



Spectro-Temporal SubNet for Real-Time Monaural Speech Denoising and Dereverberation

Feifei Xiong¹ Weiguang Chen^{1*} Pengyu Wang^{1*} Xiaofei Li² Jinwei Feng¹

¹Hummingbird Audio Lab, Alibaba Group, Hangzhou, China

²Westlake University & Westlake Institute for Advanced Study, Hangzhou, China

{xiff242920, jinwei.feng}@alibaba-inc.com, lixiaofei@westlake.edu.cn

Abstract

This paper presents an improved subband neural network applied to joint speech denoising and dereverberation for online single-channel scenarios. Preserving the advantages of subband model (SubNet) that processes each frequency band independently and requires small amount of resources for good generalization, the proposed framework named STSubNet exploits sufficient spectro-temporal receptive fields (STRFs) from speech spectrum via a two-dimensional convolution network cooperating with a bi-directional long short-term memory network across frequency bands, to further improve the neural network discrimination between desired speech component and undesired interference including noise and reverberation. The importance of this STRF extractor is analyzed by evaluating the contribution of individual module to the STSubNet performance for simultaneously denoising and dereverberation. Experimental results show that STSubNet outperforms other subband variants and achieves competitive performance compared to state-of-the-art models on two publicly benchmark test sets.

Index Terms: Subband network, spectro-temporal receptive field, speech denoising, dereverberation, real-time process

1. Introduction

Monaural speech enhancement aims at restoring the desired speech signal from its distorted version to improve perceptual speech quality and intelligibility, which is desirable for hands-free speech communications and hearing assistant systems when only a single distant microphone is available. Although significant progress has been achieved in the last four decades [1], especially in recent years with deep learning based approaches (see [2]), this active research field still remains challenging for real-world applications due to additive noise, room reverberation, computational efficiency, algorithmic latency, and model size restriction deployed on devices.

Research has been conducted recently to develop advanced technique to address either one or several of the above issues. For instance, two international challenges - the REVERB challenge [3] and the DNS challenge [4] - have respectively fostered state-of-the-art approaches of speech dereverberation and denoising. Despite the fact that there exists no optimal solution for single-channel dereverberation (as a process of inverse filtering of the room impulse response RIR) [5], spectral domain methods focusing on late reverberation suppression [6, 7], and deep learning based approaches nonlinearly mapping the reverberant to the clean spectrum [8, 9, 10] have been proposed to mitigate the reverberation effect. Speech denoising is currently often accomplished via supervised deep learning in the time-frequency (TF) domain to map noisy speech spectral fea-

ture onto clean speech target (see [11]), and the novelty ranges from *complete* time-domain method [12], over various learning targets [13], to different neural network architectures in terms of loss function [14, 15], convolutional or recurrent or generative adversarial network [16, 17, 18], multi-stage learning [19] or subband network [20]. Due to different properties of noise (addition) and reverberation (convolution), some deep learning based work [19, 21] showed that it is not easy to achieve simultaneous speech denoising and dereverberation in a single stage paradigm. Instead, the authors proposed a two stage network for joint training[22], or two pairs of learning targets are required to stabilize training[23]. Additionally, for real-time applications, online processing with tolerable computational complexity and latency (see [4]) is necessary, which meanwhile demands that the deep learning based approaches pair the inference model as lightweight as possible.

Subband long short-term memory (LSTM) network (SubNet) was proposed for online single-channel speech denoising in our previous work [20], in which SubNet requires a small amount of training resources and can be designed with few parameters without loss of generalization since it focuses on the subband information and shares the same network parameter across frequency bands. The underlying principle is to learn the frequency-wise signal characteristics from the local spectral pattern that has been proved to be informative for discriminating between speech and other signals [24]. To improve SubNet to be capable of modeling the global spectral pattern which benefits the cases when the subband has an extremely low signal-to-noise ratio (SNR), FullSubNet [25] has been proposed to fuse fullband and subband network, but at the cost of the increased network size and computational complexity.

