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/ICASSP49357.2023.10095700

NEURAL SPEECH ENHANCEMENT WITH VERY LOW ALGORITHMIC LATENCY AND COMPLEXITY VIA INTEGRATED FULL- AND SUB-BAND MODELING

Zhong-Qiu Wang[1], *Samuele Cornell*[2], *Shukjae Choi*[3], *Younglo Lee*[3], *Byeong-Yeol Kim*[3], *Shinji Watanabe*[1]

[1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
[2]Università Politecnica delle Marche, Italy [3]Hyundai Motor Group and 42dot Inc., Seoul, Korea
wang.zhongqiu41@gmail.com

ABSTRACT

We propose FSB-LSTM, a novel long short-term memory (LSTM) based architecture that integrates full- and sub-band (FSB) modeling, for single- and multi-channel speech enhancement in the short-time Fourier transform (STFT) domain. The model maintains an *information highway* to flow an over-complete input representation through multiple FSB-LSTM modules. Each FSB-LSTM module consists of a full-band block to model spectro-temporal patterns at all frequencies and a sub-band block to model patterns within each sub-band, where each of the two blocks takes a down-sampled representation as input and returns an up-sampled discriminative representation to be added to the block input via a residual connection. The model is designed to have a low algorithmic complexity, a small run-time buffer and a very low algorithmic latency, at the same time producing a strong enhancement performance on a noisy-reverberant speech enhancement task even if the hop size is as low as 2 ms.

Index Terms— Low-complexity speech enhancement, frame-online speech enhancement, deep learning, hearing aids design.

1. INTRODUCTION

Deep learning has dramatically advanced speech enhancement in the past decade [1]. However, current enhancement models reporting strong performance usually consist of many layers of convolutional, recurrent or self-attention blocks. They are often computationally-intensive, resource-demanding and suffer from large processing latency not suitable for online real-time enhancement, with low-latency, low-complexity enhancement largely being under-explored. These issues prevent the deployment of modern neural speech enhancement models into real-world products such as hearing aids which usually have very limited computing capabilities, and dramatically limit the potential application range of deep neural network (DNN) based enhancement. As is suggested in [2, 3], an ideal neural speech enhancement system needs to have a small model size and consume a small amount of memory, computation and energy at training and inference time, meanwhile achieving strong enhancement performance with very low processing latency[1].

Many recent neural speech enhancement studies [4–9] have a particular focus on using a smaller model size to achieve stronger enhancement performance. Although very small model sizes are certainly desirable, in most modern edge devices a model size below 20 megabytes (MB) is typically satisfactory as the storage and RAM are usually much larger. The more pressing issues, we believe, are in the run-time memory cost, algorithmic complexity, and computation requirements when performing one-frame-in, one-frame-out enhancement in a real-time fashion. Solving these issues requires major changes to many current DNN architectures. For example,

- Attention mechanism [10–14], especially in its original form which attends to past frames to capture long-range context, is not ideal for low-complexity, online enhancement, since it needs to buffer many past frames and hence has a sizable memory cost;
- Although two-dimensional (2D) convolution (Conv2D) features a small number of parameters and has been popular in UNet-based speech enhancement in the magnitude [15–17], complex time-frequency (T-F) [18–29] and time domain [7], it usually costs a large amount of computation.
- State-of-the-art dual-path models such as DPRNN [6] and TF-GridNet [9] are not ideal for real-time enhancement, as they are computationally expensive. TF-GridNet, for example, runs an LSTM for each frequency at each layer, and at each frame it does not process all the steps in the sequence in parallel.

Equipped with these understandings, we think that recurrent neural networks such as LSTM [30] are more suitable for online, low-complexity speech enhancement, because at run time only one past frame needs to be buffered and the memory cost and system complexity can be low. In addition, small fully-connected blocks (or one-dimensional convolutions) are usually less costly than Conv2D blocks that use large kernels and large input and output channels.

