ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096020

COMPUTATIONAL EFFICIENT MONAURAL SPEECH ENHANCEMENT WITH UNIVERSAL SAMPLE RATE BAND-SPLIT RNN

Jianwei Yu, Yi Luo

Tencent AI Lab, Shenzhen, China

{tomasyu, oulyluo}@tencent.com

ABSTRACT

While recent developments on the design of neural networks have greatly advanced the state-of-the-art of speech enhancement and separation systems, practical applications of such networks often put extra constraints on their model size and computational complexity. Moreover, as different telecommunication services may have different transmission bandwidths which result in different signal sample rates, one model is typically designed for a particular sample rate. In this paper, we extend the usage of a recently proposed frequency-domain source separation model, the band-split RNN (BSRNN), to the task of universal-sample-rate resource efficient speech enhancement. BSRNN explicitly splits the spectrogram into different frequency bands and perform interleaved band-level and sequence-level modeling, and the bandwidths can be manually designed to balance the model size, computational cost, and performance. By properly designing the band-splitting scheme and the hyperparameters, a single BSRNN model can handle signals at a wide range of sample rates, and the computational cost required to process a lower-sample-rate signal can be smaller than that of a higher-sample-rate signal. Experiment results show that compared to various benchmark systems in speech enhancement and separation, our universal-sample-rate BSRNN (USR-BSRNN) achieves comparable or better signal-to-noise ratio (SNR) performance at a same level of model size or computational cost.

Index Terms— Speech enhancement, Universal sample rate, Dynamic complexity

1. INTRODUCTION

Speech enhancement is a speech processing technique that aims to improve the quality and intelligibility of noisy speech. Over the past few years, the performance of deep neural network (DNN) based speech enhancement models has been considerably advanced due to the development of model architecture designs [1–12] and training pipelines [13,14]. To meet the model size and computation complexity constrains of practical applicants, a series of lightweight speech enhancement frameworks [15–18] have also been proposed recently. Moreover, with the increasing demand for high-quality Hi-Fi speech in online conference systems and other real-time speech communication scenarios, speech enhancement models have also been expanded to deal with super wide-band or full-band speech signals [19–21]. However, for telecommunication services with different transmission bandwidths and signal sample rates, most of the current speech enhancement models are typically designed for a particular sample rate. The potential of designing a versatile speech enhancement model that can handle signals with a wide range of sample rates is still need to be explored.

Ideally, an efficient universal-sample-rate model should have two properties. First, the model should have consistent performance at various sample rates. Second, since there is no need to model the high-frequency components for low-sample-rate inputs, the computational cost of the model required to process a lower-sample-rate signal should be smaller than that of a higher-sample-rate signal. However, the performance of most existing speech enhancement models trained on a particular sample rate usually can not be generalized to other (unseen) sample rates [22], since the training data implicitly specify the sample rate. Existing solutions [23–25] to tackle this problem includes *multi-sample-rate training* [24] and *sampling-frequency-independent (SFI) convolutional layer* [25], where the former one sequentially trained multiple models operating at different sample rates until the final sample rate is reached, and the latter one applied an analog-to-digital transform to sample discrete convolutional kernels from a continuous signal and allowed the analog filter to be jointly optimized with the entire system. However, the sequentially trained model does not satisfy the requirement that a single model can cover all sample rates, and the performance drops when dealing with an unseen sample rate [22]. The SFI method is particularly designed for learnable signal encoder and decoder, and the frequency responses of the convolutional kernels cannot exceed the maximum sample rate in the training samples [22]; moreover, the performance at lower sample rates can be lower than that at higher sample rates, and the computational complexity at all sample rates are the same.

In this paper, we attempt to address the aforementioned problem and satisfy the requirements by extending our recently work of a frequency-domain source separation model, the *band-split RNN (BSRNN)* [5], to the task of universal-sample-rate resource efficient speech enhancement. For each sample rate, the proposed USR-BSRNN model first upsamples the input signal to a pre-defined sample rate (e.g., 48k Hz) and then transforms the upsampled signal to complex-valued spectrogram via short-time Fourier transform (STFT). Then the model explicitly splits the spectrogram into a set of frequency bands and performs processing on the valid bands corresponding to the input sample rate. The selected subband spectrograms are then transformed to a series of subband features with band specify fully connection layers, and two residual recurrent neural network (RNN) layers are utilized to perform interleaved band-level and sequence-level modeling. A complex-valued time-frequency (T-F) mask is then calculated by transforming the output of the last RNN layer with band-specific multilayer perceptrons (MLPs). The two requirements mentioned above can be realized by two approaches: we train the model with input signals at various sample rates, and the band-level RNN only takes valid frequency bands as inputs to be processed. By properly setting the band-splitting scheme, the hyperparameters, and modifying the sequence-level

