noisy signal $y$ is transformed into a time-frequency domain representation via STFT and then divided into two parts, namely, $Y_{16}$ and $Y_{16-48}$, representing the wideband and highband time-frequency component below 8kHz and between 8kHz and 24kHz, respectively. The wideband signal $Y_{16}$ is then fed into two wideband SE systems to obtain enhanced signals. A

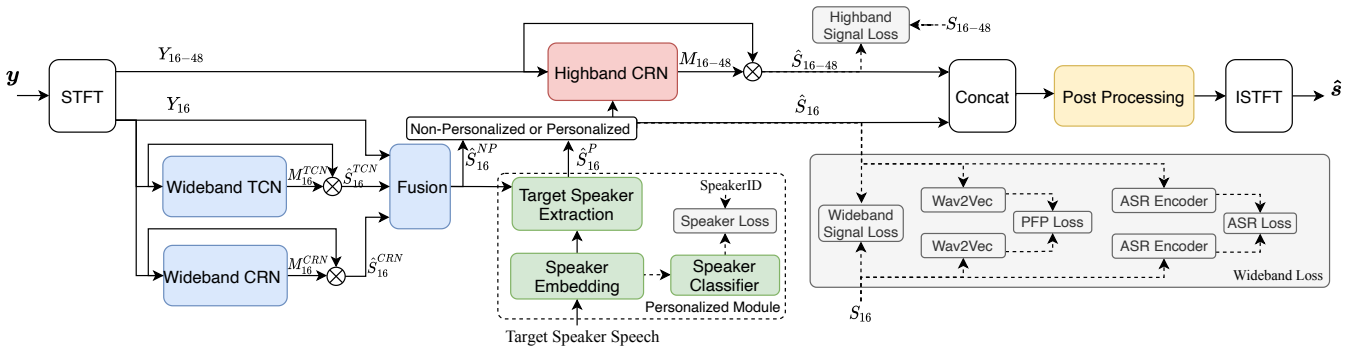


**Fig. 1**. The proposed system architecture for fullband speech enhancement.

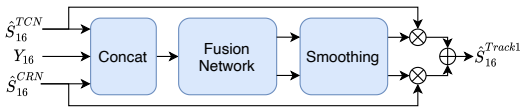


**Fig. 2**. The proposed wideband model fusion structure.

fusion network is proposed to fuse the enhanced wideband signals of two individual SE systems.

For non-personalized SE, the output of fusion module $\hat{S}_{16}^{NP}$ is selected as $\hat{S}_{16}$ and fed into the highband CRN module, which estimates the highband mask $M_{16-48}$ based on the input of noisy highband signal $Y_{16-48}$ and enhanced wideband signal $\hat{S}_{16}$. The enhanced $\hat{S}_{16}$ and $\hat{S}_{16-48}$ are concatenated, and a post-processing module is used to further suppress the residual noise. For personalized SE, the output of the fusion module is fed into the target speaker extraction module, which extracts the target speech $\hat{S}_{16}^{P}$ based on the embedding of the target speaker. Same as the non-personalized fullband SE pipeline, the highband and post-processing modules are utilized to generate the final fullband enhanced speech of the target speaker. Reference clean speech is divided into $S_{16}$ and $S_{16-48}$ for wideband and highband loss calculation, respectively.

## 3. NON-PERSONALIZED SPEECH ENHANCEMENT

### 3.1. Wideband enhancement networks

Two neural network modules with CRN and TCN structure are proposed for the wideband speech enhancement. The same CRN structure proposed in [3] is used for estimating $M_{16}^{CRN}$, which nested a RNN module inside the CNN-based encoder-decoder structure.

The wideband TCN module consists of a CNN-based encoder and a TCM [7] based mask estimator. The encoder consists of three conv2D layers with 90 channels. The kernel sizes are [1,5], [1,3] and [1,3]. The strides are [1,3], [1,2] and [1,2]. The output of the encoder is converted to a one-dimensional signal of size T × 512 by a Dense layer. The TCM operates on T × 512 and produces an output of the same size. The TCM has two causal dilation blocks stacked together. A dilation block is formed by stacking seven residual blocks [7] with exponentially increasing dilation rates (1, 2, 4, 8, 16, 32, 64). The filterSize and outputChannels of TCM [7] are set to 3 and 256, respectively. A Dense layer with 257 hidden units is added after TCM to generate the real mask $M_{16}^{TCN}$ for wideband spectrum signal.