In this context, we investigate using stacked LSTMs as the DNN backbone for frame-online speech enhancement with very low algorithmic latency and complexity. Although there have been studies exploring this direction [31–35], they are usually studied in monaural conditions and in teleconferencing scenarios where the allowed processing latency can be as high as 40 ms [36] and hence a regularly-large hop size (e.g., 8, 10 and 16 ms) is often used. In hearing aids setup, however, the requirement on algorithmic latency is usually less than 5 ms [37]. This means that the hop size cannot go beyond 2.5 ms if 50% frame-overlap is used in overlap-add, and such a small hop size would create longer frame sequences to process and requires hardware latency to be less than the hop size in order to realize real-time enhancement. In such cases, how to design a low-complexity DNN architecture that can leverage LSTMs to achieve single- and multi-channel enhancement with low algorithmic latency is an important problem to study.

In our experiments, we observe that a multi-layer unidirectional LSTM modeling full-band information performs impressively well even when the hop size is as low as 1 ms, indicating that LSTM could be very suitable for hearing aids design. We further integrate the full-band LSTM blocks with sub-band LSTM blocks so that complementary full- and sub-band information can be combined to achieve better enhancement, leading to a novel DNN architecture named FSB-LSTM for STFT-domain speech enhancement with very low algorithmic latency and low complexity. Evaluation results on single- and multi-channel speech enhancement in noisy-reverberant

[1]Processing latency consists of algorithmic latency resulting from algorithmic design (e.g., the use of overlap-add) and hardware latency for the computation at each frame [3].

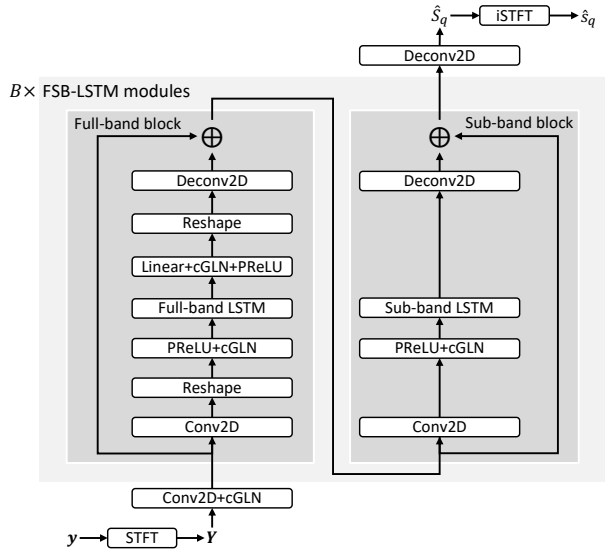

Fig. 1: Overview of proposed system.

conditions show the effectiveness of FSB-LSTM over other state-of-the-art low-latency streamable models in the time domain and in the complex T-F domain. Ablation studies also confirm the effectiveness of the proposed integrated full- and sub-band processing.

2. PROPOSED ALGORITHMS

Given a single-speaker, N-sample mixture recorded by a P-microphone array in noisy-reverberant conditions, the physical model in the time domain can be written as $\mathbf{y}[n] = \mathbf{s}[n] + \mathbf{v}[n]$, where $\mathbf{y}[n]$, $\mathbf{s}[n]$ and $\mathbf{v}[n] \in \mathbb{R}^P$ respectively denote the mixture, direct-path signal of the target speaker, and non-target signals at sample n. Our study aims at estimating the target direct-path signal captured by a reference microphone q (i.e., s_q) based on the mixture in a low-latency, low-complexity setup.

In the STFT domain, we denote the mixture as $\mathbf{Y}(t,f) = \mathbf{S}(t,f) + \mathbf{V}(t,f) \in \mathbb{C}^P$, where $\mathbf{Y}$, $\mathbf{S}$ and $\mathbf{V}$ are respectively the STFT spectra of $\mathbf{y}$, $\mathbf{s}$ and $\mathbf{v}$, t indexes T frames, and f indexes F frequencies. Our system operates in the STFT domain. Following [3], we use a regularly-large input window size (iWS) for STFT and a much smaller output window size (oWS) for overlap-add in inverse STFT (iSTFT), and both iWS and oWS are set to multiples of the hop size (HS). This way, our system can use an STFT with a regularly-high frequency resolution while still have a low algorithmic latency equal to the smaller oWS rather than the regularly-large iWS. See [3] for the details of this STFT-iSTFT mechanism.