Essentially, in an attempt to further improve SubNet, more discriminative input in each individual subband (frequency channel) is required for discriminating between desired speech and undesired interference. Motivated by the auditory findings that neurons in the primary auditory cortex of different mammal species are sensitive to specific spectro-temporal receptive fields (STRFs, e.g. the neurons exhibit high firing rates only when a specific modulation frequency is represented in the stimulus) [26], we propose an improved SubNet in this paper, namely spectro-temporal SubNet (STSubNet), which consists of a STRF extractor prior to SubNet. More specifically, both spectral and temporal context from speech spectrum are extracted via a two-dimensional (2D) convolution network for each individual subband, as the input to SubNet. Such 2D convolution layer is jointly trained with SubNet to obtain optimized parameters to extract the corresponding STRF information. However, different subband frequency channel might need individual 2D convolution parameters due to different sensitivities across frequency bands. In order to extend STRFs across

*Authors are intern students at Alibaba Group during this work.

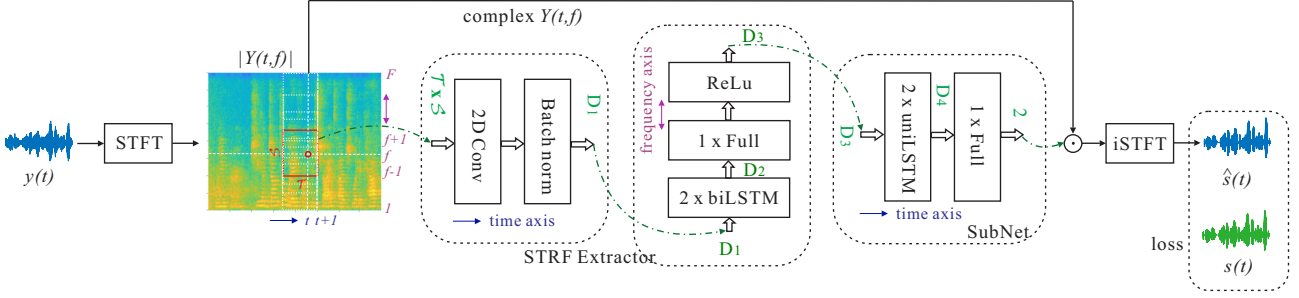


Figure 1: Diagram of the proposed STSubNet. The 2D observation window of speech spectrum is plotted with size $\mathcal{T} \times \mathcal{S}$.

spectral domain as well as to avoid using too many 2D convolution layers (one for each subband), we utilize a bi-directional LSTM along the frequency bands inspired by the proposal of convolutional and recurrent layer combination in [27].

Furthermore, STSubNet is designed to perform both speech denoising and dereverberation simultaneously within one stage training and one pair of learning target. We expect that subband network is particularly beneficial to dereverberation because the reverberation effect (represented by the RIR) is physically frequency-dependent due to the absorption coefficients vary with frequency, and subband network generalization seen by same amount of RIRs during training will be better than other fullband models that consider RIRs as fullband data. Also, like SubNet, the proposed STSubNet still meets the real-time requirement, and a lightweight version is possible by shrinking the network parameters while keeping the competitive performance.

2. Method

Let $s(t)$, $h(t)$, $n(t)$ denote anechoic speech, the RIR and additive noise, respectively. The received speech $y(t)$ with noise and reverberation can be represented as

$$y(t) = x(t) + n(t) = s(t) * h(t) + n(t), \quad (1)$$

where $x(t)$ is the reverberant speech with time index t and $*$ is the convolution operator. After the short-time Fourier transform (STFT) with frequency band f ,

$$Y(t, f) = X(t, f) + N(t, f). \quad (2)$$

The objective of speech denoising and dereverberation is to restore the anechoic speech from $y(t)$, and it is common to apply mask in STFT domain to achieve as $\hat{s}(t)$ after inverse STFT

$$\hat{S}(t, f) = Y(t, f) \cdot M(t, f). \quad (3)$$

It has been shown in [13] that a precise estimation of phase can provide more hearing perception quality improvement. Hence, we choose the real part $M_r(t, f)$ and the imaginary part $M_i(t, f)$ of uncompressed complex Ideal Ratio Mask (cIRM) $M(t, f) = M_r(t, f) + jM_i(t, f)$ as the learning target of the proposed STSubNet.