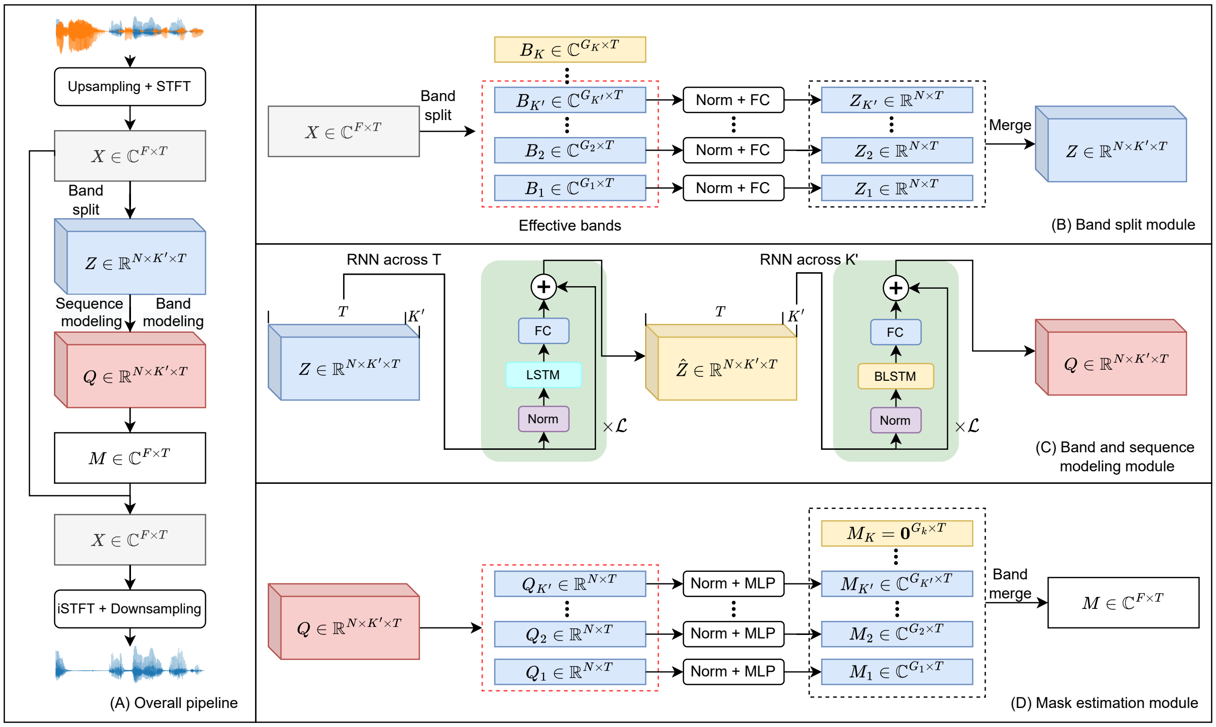

Upsampling + STFT
$X \in \mathbb{C}^{F\times T}$
Band split
$Z \in \mathbb{R}^{N\times K'\times T}$
Sequence modeling
Band modeling
$Q \in \mathbb{R}^{N\times K'\times T}$
$M \in \mathbb{C}^{F\times T}$
$X \in \mathbb{C}^{F\times T}$
iSTFT + Downsampling
(A) Overall pipeline
$B_K \in \mathbb{C}^{G_K\times T}$
$X \in \mathbb{C}^{F\times T}$
Band split
$B_{K'} \in \mathbb{C}^{G_{K'}\times T}$
$B_2 \in \mathbb{C}^{G_2\times T}$
$B_1 \in \mathbb{C}^{G_1\times T}$
Norm + FC
Norm + FC
Norm + FC
$Z_{K'} \in \mathbb{R}^{N\times T}$
$Z_2 \in \mathbb{R}^{N\times T}$
$Z_1 \in \mathbb{R}^{N\times T}$
Merge
$Z \in \mathbb{R}^{N\times K'\times T}$
Effective bands
(B) Band split module
RNN across T
T
K'
$Z \in \mathbb{R}^{N\times K'\times T}$
FC
LSTM
Norm
$\times\mathcal{L}$
RNN across K'
T
K'
$\hat{Z} \in \mathbb{R}^{N\times K'\times T}$
FC
BLSTM
Norm
$\times\mathcal{L}$
$Q \in \mathbb{R}^{N\times K'\times T}$
(C) Band and sequence modeling module
$Q \in \mathbb{R}^{N\times K'\times T}$
$Q_{K'} \in \mathbb{R}^{N\times T}$
$Q_2 \in \mathbb{R}^{N\times T}$
$Q_1 \in \mathbb{R}^{N\times T}$
Norm + MLP
Norm + MLP
Norm + MLP
$M_K = \mathbf{0}^{G_k\times T}$
$M_{K'} \in \mathbb{C}^{G_{K'}\times T}$
$M_2 \in \mathbb{C}^{G_2\times T}$
$M_1 \in \mathbb{C}^{G_1\times T}$
Band merge
$M \in \mathbb{C}^{F\times T}$
(D) Mask estimation module