Fig. 1 illustrates our proposed system. It is trained to perform multi-microphone complex spectral mapping based speech enhancement [23, 29, 38], where the real and imaginary (RI) components of the mixture $\mathbf{Y}$ are stacked as input features to predict the RI components of target speech S_q. Given an input tensor with shape $2P \times T \times F$, where $2P$ is because we stack the RI components at all the P microphones, we first use a Conv2D layer with kernel size 1×3 along time and frequency to get a D-dimensional embedding for each T-F unit, obtaining in a $D \times T \times F$ tensor. We then use B FSB-LSTM modules, each with a full-band and a sub-band block, to leverage spectral, spatial and temporal information to gradually refine the T-F embeddings. Next, a 2D deconvolution (Deconv2D) layer with kernel size 1×3 is used to predict the target RI components. Finally, inverse STFT (iSTFT) is applied for signal re-synthesis. The loss function is defined on the re-synthesized signal and its magnitude, following the Wav+Mag loss in [3]. The rest of this section describes the full- and sub-band blocks in FSB-LSTM. To avoid confusion, in Table 1 we summarize the hyper-parameters we will use to describe FSB-LSTM.

Table 1: Summary of model hyper-parameters.

Symbols	Description
B	Number of FSB-LSTM modules
D	Embedding dimension for each T-F unit
E	Output channels of Conv2D in full-band blocks
I	Kernel size along frequency in Conv2D and Deconv2D in full-band blocks
J	Stride size along frequency in Conv2D and Deconv2D in full-band blocks
H	Number of hidden units in full-band LSTMs
E'	Output channels of Conv2D in sub-band blocks
I'	Kernel size along frequency in Conv2D and Deconv2D in sub-band blocks
J'	Stride size along frequency in Conv2D and Deconv2D in sub-band blocks
H'	Number of hidden units in sub-band LSTMs

2.1. Full-Band Block

Given an input tensor with shape $D \times T \times F$, we compress the D-dimensional T-F embeddings within each frame into a frame-level embedding, use an LSTM to refine the frame embedding, and re-compute D-dimensional T-F embeddings based on the refined frame embedding. This way, the LSTM can model all the frequencies at the same time to capture full-band information.

Specifically, we first use a Conv2D layer with input channel D, output channel E, kernel size $1 \times I$, and stride $1 \times J$ to compress the $D \times T \times F$ tensor along dimension one and three to $E \times T \times (\frac{Q-I}{J} + 1)$, after zero-padding the frequency dimension to $Q = \lceil \frac{F-I}{J} \rceil \times J + I$. We then reshape it to a 2D tensor by flattening the first and third dimensions to obtain a tensor with shape $T \times A$ with $A = E \times (\frac{Q-I}{J} + 1)$, apply PReLU, and perform causal global layer normalization (cGLN) [4], which computes the mean and variance for normalization based on both dimensions in a causal way and uses two A-dimensional vectors to respectively scale and shift along the first dimension. Next, we use an LSTM with H hidden units to model the A-dimensional frame embeddings, obtaining a tensor with shape $T \times H$. After that, a linear layer is applied to map the H-dimensional embedding to A-dimensional, followed by cGLN and PReLU. Finally, we reshape the $T \times A$ tensor back to $E \times T \times (\frac{Q-I}{J} + 1)$, and use a Deconv2D layer with input channel E, output channel D, kernel size $1 \times I$, and stride $1 \times J$ to compute a $D \times T \times F$ tensor, which is added to the original input tensor to this full-band block via a residual connection, after removing padded zeros.

2.2. Sub-Band Block

In [9, 39], we find that using sub-band modules to leverage sub-band information is very effective at dereverberation and leveraging spatial information. However, TF-GridNet proposed in [9, 39] runs a sub-band module at each frequency, consuming a large amount of computation. To reduce the computation, we reduce the number of frequencies by using convolution based down-sampling, and use much fewer input, hidden and output units in LSTMs.

In detail, given an input tensor with shape $D \times T \times F$, we first use a Conv2D layer with input channel D, output channel E', kernel size $1 \times I'$, and stride size $1 \times J'$ to down-sample the $D \times T \times F$ tensor along dimension one and three to $E' \times T \times (\frac{Q'-I'}{J'} + 1)$, after zero-padding the frequency dimension to $Q' = \lceil \frac{F-I'}{J'} \rceil \times J' + I'$. Next, we apply PReLU, and perform cGLN, which, in the case of 3D tensors, computes the mean and variance for normalization based on all the three dimensions in a causal way and uses two E'-dimensional vectors to respectively scale and shift along the first dimension. After that, we view the tensor as $\frac{Q'-I'}{J'} + 1$ sequences, each with length T, and use an LSTM with H' hidden units to refine

the E'-dimensional T-F embeddings, obtaining a tensor with shape $H' \times T \times (\frac{Q'-I'}{J'}+1)$. Note that the LSTM is shared across all the sub-bands to reduce model parameters. At last, we use a Deconv2D layer with input channel H', output channel D, kernel size $1 \times I'$, and stride size $1 \times J'$ to compute a $D \times T \times F$ tensor, which is added to the input tensor of the sub-band block through a residual connection, after removing padded zeros.