2.1. STSubNet

Figure 1 shows the diagram of the proposed STSubNet. First, spectral and temporal context from speech spectrum at time slot t and frequency band f ($f \in F$ and F denotes half of the fast Fourier transform (FFT) length plus one) are extracted via a 2D observation window with temporal frames \mathcal{T} (look past) and spectral spans \mathcal{S} (look up and down symmetrically, duplicated the last context if spans are insufficient). These 2D features are

fed to a 2D convolutional network with output dimension of D_1 , followed by a batch normalization. Then a bi-directional LSTM module is appended in order to extend STRFs for each individual frequency band f , which consists of two bi-directional LSTM layers with hidden size D_2 , one linear fully connected layer and rectified linear unit (ReLU) as the output layer's activation function with unit size D_3 . Note that bi-directional LSTM used in STRF extractor is still suitable for online application due to its forward/backward direction only along the frequency axis. For the SubNet part, we use two stacked unidirectional LSTM layers (fitting to real-time process) with hidden size D_4 and one linear fully connected layer to yield two values $M_r(t, f)$ and $M_i(t, f)$.

It is worth noting that the number of STSubNet parameters mainly depends on the settings of (D_1, D_2, D_3, D_4) , and a lightweight version (see Section 3.1) is possible thanks to the SubNet advantage that training procedure converges normally in a few epochs, and overfitting rarely happens [20].

2.2. Loss Function

Instead of calculating the loss function directly in TF domain when cIRM is obtained, we use the signal-to-distortion ratio (SDR) loss as the criterion for maximizing to train the model with Adam optimizer, computed as

$$\mathcal{L} = 10 * \log_{10} \frac{\|s(t)\|^2}{\|s(t) - \hat{s}(t)\|^2}. \quad (4)$$

Our ablation experimental results showed that this time-domain loss performs more effective compared to scale-invariant SDR (SI-SDR) loss [14] or mean square error (MSE) criterion of complex spectrum as applied in [20].

2.3. Online Inference

To facilitate the network training, the input sequence has to be normalized to obtain an equalized input level, which is usually accomplished via dividing the mean value $\bar{\mu}$ of the magnitude spectral features in a batch manner as $|Y(t, f)|/\bar{\mu}$ [25]. However, such normalization does not fit to online inference because the signal is received and processed frame by frame. In order to keep the normalization strategy consistent during training and inference, we adopt the cumulative normalization method [28] across the fullband magnitude, that is, at each time, the mean value used for normalization is computed using the available frames within a sliding window. This can be also implemented as a recursively computed mean value $\mu(t)$ [20] as

$$\mu(t) = \alpha\mu(t-1) + (1-\alpha) \left(\frac{\sum_{f=1}^F |Y(t, f)|}{F} \right), \quad (5)$$

where $\alpha = \frac{L-1}{L+1}$ is the smoothing factor and L represents the sliding window length.

3. Experimental Setup

3.1. Training Dataset

To generate the training data according to (1), we used the speech and noise data from the DNS challenge [4], including clean speech data 500 hours of clips from 2150 speakers, and the noise dataset over 180 hours of clips from 150 classes. Without the use of real-recorded RIRs provided by open resources, we only used 12828 synthetic RIRs generated using image method [29] with the reverberation time ranging from 0.2 to 2.0 s in the room space from 90 to 450 m³. Noise signals were added with the segmental SNR levels sampled from a uniform distribution between -5 and 20 dB. The mixed signal was then set to target root mean square (RMS) level sampled from a uniform distribution between -15 and -35 dBFS.

All the utterances were sampled at 16 kHz. The 20 ms Hanning window was utilized in STFT, with 50% overlap between adjacent frames. FFT length was set to 512 to obtain the complex spectrum, resulting in $F = 257$ (see Section 2.1). The input-target sequence pairs are usually generated with a constant-length sequence. We set this length to 4 s, which was also applied to $L = 400$ for online inference in (5). We adopted the real-time requirement rule from the DNS challenge [4] - the total algorithmic latency including frame size, stride time and look ahead should not exceed 40 ms - which allows for one frame (10 ms) look ahead in our experiments. The initialized learning rate was set to 1e-3, and we halved such rate when validation loss did not increase until the final learning rate 6.25e-5. The batch size was set to 32 at the utterance level.