Fig. 1. (A) The overall pipeline for the USR-BSRNN framework. (B) The design of the band split module. (C) The design of the sequence and band modeling module. (D) The design of the mask estimation module.

RNN layers to uni-directional ones, the proposed USR-BSRNN can be causal and lightweight enough to be deployed to resource-constrained devices and platforms. Experiments conducted on the Voice Bank and DEMAND (VCTK) dataset [26] show that compared to various benchmark systems in speech enhancement and separation, our USR-BSRNN achieves comparable or better signal-to-noise ratio (SNR) performance at a same level of model size or complexity. In addition, the computation cost at 8k Hz is only 53.4% of that at 48k Hz.

The rest of the paper is organized as follows. Section 2 introduces the proposed USR-BSRNN architecture. Section 3 provides the configurations for training and evaluation. Section 4 presents the experiment results and analysis. Section 5 concludes the paper.

2. UNIVERSAL SAMPLE-RATE BAND-SPLIT RNN

The overall pipeline of the proposed USR-BSRNN model is demonstrated in Figure 1 (A). First, the input waveform is upsampled to 48k Hz sample rate and transformed into a complex-valued spectrogram by short-time Fourier transform (STFT). The spectrograms are then passed through a *band split module*, a *band and sequence modeling module* and a *mask estimation module* in turn to compute the complex-valued time-frequency (T-F) mask. The output is obtained by multiplying the mask with the input spectrogram followed by an inverse STFT operation, and the signal is finally downsampled back to the original input sample rate.

2.1. Band Split Module

The design of the band split module is illustrated in Figure 1 (B). The complex-valued spectrogram $\mathbf{X} \in \mathbb{C}^{F\times T}$, where F and T represent the frequency and temporal dimensions, respectively, are divided into K nonoverlapped frequency bands $\mathbf{B}_i \in \mathbb{C}^{G_i\times T}, i = 1, \ldots, K$ with pre-defined bandwidths $\{G_i\}_{i=1}^{K}$ satisfying $\sum_{i=1}^{K} G_i = F$. The valid frequency bands are defined as the subbands whose maximum frequency is not larger than the Nyquist frequency of the input signal, and we denote the number of valid frequency bands as K'. Then the real and imaginary parts of each valid frequency band $\{\mathbf{B}_i\}_{i=1}^{K'}$ are concatenated and passed to a layer normalization module [27] and a fully-connected (FC) layer to generate a real-valued subband feature $\mathbf{Z}_i \in \mathbb{R}^{N\times T}$. Note that each subband spectrogram has its own normalization module and FC layer as the bandwidths $\{G_i\}_{i=1}^{K}$ can all be different. All K' subband features $\{\mathbf{Z}_i\}_{i=1}^{K'}$ are then merged to generate a transformed feature tensor $\mathbf{Z} \in \mathbb{R}^{N\times K'\times T}$.