2.3. Discussion on Network Design

In our network, we maintain an *information highway*, which flows an over-complete T-F representation (i.e., the D-dimensional T-F embeddings) of multi-channel input signals inside the network through residual connections. When refining the T-F embeddings, we first perform down-sampling to extract desired discriminative features, then use LSTM layers to perform full- or sub-band temporal modeling, and finally up-sample and add it back to the input tensor. This could be a good strategy especially for multi-microphone enhancement, as it can maintain the fine-grained information of multiple input signals (e.g., spectro-temporal and spatial patterns) and at the same time extract different discriminative information of interest at different blocks for better enhancement.

In comparison, an early popular way of using stacked LSTMs feeds $2P \times F$-dimensional input features directly to a multi-layer LSTM (in a way similar to that in the deep clustering [40, 41] and permutation invariant training [42] studies). However, since the hidden dimension of LSTM is usually much smaller than the input dimension when P is large, the resulting model would be limited at exploiting spatial information due to the compression of input features. Similarly, modern time-domain models such as Conv-TasNet [4, 43] tend to create a bottleneck representation immediately after the encoder. Such a bottleneck could lead to loss of information when the input dimension is high (e.g., in multi-microphone cases). One solution is to use recurrent U-Net based models for multi-channel separation [23,29,38], where the lower layers in the U-Net encoder (and the corresponding layers in the decoder) can have an over-complete representation of input features to maintain fine-grained patterns, and the input features are gradually down-sampled to a dimension suitable for recurrent networks. In our experiments, we will show that FSB-LSTM produces better performance than a strong recurrent U-Net based model [44] and a multi-channel Conv-TasNet [43, 45].

All the convolutions in our models have a kernel size of one along time. This way, we avoid buffering past frames due to the use of causal convolution, and just use LSTMs to model temporal information. The Conv2D and Deconv2D layers in our models are not very costly, as they are used with a large stride size, a small kernel size, and few input and output channels. For deconvolution, we use a custom implementation[2] to reduce the number of multiply–accumulate (MAC) operations.

We emphasize that the proposed network only needs to buffer the hidden and cell states of LSTMs in the past frame. The run-time memory cost and complexity of maintaining the buffer is low.

[2]Deconvolution (*a.k.a* transposed convolution) [30] is typically implemented by first interleaving zeros to the input tensor based on the stride size and then performing regular convolution. This increases the MAC operations when the stride is larger than one, because the new input tensor would have more elements to convolve due to the interleaved zeros. In our study, we implement deconvolution as a linear layer followed by overlap-add along frequency (please do not confuse this overlap-add with that in iSTFT). This way, we can save the computation wasted on the interleaved zeros. On the other hand, the overlap-add usually costs negligible MAC operations compared to the linear layer. The number of MAC operations of our implementation is roughly $1/J$ of that of the typical implementation, where J is the stride size.

3. EXPERIMENTAL SETUP

We validate our algorithms on a simulated noisy-reverberant speech enhancement task. This section describes the dataset, system configurations, baseline systems, and evaluation metrics.