**Table 1**. Hyper-parameters for fusion network.

| Layer | CNN | | | RNN | FC |
|---|---|---|---|---|---|
| | Channels | Size | Stride | Units | Units |
| Conv2D | 32 | [1,5] | [1,3] | | |
| Conv2D | 32 | [1,3] | [1,2] | | |
| Conv2D | 32 | [1,3] | [1,2] | | |
| GRU | | | | 256 | |
| Dense | | | | | 16 * 2 |

### 3.2. Wideband model fusion

The two wideband SE structures may have complementary strengths and weaknesses. For example, the enhanced output may have residual noise or speech distortion in different time-frequency regions for the two different SE structures. Instead of selecting one structure rather than another according to some criterion, inspired by the work of [19], we proposed a sub-band-based fusion module to fuse the output of TCN and CRN wideband SE.

As shown in Figure 2, the noisy wideband input $Y_{16}$, the wideband enhanced outputs $\hat{S}_{16}^{TCN}$ and $\hat{S}_{16}^{CRN}$ are concatenated in the channel domain to form the input of the fusion network. The fusion network has three conv2D layers, a GRU layer, and a Dense layer with softmax activation. The hyper-parameters are shown in Table 1. The sub-band fusion weights for the two SE systems are estimated by the fusion network. The number of sub-bands is set to 16. To avoid artifacts introduced by fast switching between the two systems, the weights are further smoothed along the time domain by exponential moving average method with decay set to 0.95. The enhanced signals are multiplied with the corresponding weights, and the weighted signals are summed together to generate the fusion output.

### 3.3. Wideband loss function

As shown in Figure 1, the wideband loss contains three loss functions to measure the estimation error from the signal, perceptual, and ASR aspects, respectively. For the wideband signal loss $\mathcal{L}_S$, the IAM-MALE loss [3] is used to suppress

the noise for the low-SNR time-frequency bins. Besides, the SISDR loss is also applied since it can achieve good general performance for SE [20].

$$\mathcal{L}_{\text{S}} = \mathcal{L}_{\text{IAM-MALE}} + \mathcal{L}_{\text{SISDR}} \quad (1)$$

The Phone-Fortified Perceptual Loss (PFPL) proposed in [14] is applied to take phonetic information into account for training the wideband SE. The $\mathcal{L}_{\text{PFPL}}$ is calculated by Wasserstein distance between the latent representations of wav2vec [15] model for clean and enhanced speech.

Moreover, an ASR-oriented loss $\mathcal{L}_{\text{ASR}}$ is used to reduce the speech distortion of enhanced speech and reduce the WER for ASR. More specifically, Wasserstein distance is calculated between the embeddings of the ASR encoder for clean and enhanced speech. A pre-trained ASR encoder with LibriSpeech dataset [21] by WeNet toolkit [16] is used to extract the embeddings. The overall wideband loss $\mathcal{L}_{\text{WB}}$ is calculated by,

$$\mathcal{L}_{\text{WB}} = \mathcal{L}_{\text{S}} + \mathcal{L}_{\text{PFPL}} + \mathcal{L}_{\text{ASR}} \quad (2)$$

### 3.4. Highband enhancement

As proposed in our work [6], the estimated wideband speech signal $\hat{S}_{16}$ is fed to the highband SE network alongside with $Y_{16-48}$. Using the estimated signal $\hat{S}_{16}$ to assist in producing $\hat{S}_{16-48}$ can significantly improve the highband SE quality. In the highband network, two CNN branches are utilized to extract the wideband and highband features from the estimated wideband speech and the noisy highband speech, respectively. The wideband and highband features are then combined and fed into a recurrent layer and a feed-forward layer to estimate the highband mask $M_{16-48}$. The details of the hyperparameters can be found in [6]. The IAM-MALE loss is utilized as the highband signal loss.

### 3.5. Post-processing

To further improve the subjective quality, envelope post-filtering [3] is employed to refine the amplitude mask estimated of wideband and highband SE models. Besides, to further suppress the highband residual noise, we adopt a simple but effective method to reduce the highband amplitude mask $M_{16-48}$ for the frames with a small average amplitude mask in frequency region between 4kHz and 8kHz.

$$M_{16-48}(t,f) = \begin{cases} 0.001, & \frac{\sum_f M_{8-16}(t,f)}{N_{8-16}} < 0.1 \\ M_{16-48}(t,f), & \frac{\sum_f M_{8-16}(t,f)}{N_{8-16}} >= 0.1 \end{cases} \quad (3)$$

where $M_{8-16}$ is the mask for frequency region between 4kHz and 8kHz. $N_{8-16}$ is the number of frequency bins of $M_{8-16}$.