2.2. Band and Sequence Modeling Module

Figure 1 (C) shows the band and sequence modeling module. Similar to the dual-path RNN architecture [28], BSRNN performs interleaved sequence-level and band-level modeling via two different residual LSTM layers. To model the temporal information, the sequence-level RNN is first applied to $\mathbf{Z}$ across the temporal dimension T. Note that, all the K' subband features with identical feature dimension N share the same RNN layer to save the overall model size, which also enables parallel computation. To model the inter-band feature dependencies at each time step, the band-level RNN is applied to $\mathbf{Z}$ across the band dimension K'. Both sequence-level and band-level RNN modules share the same micro-design, where a batch normalization module is first applied to the input, and a (B)LSTM layer and an FC layer is applied to perform the modeling. Residual connection is added between the original input and the output of the FC layer. Multiple such RNNs can be stacked to create a deeper architecture, and the output of the last layer is denoted by $\mathbf{Q} \in \mathbb{R}^{N \times K' \times T}$. In our causal configuration, the uni-direction LSTM layer is applied in sequence-level RNN module, and we will compare the effect of uni-directional and bi-direction LSTM layers in the band-level RNN in Section 4.

2.3. Mask Estimation Module

The mask generation model is used to generate a complex-valued time-frequency (T-F) mask for extracting the target source. As shown in Figure 1 (D), $\mathbf{Q}$ is first evenly divided into K' features $\{\mathbf{Q}_i\}_{i=1}^{K'} \in \mathbb{R}^{N \times T}$ where each feature corresponds to the transformed feature of a valid subband, and each subband feature is passed to a batch normalization module followed by a multilayer perceptron (MLP) with one hidden layer to generate the real and imaginary parts of the T-F masks $\mathbf{M}_i \in \mathbb{C}^{G_i \times T}, i = 1, \ldots, K'$. Similar to the band split module, each frequency band has its own normalization module and MLP, and the masks for invalid frequency bands are set to zero. All $\mathbf{M}_i$ are then merged into the fullband T-F mask $\mathbf{M} \in \mathbb{C}^{F \times T}$ and multiplied with $\mathbf{X}$ to generate the target spectrogram $\mathbf{S} \in \mathbb{C}^{F \times T}$.

3. EXPERIMENT CONFIGURATIONS

3.1. Data Configuration

To verify the effectiveness of the proposed USR-BSRNN model, we conduct our experiments on the speech enhancement task with the public available Voice Bank and DEMAND (VCTK) dataset [26]. Specifically, the source speech comes from the VoiceBank corpus [26], which contains 28 speakers for training (≈11.5k utterances, ≈ 9.4 hrs) and another 2 speakers for testing (824 utterances, ≈ 0.6 hrs). The noisy speech is simulated using ten noise types with two artificial and eight real recordings from DEMAND with signal-to-noise (SNR) level of $[0, 5, 10, 15]$ dB. Note that the original 48 kHz VCTK data is also downsampled to 8 kHz, 16 kHz, and 32 kHz for both training and testing in our experiments.

3.2. Implementation Details

3.2.1. Hyperparameter Configurations

Band-spilt bandwidths: Given the observation from our previous work [5] that lower frequency bands are more important for vocal sounds, we split the frequency band below 1k Hz by a 100 Hz bandwidth, the frequency band between 1k Hz and 4k Hz by a 250 Hz bandwidth, the frequency band between 4k Hz and 8k Hz by a 500 Hz bandwidth, the frequency band between 8k Hz and 16k Hz by a 1k Hz bandwidth, the frequency band between 16k Hz and 20k Hz by a 2k Hz bandwidth, and treat the rest as one subband. This results in 41 subbands.

Model configuration: For all frequency-domain models in our experiments, the complex-valued spectrogram is computed by STFT with 2048-point and 512-point window and hop sizes, respectively, with a Hanning window. For the proposed USR-BSRNN, we set the feature dimension N to be 16 in all experiments, and use 6 band and sequence modeling modules with a total of 12 residual BLSTM layers. We set the hidden unit of BLSTM layers to be $2N = 32$, the hidden size in the mask estimation MLP to be $4N = 128$, and use the hyperbolic tangent function as the nonlinear activation function in the MLP. We use a gated linear unit (GLU) [29] for the output layer of the MLP.

3.2.2. Training Pipeline

To allow the model to handle signals with various sample rates, we use a multi-sample-rate training scheme different from prior works [24]. Specifically, during each iteration in the training phase, the original 48 kHz utterances are first randomly downsampled to [8, 16, 32] kHz then upsampled back to 48 kHz. This allows all models to utilize input signals at various sample rates for optimization. For the USR-BSRNN model, the valid frequency bands are determined by the sample rate used in the downsampling operation. All models in our experiment are trained for 144 epochs with Adam optimizer with an initial learning rate of $1e^{-3}$ and 0.99 exponential learning rate decay for every epoch. 8 Tesla P40 cards are used for model training with batch size of 12 per GPU. Early stopping is applied when the best validation is not found in 15 consecutive epochs.