3.1. Dataset

We use a simulated data, which was used in recent studies [3, 44], to evaluate the proposed algorithms. Using the split of clean speech in WSJCAM0, the dataset simulates 39,245 ($\sim$77.7 h), 2,965 ($\sim$5.6 h) and 3,260 ($\sim$8.5 h) noisy-reverberant mixtures respectively for training, validation and testing. The clips in the development set of FSD50k [46] are sampled to simulate the noises for training and validation, and those in the evaluation set for testing. Each simulated mixture contains up to seven noise clips, with one longer than ten seconds as background and the others as foreground noises. The simulated microphone array contains six microphones arranged uniformly on a circle with a diameter of 20 cm. The direction of each source to the array center is sampled from the range $[0, 2\pi)$, distance from $[0.75, 2.5]$ m, and the reverberation time from $[0.2, 1.0]$ s. We treat each sound source as a point source, convolve each source with a simulated room impulse response, and summate the convolved sources to create the mixture. The signal-to-noise ratio between the target direct-path speech and reverberant noise is drawn from $[-8, 3]$ dB. The sampling rate is 16 kHz. For two-channel processing, we use signals at the first and the fourth microphones; and for monaural processing, the first microphone is used. The target direct-path signal captured at the first microphone is used as the label for model training and as the reference for metric computation.

3.2. System Configurations

We aim at an enhancement system with an algorithmic latency of 4 ms, which is slightly shorter than the 5 ms requirement suggested in the recent Clarity challenge [37] proposed for hearing aids design. For STFT and iSTFT, in default the iWS is set to 16 ms, HS to 2 ms, and oWS to 4 ms, resulting in an algorithmic latency of 4 ms [3]. The rectangular window is used as the analysis window. Given a sampling rate of 16 kHz, a 256-point discrete Fourier transform is used to extract 129-dimensional complex spectra at each frame. Through the validation set, we set $B = 3$, $D = 32$, $E = 8$, $I = 8$, $J = 4$, $H = 256$ $E' = 64$, $I' = 5$, $J' = 5$, and $H' = 64$ (see Table 1 for the definition of the notations). In this configuration, there are 26 sub-bands and the MAC operations of the sub-band module are around twice as many as the full-band module.

3.3. Baseline Systems

We consider Conv-TasNet [4], its multi-channel extension MC-Conv-TasNet [43, 45], LSTM-ResUNet [3], and a full-band only LSTM model as the major baselines. All of them are trained with the same loss as FSB-LSTM.

Conv-TasNet [4] is an excellent time-domain model in speech separation. It uses learned bases on very short windows of signals to achieve separation with very low algorithmic latency. Using the symbols listed in Table I of the Conv-TasNet paper [4], we set the hyper-parameters of Conv-TasNet and MC-Conv-TasNet to $N = 512, B = 158, S_c = 158, H = 512, P = 3, X = 8$, and $R = 3$ (please do not confuse these symbols with those defined in this paper). B and S_c are set slightly larger than the default 128 suggested in [4], considering that in MC-Conv-TasNet there are additional spatial embeddings concatenated to the spectral embeddings as the input to the separator of Conv-TasNet. Following [43, 45], the spatial embedding dimension is set to 60 for two-channel enhancement and to 360 for six-channel enhancement.

Table 2: Results of FB-LSTM at various hop sizes (6ch).

Systems	iWS (ms)	oWS (ms)	HS (ms)	#params (M)	GMAC/s	SI-SDR (dB)	PESQ	eSTOI
Unprocessed	-	-	-	-	-	−6.2	1.44	0.411
FB-LSTM (6-layer)	16	16	8	3.59	0.58	3.9	2.06	0.721
FB-LSTM (6-layer)	16	8	4	3.59	1.16	5.8	2.28	0.776
FB-LSTM (6-layer)	16	4	2	3.59	2.33	6.8	2.37	0.795
FB-LSTM (6-layer)	16	2	1	3.59	4.65	**8.0**	**2.54**	**0.821**

Table 3: Results on speech enhancement (6ch).

Systems	#params (M)	GMAC/s	SI-SDR (dB)	PESQ	eSTOI
Unprocessed	-	-	−6.2	1.44	0.411
FSB-LSTM	1.96	3.37	**7.8**	**2.61**	**0.830**
FB-LSTM (6-layer)	3.59	2.33	6.8	2.37	0.795
FB-LSTM (9-layer)	5.38	3.43	7.6	2.51	0.816
MC-Conv-TasNet [43, 45]	6.37	3.76	5.2	2.24	0.764
LSTM-ResUNet [3]	2.33	3.54	6.2	2.23	0.767

LSTM-ResUNet [3] is a representative complex T-F domain model, consisting of a multi-layer LSTM sandwiched by a UNet with residual net blocks inserted at multiple frequency scales. It uses a shorter oWS than the iWS in overlap-add to realize enhancement with low algorithmic latency [3]. We emphasize that its network architecture shares many similarities with recent complex T-F domain models [17, 18, 20, 29] in speech enhancement. We therefore consider it as a major baseline in addition to Conv-TasNet.