## 4. PERSONALIZED SPEECH ENHANCEMENT

### 4.1. Speaker embedding extraction

To extract the target speaker's voice from a mixture of multiple speakers and noise, we need a reference recording of the target speaker. In this work, we adopt a speaker embedding extraction network to encode the reference speech into a vector that characterizes the target speaker. The speaker embedding extraction network consists of a 1D-Convolution layer, three GRU layers with 512 units, and three feed-forward layers. The 1D-Convolution layer has a kernel size of 1, a stride of 1, and a channel size of 512. The first feed-forward layer has 192 nodes followed by a mean pooling, where the speaker embedding vector is extracted from. The second and third feed-forward layers have 512 and 3236 nodes with ReLU and Softmax activation functions, respectively. The third feed-forward layer is the output layer of the speaker classification.

### 4.2. Target speech extraction

The target speech extraction network takes the mixture input and the encoded speaker embedding vector. It mimics a human's selective auditory attention that passes through only the target speaker's voice [22]. The target speaker extraction network firstly transforms the mixture input into a feature representation by a 1D-Convolution layer with a kernel size of 1, a stride of 1, and a channel size of 256. Then the feature representation is concatenated with the speaker embedding vector by repeating the vector along the temporal dimension of the feature. A linear layer is adopted to reduce the dimension of the concatenated features to 256. Then three GRU layers with 256 units are further employed to obtain the representations of the target speaker. The process of concatenation, linear transformation, and three GRU layers is repeated once again. Finally, a mask is estimated by a feed-forward layer with a Sigmoid activation function, which is used to filter the mixture input and remove the noise and interference speech.

### 4.3. Loss

Similar to the non-personalized speech enhancement, the overall wideband loss $\mathcal{L}_{\text{WB}}$ in eq. 2 is kept the same in personalized speech enhancement. Besides the wideband loss, a cross-entropy loss $\mathcal{L}_{\text{CE}}$ is used for the speaker classification to jointly optimize the speaker embedding network.

## 5. DATASETS, EXPERIMENTS, AND RESULTS

### 5.1. Datasets

The clean speech and noise datasets from the DNS Challenge (ICASSP 2022) [18] are used for data simulation. For the training dataset, we synthesize 1250 hours samples with the SNR levels of -5dB to 30dB. To improve the robustness of our system, the noisy and target signals are simultaneously convolved with the room impulse responses and various EQ filters, which are measured or simulated. To avoid speech distortion introduced by dereverberation, speech with 75ms early reflection is used as a training target.

### 5.2. Setup

The fullband SE system in Figure 1 is trained in multi-stage.

- Stage 1: The wideband TCN and CRN models are trained separately using wideband signal loss $\mathcal{L}_{\text{S}}$.
- Stage 2: The wideband TCN and CRN models are fine-tuned separately using the overall wideband loss $\mathcal{L}_{\text{WB}}$.
- Stage 3: The fusion network is trained based on the output of wideband models in stage2 using wideband loss $\mathcal{L}_{\text{WB}}$.

**Table 2**. Comparisons of wideband non-personalized SE systems.

| ID | System | Loss | PF | PSEQ-NB | PESQ-WB | WER |
|---|---|---|---|---|---|---|
| 0 | Clean | - | - | 4.5 | 4.64 | 2.04 |
| 1 | Noisy | - | - | 2.45 | 1.58 | 7.82 |
| 2 | DCCRN[1] | - | - | 3.27 | - | - |
| 3 | FullSubNet[23] | - | - | 3.31 | 2.78 | - |
| 4 | TCN | $\mathcal{L}_S$ | 0 | 3.41 | 2.92 | 9.03 |
| 5 | TCN | $\mathcal{L}_S + \mathcal{L}_{PFPL}$ | 0 | 3.45 | 2.98 | 8.43 |
| 6 | TCN | $\mathcal{L}_S + \mathcal{L}_{ASR}$ | 0 | 3.42 | 2.96 | 7.22 |
| 7 | TCN | $\mathcal{L}_{WB}$ | 0 | 3.44 | 2.98 | 7.39 |
| 8 | TCN-large | $\mathcal{L}_{WB}$ | 0 | 3.44 | 3.00 | 7.57 |
| 9 | CRN | $\mathcal{L}_{WB}$ | 0 | 3.40 | 2.91 | 8.55 |
| 10 | CRN-large | $\mathcal{L}_{WB}$ | 0 | 3.42 | 2.93 | 8.03 |
| 11 | TCN+RCN | $\mathcal{L}_{WB}$ | 0 | **3.47** | **3.04** | **7.22** |
| 12 | TCN+RCN | $\mathcal{L}_{WB}$ | 1 | 3.43 | 2.93 | 7.70 |