The STSubNet parameter settings (see Section 2.1) were $D_1, D_2, D_3, D_4 = 16, 64, 32, 128$, resulting in the number of network parameters 0.36 million. Additionally, we summarized other parameter settings with larger network size in Table 1, and these models still met the requirement for online applications which was measured using real-time factor (RTF). It is calculated as the computation time of processing one frame divided by the duration of one STFT frame (10 ms here) and needs to be smaller than one. We tested on Intel Xeon CPU E5-2682 v4 (2.50 GHz) with PyTorch implementation of the proposed STSubNet.

Table 1: STSubNet with different network parameter settings.

| Model | D_1, D_2, D_3, D_4 | #Para(M) | RTF |
|-----------------------|----------------------|----------|-------|
| STSubNet | 16, 64, 32, 128 | 0.36 | 0.537 |
| STSubNet ₂ | 32, 128, 64, 256 | 1.44 | 0.576 |
| STSubNet ₃ | 64, 256, 128, 512 | 5.66 | 0.705 |

3.2. Simulated Test Dataset

For speech denoising, two untrained noises were employed to demonstrate the model generalization capability, namely babble, and factory1 from NOISEX92 [30], with three segmental SNRs of $[-6, 0, 6]$ dB, denoted as test set '*bnm6, bn0, bn6, fnm6, fn0, fn6*'. For speech dereverberation, the RIRs from REVERB challenge [3] simulated test set were employed, representing six different acoustic conditions, meaning three different room sizes (small1, median2, large3) with two different speaker-to-microphone distances (near with 0.5 m, far with 2 m), denoted as test set '*r1n, r1f, r2n, r2f, r3n, r3f*'. The anechoic speech came from the non-blind DNS development set [4] which contains 150 utterances.

3.3. Public Test Dataset

We further tested the proposed model effectiveness by comparing with other state-of-the-art methods under the same publicly

available test datasets. For speech denoising, the DNS challenge non-blind test set [4] was used for speech denoising, which has 150 noisy clips with SNR levels distributed randomly between 0 and 20 dB. The real recordings from the evaluation test set of the REVERB challenge [3] were used for speech dereverberation task, which contains '*near*' (187 utterances) and '*far*' (186 utterances) speaker-to-microphone distances with ambient room noises from the MC-WSJ-AV corpus [31].

4. Results

Wide-band perceptual evaluation of speech quality score (PESQ) [32], short-time objective intelligibility score (STOI) [33] and SI-SDR [34] were used as evaluation metrics for speech denoising, while for speech dereverberation, SI-SDR was replaced by normalized speech-to-reverberation modulation energy ratio (SRMR) [35].

4.1. Effect of STRF Extractor

In order to analyze the effect of the proposed STRF extractor, composed of 2D convolution layer with $(\mathcal{T}, \mathcal{S})$ input, and the bilSTM module along frequency axis (see Figure 1), systematic experiments were conducted on the simulated test dataset.

As illustrated in Figure 2, in general, STSubNet trained in a single stage is capable of providing consistent improvement in both denoising and dereverberation tasks in terms of all the evaluation metrics. For instance, performance increased by 0.1 and 0.12 absolutely on average in terms of STOI for the more challenging scenarios with SNR of -6 dB and *far* position, respectively. It is also interesting to observe that the gain achieved by STSubNet for dereverberation on *near* test sets is not as large as the case of *far*, particularly in terms of SRMR. Actually, *near* test sets usually have higher early-to-late reverberation ratio than *far* test sets and early reflections are always beneficial to speech perception (see [7]). This could give us a hint that it still remains a big challenge for model-based approaches to achieve a *complete* dereverberation, and future research can be focused on modeling early reflections which own more correlation to the anechoic signal itself than late reverberation.