To show the effectiveness of including sub-band LSTMs, we replace the sub-band module in Fig. 1 with the full-band module. This way, the system essentially stacks multiple full-band LSTMs. We denote this system as **FB-LSTM**, where "FB" means full-band. We experiment FB-LSTM with 6 and 9 full-band LSTM blocks, since we use $B = 3$ full- and sub-band modules in FSB-LSTM (totalling $2 \times 3 = 6$ LSTMs) and each sub-band module costs roughly twice as many MAC operations as the full-band module.

3.4. Evaluation Metrics

The evaluation metrics include scale-invariant signal-to-distortion ratio (SI-SDR), perceptual evaluation of speech quality (PESQ), and extended short-time objective intelligibility (eSTOI). For PESQ, we use the *python-pesq* (v0.0.2) toolkit to report narrow-band MOS-LQO scores. The number of model parameters is reported in millions (M). Using the *ptflops* toolkit, we report the amount of computation by counting MAC in giga-operations per second (GMAC/s).

4. EVALUATION RESULTS

4.1. Effectiveness of LSTM at Dealing with Small Hop Sizes

Although LSTM has been criticized for not being good enough at modeling long sequences resulted from small hop sizes [4], in our experiments (which focus on frame-online enhancement) we find FB-LSTM performing surprisingly well even if the hop size is as low as 1 ms. See Table 2 for the results. The iWS is always 16 ms. We reduce HS together with oWS so that the frame-overlap ratio in overlap-add is always 50% when the algorithmic latency (equal to oWS) becomes smaller. Every time HS is halved, the amount of computation is approximately doubled as the number of frames to process is doubled. From the results, we observe that a smaller HS (and oWS) leads to better performance, even though the resulting frame sequence gets much longer and the future context information that can be utilized (up to oWS to the future) becomes less.

Although using smaller hop sizes was found effective in time-domain speaker separation studies such as Conv-TasNet [4] and DPRNN [6], they are often used with more advanced architectures rather than simple uni-directional LSTMs that model frame sequences from left to right. Our study observes that such simple causal LSTMs can perform reasonably well for hop sizes as low as 1 ms. This finding is very significant, as it indicates that simple LSTMs, which have very low run-time complexity, can produce promising enhancement results in a hearing aid setup which requires very low processing latency.

Table 4: Results on speech enhancement (2ch).

Systems	#params (M)	GMAC/s	SI-SDR (dB)	PESQ	eSTOI
Unprocessed	-	-	−6.2	1.44	0.411
FSB-LSTM	1.96	3.31	**4.9**	**2.20**	**0.753**
FB-LSTM (6-layer)	3.59	2.27	4.2	2.07	0.729
FB-LSTM (9-layer)	5.38	3.37	4.5	2.14	0.744
MC-Conv-TasNet [43, 45]	6.19	3.68	3.6	2.00	0.711
LSTM-ResUNet [3]	2.33	3.48	4.3	2.06	0.726

Table 5: Results on speech enhancement (1ch).

Systems	#params (M)	GMAC/s	SI-SDR (dB)	PESQ	eSTOI
Unprocessed	-	-	−6.2	1.44	0.411
FSB-LSTM	1.96	3.30	**3.1**	**1.92**	**0.688**
FB-LSTM (6-layer)	3.59	2.25	2.6	1.84	0.671
FB-LSTM (9-layer)	5.38	3.35	2.5	1.83	0.667
Conv-TasNet [4]	6.18	3.67	2.2	1.78	0.657
LSTM-ResUNet [3]	2.32	3.47	2.8	1.90	0.682

4.2. Results of FSB-LSTM

Table 3, 4 and 5 respectively present the results of FSB-LSTM on six-, two- and one-channel speech enhancement. We can see that, with a smaller model size and using fewer MAC/s operations, FSB-LSTM produces better enhancement than Conv-TasNet, MC-Conv-TasNet and LSTM-ResUNet. Note that both Conv-TasNet and LSTM-ResUNet contain convolutions dilated along time and need to buffer many past frames at run time.

FSB-LSTM produces better results than 6-layer and 9-layer FB-LSTM. This shows the benefits of using the sub-band blocks.