3.2.3. Training Objective

All the models in our experiment are optimized by minimizing the sum of a frequency-domain mean-absolute-error (MAE) loss and a time-domain MAE loss:

$$\mathcal{L}_{obj} = |\mathbf{S}_r - \bar{\mathbf{S}}_r|_1 + |\mathbf{S}_i - \bar{\mathbf{S}}_i|_1 + |\text{iSTFT}(\mathbf{S}) - \text{iSTFT}(\bar{\mathbf{S}})|_1 \quad (1)$$

where $\bar{\mathbf{S}} \in \mathbb{C}^{F \times T}$ denotes the complex-valued spectrogram of the clean target, subscript r and l denote the real and imaginary parts, respectively.

3.2.4. Evaluation

To evaluate the performance of the proposed USR-BSRNN model, we select signal-to-Noise Ratio (SNR), the narrow-band and wide-band versions of perceptual evaluation of speech quality (PESQ-NB and PESQ-WB) [30], and the short-time objective intelligibility (STOI) [31] as the evaluation metrics. The calculation of PESQ-WB is done by downsample signals higher than 16k Hz sample rate to 16k Hz sample rate. Since SNR is a sample-rate-independent metric while PESQ-WB is particularly designed for evaluating 16k Hz signals, we mainly focus on the SNR results in our experiment. In addition, the computational complexity of each model is measured using multiply-and-accumulate per second[1] (MACs).

4. RESULTS AND ANALYSIS

Table 1 compares the model sizes, complexity and performance between USR-BSRNN and several recent benchmark systems, including CLDNN [32], DCCRN[2] [3], C-SuDoRm-RF[3] [17], and DPTFS-NET [6], on the VCTK dataset. We use the officially released implementation of all the models and train them with the same pipeline, while we modify the width and the depth of the models so that all models share a similar level of model size and complexity. We can see that USR-BSRNN outperforms other benchmarks models

[1] https://github.com/Lyken17/pytorch-OpCounter

[2] https://github.com/huyanxin/DeepComplexCRN/blob/master/dc_crn.py

[3] https://github.com/etzinis/sudo_rm_rf/blob/master/sudo_rm_rf/dnn/models/causal_improved_sudormrf_v3.py

in terms of SNR by at least 0.39 dB, and achieves comparable performance on PESQ-WB and STOI as the CLDNN model with a 14 times smaller model size.

Table 1. Comparison of different real-time speech enhancement systems on 48 kHz VCTK test set.

Model	Param (M)	MACs (G)	SNR (dB)	PESQ-WB	STOI (%)	RTF	MEM (M)
Noisy	–	–	8.4	1.97	92.1	–	–
RNNoise [15]	0.06	–	–	2.29	92.2	–	–
C-SuDoRM-RF [17]	0.61	0.53	16.4	2.11	92.1	0.36	305
DPTFSNET [6]	0.53	0.42	18.7	2.51	92.8	0.10	134
DCCRN [3]	0.72	0.53	18.6	2.53	93.3	**0.08**	91
CLDNN	4.0	0.52	18.6	**2.59**	**93.6**	0.09	52
USR-BSRNN	**0.35**	0.52	**19.1**	**2.59**	93.5	0.10	160 (10.5*)

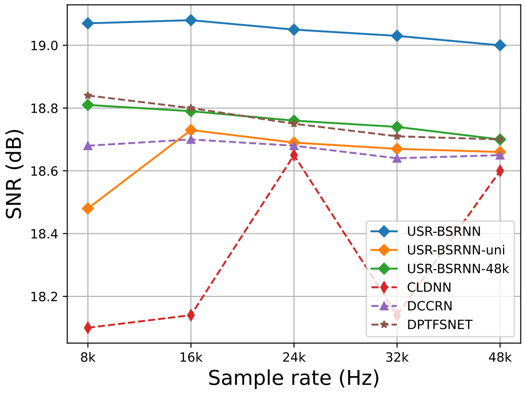
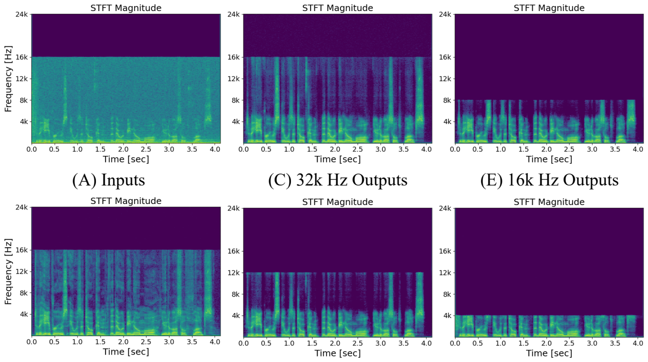


Fig. 2. SNR results of different USR-BSRNN configurations and other models at different sample rates.