With an increased temporal context \mathcal{T} (e.g., 15 means 13 frames look past, one current frame and one frame look ahead),

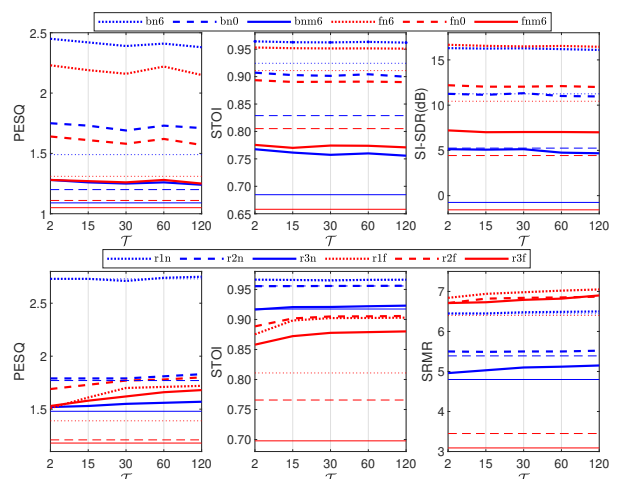


Figure 2: STSubNet Performance on 12 simulated test sets in terms of different \mathcal{T} with fixed $\mathcal{S} = 1$. Thin lines (with the same line type to thick lines) denote the corresponding metrics with unprocessed noisy or reverberant signals.

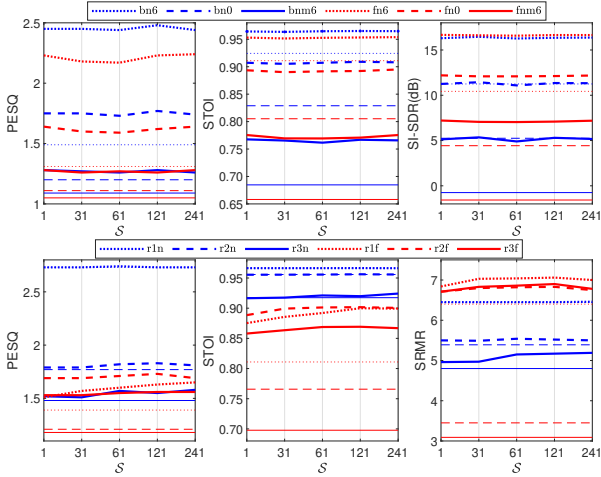


Figure 3: *STSubNet* Performance on 12 simulated test sets in terms of different S with fixed $\mathcal{T} = 15$.

performance for denoising task slightly degrades, while performance for dereverberation improves, particularly for *far* positions (*r1f*, *r2f*, *r3f*) represented by red lines) with severe reverberation effect. This indicates that a long temporal context seems to be more important for *STSubNet* to better eliminate reverberation effect, probably due to the strong correlation of reverberant signals across frames. A good compromised \mathcal{T} to balance these two different tasks (denoising and dereverberation) can be set around 15, i.e., 150 ms.

On the other hand, as seen in Figure 3, the effect of spectral context S seems to affect more the performance of test sets '*r1f*, *r2f*, *r3n*, *r3f*' with severe reverberation effect, showing that there still exists room for improvement of dereverberation via a large spectral receptive field, even if biLSTM is applied to extend the spectral patterns. However, a larger S does not guarantee better performance, indicating that large S cooperating with biLSTM would cause redundant spectral information extraction to hinder the model generalization improvement. $S = 31$ (spanning around 1000 Hz) seems to be a good choice.

Furthermore, the contribution of individual \mathcal{T} , S and biLSTM is summarized in Table 2 via a *leave-one-out* procedure to compare to the baseline ($\mathcal{T} = 15$, $S = 31$, biLSTM), i.e., temporal context degraded to $\mathcal{T} = 2$, spectral context degraded to $S = 1$, or without biLSTM. The change (denoted as Δ) of the averaged metric on simulated test sets thus reflects the significance of each individual module, i.e. more negative value in Table 2 represents more significance of the corresponding module. Results show that biLSTM plays an essential role in *STSubNet* improvement for both denoising and dereverberation, and temporal context is more important for dereverberation.