4.3. Run-Time Complexity

We compute the run-time buffer size of each model in an online, streaming setup, based on single-precision floating-point operations. FSB-LSTM only needs to buffer LSTMs' hidden and cell states in the past frame. It has a buffer size of 46.2 kilobytes (KB) to maintain, while the buffer sizes of Conv-TasNet and LSTM-ResUNet are respectively 3133.4 and 1815.1 KB. Such a small buffer size makes it possible to have the buffered tensors reside in a higher cache hierarchy, which has very limited space (e.g., tens of KB at Level 1 and several MB at Level 2) even in modern processors. The small buffer size and the low algorithmic complexity also make it easier for the hardware latency to be smaller than the hop size to realize real-time enhancement in resource-constrained hearing-aid scenarios.

5. CONCLUSION

We have proposed a novel FSB-LSTM architecture that integrates full- and sub-band modeling for low-complexity, low-algorithmic-latency speech enhancement. Our experiments show that FSB-LSTM outperforms previously proposed state-of-the art streamable, low-latency models with much less buffer memory and less computational burden in MAC operations. Future research will further reduce algorithmic latency and explore DNN quantization and distillation to further reduce complexity, at the same time maintaining a strong enhancement performance.

6. REFERENCES

[1] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, pp. 1702–1726, 2018.

[2] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo RM -RF: Efficient Networks for Universal Audio Source Separation," in *Proc. MLSP*, 2020.

[3] Z.-Q. Wang, G. Wichern, S. Watanabe, and J. Le Roux, "STFT-Domain Neural Speech Enhancement with Very Low Algorithmic Latency," in *arXiv preprint arXiv:2204.09911*, 2022.

[4] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[5] Y. Liu and D. Wang, "Causal Deep CASA for Monaural Talker-Independent Speaker Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, pp. 1270–1279, 2020.

[6] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation," in *Proc. ICASSP*, 2020, pp. 46–50.

[7] A. Pandey and D. Wang, "Dense CNN with Self-Attention for Time-Domain Speech Enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1270–1279, 2021.

[8] L. Yang, W. Liu, and W. Wang, "TFPSNet: Time-Frequency Domain Path Scanning Network for Speech Separation," in *Proc. ICASSP*, 2022, pp. 6842–6846.

[9] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee *et al.*, "TF-GridNet: Making Time-Frequency Domain Models Great Again for Monaural Speaker Separation," in *Proc. ICASSP*, 2023.

[10] A. Vaswani, , N. Shazeer, N. Parmar *et al.*, "Attention Is All You Need," in *Proc. NIPS*, 2017.

[11] C. Subakan, M. Ravanelli, S. Cornell *et al.*, "Attention Is All You Need In Speech Separation," in *Proc. ICASSP*, 2021, pp. 21–25.

[12] ——, "Resource-Efficient Separation Transformer," *arXiv preprint arXiv:2206.09507*, 2022.

[13] G. Zhang, C. Wang, L. Yu, and J. Wei, "Multi-Scale Temporal Frequency Convolutional Network with Axial Attention for Multi-Channel Speech Enhancement," in *Proc. ICASSP*, 2022, pp. 9206–9210.

[14] A. Pandey and D. Wang, "Self-Attending RNN for Speech Enhancement to Improve Cross-Corpus Generalization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1374–1385, 2022.

[15] K. Tan and D. Wang, "A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement," in *Proc. Interspeech*, 2018, pp. 3229–3233.

[16] O. Ernst, S. E. Chazan *et al.*, "Speech dereverberation using fully convolutional networks," in *Proc. EUSIPCO*, 2018, pp. 390–394.

[17] K. Tan and D. Wang, "Learning Complex Spectral Mapping With Gated Convolutional Recurrent Networks for Monaural Speech Enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 380–390, 2020.

[18] Y. Liu and D. Wang, "Divide and Conquer: A Deep CASA Approach to Talker-Independent Monaural Speaker Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2092–2102, 2019.

[19] U. Isik, R. Giri *et al.*, "PoCoNet: Better Speech Enhancement with Frequency-Positional Embeddings, Semi-Supervised Conversational Data, and Biased Loss," in *Proc. Interspeech*, 2020, pp. 2487–2491.

[20] Y. Hu, Y. Liu, S. Lv, M. Xing *et al.*, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.