**Table 3**. Comparisons of wideband target speaker extraction with/without pre-enhancement for personalized SE.

| Condition | System | Pre-enhance | PSEQ-NB | PESQ-WB |
|---|---|---|---|---|
| Noisy 1-talker | Noisy | - | 1.97 | 1.16 |
| | Extraction | 0 | 2.52 | 1.62 |
| | Extraction | 1 | **2.87** | **2.01** |
| Noisy 2-talker | Noisy | - | 1.51 | 1.08 |
| | Extraction | 0 | 1.67 | 1.19 |
| | Extraction | 1 | **1.75** | **1.22** |

- Stage 4: The highband CRN model is trained based on the output of non-personalized wideband SE using highband signal loss.
- Stage 5: For personalized SE, the wideband speaker embedding and target speaker extraction networks are jointly optimized with the loss $\mathcal{L}_S + \mathcal{L}_{CE}$.
- Stage 6: The model in Stage 5 is further fine-tuned with the loss $\mathcal{L}_{WB} + \mathcal{L}_{CE}$.

All networks use ReLU activation and batch norm each layer except for the output layer. The Adam optimizer is employed for all networks. The learning rate is set to 0.0002 for Stage 2 and Stage 6, and 0.001 for all the other stages. Pre-processing method proposed in [3] is used to create a perceptually optimal magnitude spectrum as the training target. The parameters of STFT are the 1536 points (32 ms) FFT length with Hanning window and 384 points (8 ms) shift length.

### 5.3. Non-personalized speech enhancement on DNS2020

The wideband SE systems are evaluated on the synthetic test set from DNS Challenge (INTERSPEECH 2020) without reverberations. The 150 clips are labeled, and the pre-trained ASR system of WeNet is used to calculate the WER. As shown in Tabel 2, the TCN with wideband signal loss outperforms the baseline systems on PESQ-NB and PESQ-WB, indicating the effectiveness of the proposed wideband system. Combining the PFPL loss and ASR loss in system ID7 can further improve PESQ scores while reducing the WER, compared to the system ID4 with signal loss. Based on system ID7 and ID9, the proposed fusion method (system ID11) can further improve the performance. We also evaluate TCN and CRN (system ID8 and ID10) with the same complexity as system ID11. The results demonstrate that the performance improvement of fusion system is not introduced by higher complexity. With the post-filtering, although the subjective quality is improved, the PESQ of system ID12 is slightly decreased and the WER is increased.

**Table 4**. Performance of fullband non-personalized (Track 1) and personalized (Track 2) SE for DNS2022.

| Track | System | SIG-MOS | BAK-MOS | OVRL-MOS | WAcc | Score |
|---|---|---|---|---|---|---|
| 1 | Noisy | **4.29** | 2.15 | 2.63 | **0.72** | 0.56 |
| | Baseline | 3.62 | 3.93 | 3.26 | 0.63 | 0.60 |
| | Ours | 3.97 | **4.25** | **3.61** | 0.68 | **0.66** |
| 2 | Noisy | **4.25** | 2.14 | 2.56 | **0.72** | 0.55 |
| | Baseline | 3.64 | 4.24 | 3.4 | 0.64 | 0.62 |
| | Ours | 3.88 | **4.32** | **3.63** | 0.68 | **0.67** |

### 5.4. Personalized speech enhancement on Libri2Mix

We evaluate the effectiveness of target speaker extraction for personalized SE on the synthetic wideband test set of Libri2Mix [24]. From Table 3, we observe that the target speaker extraction could significantly improve the perceptual quality of the noisy speech from one or two simultaneous speakers. With an enhancement module as a pre-processing, the performance of the cascaded target speaker extraction system could be further improved.

### 5.5. Track1 and Track2 Results on DNS2022

From Table 4, we observe that the proposed multi-stage and multi-loss framework significantly improves the perceptual quality and outperforms the baseline systems for both non-personalized and personalized SE.

### 5.6. Delay and Complexity

With 1536-point FFT (32 ms) and 384-point stride (8 ms), the total system delay is 32ms + 8ms = 40ms, which satisfies the latency requirement of this challenge. We evaluate the one-frame inference time of proposed systems using a privately modified version of TFLite 2.3 on Intel Core i5 (2.4 GHz) CPU (single-threaded). For non-personalized SE system, the proposed model has 12.41M parameters, the one-frame inference time is 1.673 ms. For the personalized SE system, the proposed model has 15.14M parameters, the one-frame inference time is 3.455 ms.