Table 2: Contribution of individual \mathcal{T} , S and biLSTM to *STSubNet* performance on simulated test sets for two tasks (denoising and dereverberation, separated by /).

| Denoising / Dereverberation | Δ PESQ | Δ STOI | Δ SI-SDR Δ SRMR |
|---|--------------------------------|------------------------------------|----------------------------------|
| $\mathcal{T} = 2$ (not 15), $S = 31$, biLSTM | +0.03 / -0.04 | +0.0036 / -0.0091 | +0.12 / -0.05 |
| $S = 1$ (not 31), $\mathcal{T} = 15$, biLSTM | +0.01 / -0.02 | +0.0017 / -0.0046 | -0.04 / -0.01 |
| $\mathcal{T} = 15$, $S = 31$, (without biLSTM) | -0.21 / -0.11 | -0.0352 / -0.0235 | -1.26 / -0.02 |

4.2. Performance Comparison with State-of-the-Art

As listed in Table 3 for real-time speech denoising, *STSubNet* consistently surpasses other types of SubNet in terms of all evaluation metrics, showing that the proposed STRF extractor can efficiently provide SubNet more discriminative input for each frequency band with fewer network parameters. In comparison to the competitive lightweight TRUNet [23], *STSubNet* obtains better performance in terms of STOI and SI-SDR with only slightly worse PESQ. Further, performance of *STSubNet* consistently improves as network parameters (see Table 1) increase. In comparison to state-of-the-art GaGNet [36], *STSubNet*₃ achieves comparable STOI, and even better SI-SDR.

Table 3: Comparison with other state-of-the-art systems on the DNS challenge non-blind test set.

| Method | #Para(M) | PESQ | STOI | SI-SDR(dB) |
|------------------------------|-------------|-------------|---------------|--------------|
| Noisy | - | 1.58 | 0.9152 | 9.07 |
| NSNet [15] | 5.1 | 2.15 | 0.9447 | 15.61 |
| DTLN [17] | 0.99 | - | 0.9476 | 16.34 |
| TRUNet [23] | 0.38 | 2.86 | 0.9632 | 17.55 |
| CTSNNet [19] | 4.99 | 2.94 | 0.9666 | 17.99 |
| GaGNet [36] | 5.94 | 3.17 | 0.9713 | 18.91 |
| SubNet [20] | 1.3 | 2.37 | 0.9424 | 16.15 |
| Fullband [25] | 6.0 | 2.73 | 0.9571 | 16.19 |
| FullSubnet [25] | 5.6 | 2.78 | 0.9611 | 17.29 |
| <i>STSubNet</i> | 0.36 | 2.84 | 0.9645 | 18.59 |
| <i>STSubNet</i> ₂ | 1.44 | 2.91 | 0.9680 | 19.12 |
| <i>STSubNet</i> ₃ | 5.66 | 3.00 | 0.9703 | 19.64 |

With the same *STSubNet* model, performance of speech dereverberation is compared to other state-of-the-art dereverberation approaches. As summarized in Table 4, *STSubNet* further increases SRMR by nearly 1.0 on average when compared to the best model-based method [10] which was specifically designed for dereverberation using the same test set (Section 3.3). Competitive performance is even achieved by *STSubNet*₃ when compared to the standard weighted prediction error (WPE) dereverberation algorithm [6] but with two channels recording. Note that a subjective evaluation was not carried out yet (for future work). The interested reader can check a few examples in <https://github.com/ffxiang/stsubnet>.

Table 4: Comparison with other state-of-the-art systems on the REVERB challenge real recorded evaluation test set (no reference signals) in terms of SRMR.

| Method | <i>near</i> | <i>far</i> |
|------------------------------|-------------|-------------|
| Reverberant | 3.17 | 3.19 |
| NTU-ADSC [8] | 4.29 | 4.42 |
| U-Net [10] | 5.47 | 5.68 |
| WPE(2ch) [6] | 6.55 | 6.71 |
| <i>STSubNet</i> | 6.36 | 6.37 |
| <i>STSubNet</i> ₂ | 6.48 | 6.49 |
| <i>STSubNet</i> ₃ | 6.54 | 6.53 |

5. Conclusions

This paper proposed *STSubNet*, a novel subband network incorporating an efficient spectro-temporal receptive field extractor to achieve simultaneous denoising and dereverberation in a single stage for real-time applications. Experimental results showed that *STSubNet* outperforms other subband network variants and achieves both comparable performance with individual state-of-the-art denoising and dereverberation approaches on two public test datasets. Further, *STSubNet* can be designed in a lightweight manner using fewer network parameters, and meanwhile maintains competitive performance.