Figure 2 further demonstrates the performance of different USR-BSRNN configurations and other models on different sample rates. Similar to the training phase, signals from lower sample rates are first upsampled to 48 kHz and passed to the models. The outputs are then downsampled back to calculate the evaluation metrics. The *USE-BSRNN-uni* model corresponds to the configuration where the bi-directional band-level RNN is replaced by a uni-directional RNN, and the *USR-BSRNN-48k* model corresponds to the configuration where only the 48 kHz signals are used during training while all sample rate signals are evaluated during testing. The *USR-BSRNN* model corresponds to the one described in Table 1, and the bi-directional band-level RNN always scans the valid number of frequency bands defined by the actual sample rate. The performance difference between *USR-BSRNN* and *USR-BSRNN-uni* shows that the bi-directional band-level modeling can effectively lead to performance gain compared to the uni-directional band-level modeling, and the performance difference between *USR-BSRNN* and *USR-BSRNN-48k* indicates the importance of multi-sample-rate training in this band-level modeling scheme. Moreover, *USR-BSRNN* performs consistently better across all sample rates than other systems which are all trained by multi-sample-rate training, and the performance of *USR-BSRNN* is also stable across all conditions. This proves the effectiveness of USR-BSRNN architecture. We also find that while all models have not seen any 24 kHz signals during training phase (as we mentioned in Section 3.1), the performance of all the systems at 24 kHz is consistent with it at other sample rates. This shows that the our multi-sample-rate training scheme is also helpful in other models.



(A) Inputs (C) 32k Hz Outputs (E) 16k Hz Outputs

(B) Targets (D) 24k Hz Outputs (D) 8k Hz Outputs

Fig. 3. Visualization of outputs from USR-BSRNN under different sample rates.

Table 2 shows the model complexity of USR-BSRNN across different sample rates. Unlike other models that have an identical model complexity at all sample rates, USR-BSRNN can achieve a dynamic computational complexity and can be more computational efficient at lower sample rates. For instance, the computation cost at 8 kHz is only 53.4% of that at 48 kHz.

Figure 3 provides the visualization of the magnitude spectrograms of a 32 kHz noisy input and its clean target together with the outputs from USR-BSRNN at different sample rates. We can see that the USR-BSRNN outputs can be accurately constrained to the target samples rates while generating consistent outputs at lower frequency bands. This further proves that USR-BSRNN is able to maintain similar model behavior at different sample rates.

Table 2. Performance and inference-time statistics of USR-BSRNN on VCTK test set under various sample rates.

Sample rate	MACs (G)	SNR (dB)	PESQ-WB	PESQ-NB	RTF	MEM (M)
8 kHz	0.29	19.14	–	3.35	0.06	88 (6*)
16 kHz	0.39	19.15	2.54	3.43	0.08	117 (7.9*)
24 kHz	0.44	19.12	2.56	3.43	0.09	135 (9.1*)
32 kHz	0.49	19.09	2.59	3.45	0.09	152 (9.9*)
48 kHz	0.52	19.07	2.59	3.44	0.10	160 (10.5*)

5. CONCLUSION

In this paper, we proposed universal-sample-rate band-split RNN (USR-BSRNN), a lightweight and computational efficient model architecture for real-time resource-efficient speech enhancement tasks. With band-split operation and interleaved band-level and sequence-level modeling, a single BSRNN model can handle signals at a wide range of sample rates, and the computational cost required to process a lower-sample-rate signal can be smaller than that of a higher-sample-rate signal. Experiment conducted on the VCTK dataset indicated that compared to various benchmark speech enhancement systems, the proposed USR-BSRNN model together with a multi-sample-rate training pipeline achieved comparable or better signal-to-noise ratio (SNR) performance at a same level of model size and complexity.

6. REFERENCES

[1] Ke Tan and DeLiang Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2019.