## 6. CONCLUSIONS

We present a multi-stage multi-loss training framework for non-personalized and personalized full-band speech enhancement. We extend the existing wideband (16k) system to enable fullband (48k) SE. Our approach significantly outperforms the baseline and other systems and improves the perceptual signal quality in real-time.

## 7. REFERENCES


[1] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. of Interspeech*, 2020.

[2] A. Li, W. Liu, X. Luo, G. Yu, C. Zheng, and X. Li, "A Simultaneous Denoising and Dereverberation Framework with Target Decoupling," in *Proc. of Interspeech*, 2021, pp. 2801–2805.

[3] X. Zhang, X Ren, X. Zheng, L. Chen, C. Zhang, L. Guo, and B. Yu, "Low-Delay Speech Enhancement Using Perceptually Motivated Target and Loss," in *Proc. of Interspeech*, 2021, pp. 2826–2830.

[4] J. Valinc, "A hybrid dsp/deep learning approach to real-time full-band speech enhancement," in *Proc. of MMSP*, 2018, pp. 1–5.

[5] J.-M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy, "A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech," in *Interspeech*, 2020, pp. 2482–2486.

[6] X. Zhang, L. Chen, X. Zheng, X. Ren, C. Zhang, L. Guo, and B. Yu, "A two-step backward compatible fullband speech enhancement system," in *Proc. of ICASSP 2022*, In Press, https://arxiv.org/abs/2201.10809.

[7] A. Pandey and D. Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. of ICASSP*. IEEE, 2019, pp. 6875–6879.

[8] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement.," in *Proc. of Interspeech*, 2018, pp. 3229–3233.

[9] M. Kim, W. Choi, J. Chung, D. Lee, and S. Jung, "Kuielab-mdx-net: A two-stream neural network for music demixing," in *Proc. of MDX Workshop*, 2021.

[10] S. Seo and J. Kim, "Mobilenet using coordinate attention and fusions for low-complexity acoustic scene classification with multiple devices," Tech. Rep., DCASE2021 Challenge, June 2021.

[11] J. Yu, B. Wu, R. Gu, S.-X. Zhang, L. Chen, Y. X. Yu, D. Su, D. Yu, X. Liu, H. Meng, et al., "Audio-visual multi-channel recognition of overlapped speech," in *Proc. of Interspeech*, 2020.

[12] J. Wu, Z. Chen, S. Chen, Y. Wu, T. Yoshioka, N. Kanda, S. Liu, and J. Li, "Investigation of practical aspects of single channel speech separation for asr," in *Proc. of Interspeech*, 2021.

[13] S. E. Eskimez, X. Wang, M. Tang, H. Yang, Z. Zhu, Z. Chen, H. Wang, and T. Yoshioka, "Human listening and live captioning: Multi-task training for speech enhancement," in *Proc. of Interspeech*, 2021.

[14] T.-A. Hsieh, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Improving perceptual quality by phone-fortified perceptual loss using wasserstein distance for speech enhancement," in *Proc. of Interspeech*, 2020.

[15] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. of Interspeech*, 2019.

[16] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, "WeNet: Production Oriented Streaming and Non-Streaming End-to-End Speech Recognition Toolkit," in *Proc. of Interspeech*, 2021, pp. 4054–4058.

[17] ITU, "ITU-R Rec. P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," *ITU Standard*, 2007.

[18] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusevych, S. Braun, S. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "Icassp 2022 deep noise suppression challenge," in *Proc. of ICASSP*. IEEE, 2022.

[19] X. Jaureguiberry, E. Vincent, and G. Richard, "Fusion methods for speech enhancement and audio source separation," *IEEE/ACM TASLP*, vol. 24, no. 7, pp. 1266–1279, 2016.

[20] Morten Kolbæk, Zheng-Hua Tan, Søren Holdt Jensen, and Jesper Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *IEEE/ACM TASLP*, vol. 28, pp. 825–838, 2020.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. of ICASSP*. IEEE, 2015, pp. 5206–5210.

[22] C. Xu, W. Rao, E. S. Chng, and H. Li, "Spex: Multi-scale time domain speaker extraction network," *IEEE/ACM TASLP*, vol. 28, pp. 1370–1384, 2020.

[23] 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 *Proc. of ICASSP*. IEEE, 2021, pp. 6633–6637.

[24] 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.