6. References

- [1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.
- [2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.
- [3] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, no. 7, 2016.
- [4] C. K. A. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matuselych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Interspeech*, 2020, pp. 2492–2496.
- [5] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 2, pp. 145–152, 1988.
- [6] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, and A. Nakamura, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge," in *REVERB Challenge Workshop*, 2014.
- [7] F. Xiong, B. T. Meyer, N. Moritz, R. Rehr, J. Anemüller, T. Gerkmann, S. Doclo, and S. Goetze, "Front-end technologies for robust ASR in reverberant environments - spectral enhancement-based dereverberation and auditory modulation filterbank features," *EURASIP Journal on Advances in Signal Processing*, no. 70, 2015.
- [8] X. Xiao, S. Zhao, D. Hoang, H. Nguyen, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "The NTU-ADSC systems for reverberation challenge 2014," in *REVERB Challenge Workshop*, 2014.
- [9] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.
- [10] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," in *European Signal Processing Conference*, 2018, pp. 390–394.
- [11] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.
- [12] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *ICASSP*, 2018, pp. 5069–5073.
- [13] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.
- [14] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 1, pp. 825–838, 2020.
- [15] Y. Xia, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *ICASSP*, 2020, pp. 871–875.
- [16] 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 *Interspeech*, 2020, pp. 2472–2476.
- [17] N. L. Westhausen and B. T. Meyer, "Dual-signal transformation LSTM network for real-time noise suppression," in *Interspeech*, 2020, pp. 2477–2481.
- [18] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An improved version of MetricGAN for speech enhancement," in *Interspeech*, 2021, pp. 201–205.
- [19] A. Li, W. Liu, C. Zheng, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, no. 5, pp. 1829–1843, 2021.
- [20] X. Li and R. Horaud, "Online monaural speech enhancement using delayed subband LSTM," in *Interspeech*, 2020, pp. 2462–2466.
- [21] J. Li, D. Luo, Y. Liu, Y. Zhu, Z. Li, G. Cui, W. Tang, and W. Chen, "Densely connected multi-stage model with channel wise subband feature for real-time speech enhancement," in *ICASSP*, 2021, pp. 6638–6642.
- [22] C. Fan, J. Tao, B. Liu, J. Yi, and Z. Wen, "Joint training for simultaneous speech denoising and dereverberation with deep embedding representations," in *Interspeech*, 2020, pp. 4536–4540.
- [23] H.-S. Choi, S. Park, J. H. Lee, H. Heo, D. Jeon, and K. Lee, "Real-time denoising and dereverberation with tiny recurrent U-Net," in *ICASSP*, 2021, pp. 5789–5793.
- [24] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 270–279, 2013.
- [25] 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, pp. 6633–6637.
- [26] A. Qui, C. Schreiner, and M. Escabi, "Gabor analysis of auditory midbrain receptive fields: Spectro-temporal and binaural composition," *J. Neurophysiol.*, vol. 90, no. 1, pp. 456–476, 2003.
- [27] T. Grzywalski and S. Drgas, "Using recurrences in time and frequency within U-net architecture for speech enhancement," in *ICASSP*, 2019, pp. 6970–6974.
- [28] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [29] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. of the Acoustical Society of America*, vol. 64, no. 4, p. 943, 1979.
- [30] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.
- [31] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "Multi-channel wall street journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," in *ASRU*, 2005, pp. 357–362.
- [32] Rec. ITU-T P. 862.2, *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, Std., 2001.
- [33] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.
- [34] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *ICASSP*, 2019, pp. 626–630.
- [35] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An updated objective intelligibility estimation metric for normal hearing listeners under noise and reverberation," in *IWAENC*, 2014.
- [36] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," in *arXiv:2106.11789*, 2021.