[2] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui 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, 2014.

[3] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv:2008.00264*, 2020.

[4] Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei 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.

[5] Yi Luo and Jianwei Yu, "Music source separation with band-split rnn," *arXiv:2209.15174*, 2022.

[6] Feng Dang, Hangting Chen, and Pengyuan Zhang, "Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6857–6861.

[7] Jean-Marc Valin, Umut Isik, Neerad Phansalkar, Ritwik Giri, Karim Helwani, and Arvindh Krishnaswamy, "A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech," *arXiv:2008.04259*, 2020.

[8] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi, "Real time speech enhancement in the waveform domain," *arXiv:2006.12847*, 2020.

[9] Andong Li, Wenzhe Liu, Chengshi Zheng, Cunhang Fan, and Xiaodong 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, pp. 1829–1843, 2021.

[10] Santiago Pascual, Antonio Bonafonte, and Joan Serra, "Segan: Speech enhancement generative adversarial network," *arXiv:1703.09452*, 2017.

[11] Andong Li, Chengshi Zheng, Lu Zhang, and Xiaodong Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Applied Acoustics*, vol. 187, pp. 108499, 2022.

[12] Eesung Kim and Hyeji Seo, "Se-conformer: Time-domain speech enhancement using conformer.," in *Interspeech*, 2021, pp. 2736–2740.

[13] Efthymios Tzinis, Yossi Adi, Vamsi K Ithapu, Buye Xu, Paris Smaragdis, and Anurag Kumar, "Remixit: Continual self-training of speech enhancement models via bootstrapped remixing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[14] Yangyang Xia, Sebastian Braun, Chandan KA Reddy, Harishchandra Dubey, Ross Cutler, and Ivan Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 871–875.

[15] Jean-Marc Valin, "A hybrid dsp/deep learning approach to real-time full-band speech enhancement," in *2018 IEEE 20th international workshop on multimedia signal processing (MMSP)*. IEEE, 2018, pp. 1–5.

[16] Hendrik Schroter, Alberto N Escalante-B, Tobias Rosenkranz, and Andreas Maier, "Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7407–7411.

[17] Efthymios Tzinis, Zhepei Wang, Xilin Jiang, and Paris Smaragdis, "Compute and memory efficient universal sound source separation," *Journal of Signal Processing Systems*, vol. 94, no. 2, pp. 245–259, 2022.

[18] Krishna Subramani and Paris Smaragdis, "Point cloud audio processing," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 31–35.

[19] Shubo Lv, Yihui Fu, Mengtao Xing, Jiayao Sun, Lei Xie, Jun Huang, Yannan Wang, and Tao Yu, "S-dccrn: Super wide band dccrn with learnable complex feature for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7767–7771.

[20] Xu Zhang, Lianwu Chen, Xiguang Zheng, Xinlei Ren, Chen Zhang, Liang Guo, and Bing Yu, "A two-step backward compatible full-band speech enhancement system," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7762–7766.

[21] Guochen Yu, Yuansheng Guan, Weixin Meng, Chengshi Zheng, and Hui Wang, "Dmf-net: A decoupling-style multi-band fusion model for real-time full-band speech enhancement," *arXiv:2203.00472*, 2022.

[22] Koichi Saito, Tomohiko Nakamura, Kohei Yatabe, Yuma Koizumi, and Hiroshi Saruwatari, "Sampling-frequency-independent audio source separation using convolution layer based on impulse invariant method," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 321–325.

[23] Jouni Paulus and Matteo Torcoli, "Sampling frequency independent dialogue separation," in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 160–164.

[24] David Samuel, Aditya Ganeshan, and Jason Naradowsky, "Meta-learning extractors for music source separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 816–820.

[25] Koichi Saito, Tomohiko Nakamura, Kohei Yatabe, and Hiroshi Saruwatari, "Sampling-frequency-independent convolutional layer and its application to audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2928–2943, 2022.

[26] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech.," in *SSW*, 2016, pp. 146–152.

[27] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, "Layer normalization," *arXiv:1607.06450*, 2016.

[28] Yi Luo, Zhuo Chen, and Takuya Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[29] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*. PMLR, 2017, pp. 933–941.

[30] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*. IEEE, 2001, vol. 2, pp. 749–752.

[31] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[32] Tara N Sainath, Oriol Vinyals, Andrew Senior, and Haşim Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 4580–4584.