[21] Z.-Q. Wang and D. Wang, "Deep Learning Based Target Cancellation for Speech Dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 941–950, 2020.

[22] Z.-Q. Wang, P. Wang, and D. Wang, "Complex Spectral Mapping for Single-and Multi-Channel Speech Enhancement and Robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1778–1787, 2020.

[23] ——, "Multi-Microphone Complex Spectral Mapping for Utterance-Wise and Continuous Speech Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2001–2014, 2021.

[24] H. Taherian, Z.-Q. Wang, J. Chang, and D. Wang, "Robust Speaker Recognition Based on Single-Channel and Multi-Channel Speech Enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1293–1302, 2020.

[25] Z.-Q. Wang, G. Wichern, and J. Le Roux, "Convolutive Prediction for Monaural Speech Dereverberation and Noisy-Reverberant Speaker Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3476–3490, 2021.

[26] S. Zhao, B. Ma, K. N. Watcharasupat, and W.-S. Gan, "FRCRN: Boosting Feature Representation using Frequency Recurrence for Monaural Speech Enhancement," in *Proc. ICASSP*, 2022, pp. 9281–9285.

[27] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang *et al.*, "Personalized Speech Enhancement: New Models and Comprehensive Evaluation," in *Proc. ICASSP*, 2022.

[28] H. Taherian, S. E. Eskimez, T. Yoshioka, H. Wang *et al.*, "One Model to Enhance Them All: Array Geometry Agnostic Multi-Channel Personalized Speech Enhancement," in *Proc. ICASSP*, 2022, pp. 271–275.

[29] K. Tan *et al.*, "Neural Spectrospatial Filtering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 605–621, 2022.

[30] A. Courville, I. Goodfellow, and Y. Bengio, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[31] H. Zhao, S. Zarar *et al.*, "Convolutional-Recurrent Neural Networks for Speech Enhancement," in *Proc. ICASSP*, 2018, pp. 2401–2405.

[32] J. M. Valin, U. Isik, N. Phansalkar, R. Giri *et al.*, "A Perceptually-Motivated Approach for Low-Complexity, Real-Time Enhancement of Fullband Speech," in *Proc. Interspeech*, 2020, pp. 2482–2486.

[33] Y. Xia, S. Braun, C. K. Reddy, H. Dubey *et al.*, "Weighted Speech Distortion Losses for Neural-Network-Based Real-Time Speech Enhancement," in *Proc. ICASSP*, 2020, pp. 871–875.

[34] S. Braun, H. Gamper *et al.*, "Towards Efficient Models for Real-Time Deep Noise Suppression," in *Proc. ICASSP*, 2021, pp. 656–660.

[35] M. Thakker, S. E. Eskimez, T. Yoshioka, and H. Wang, "Fast Real-Time Personalized Speech Enhancement: End-to-End Enhancement Network (E3Net) and Knowledge Distillation," in *Proceedings of Interspeech*, 2022, pp. 991–995.

[36] C. K. Reddy, V. Gopal *et al.*, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Speech Quality and Testing Framework," in *Proc. Interspeech*, 2020, pp. 2492–2496.

[37] "Clarity Challenge: Machine learning Challenges for Hearing Devices." [Online]. Available: http://claritychallenge.org/

[38] Z.-Q. Wang and D. Wang, "Multi-Microphone Complex Spectral Mapping for Speech Dereverberation," in *Proc. ICASSP*, 2020, pp. 486–490.

[39] Z.-Q. Wang, S. Cornell, S. Choi *et al.*, "TF-GridNet: Integrating Full- and Sub-Band Modeling for Speech Separation," *arXiv preprint arXiv:2211.12433*, 2022.

[40] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep Clustering: Discriminative Embeddings for Segmentation and Separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[41] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-Channel Deep Clustering: Discriminative Spectral and Spatial Embeddings for Speaker-Independent Speech Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 1–5.

[42] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multi-Talker Speech Separation with Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[43] J. Zhang, C. Zorila, R. Doddipatla, and J. Barker, "On End-to-End Multi-Channel Time Domain Speech Separation in Reverberant Environments," in *Proc. ICASSP*, 2020, pp. 6389–6393.

[44] Z.-Q. Wang, G. Wichern, and J. Le Roux, "Leveraging Low-Distortion Target Estimates for Improved Speech Enhancement," *arXiv preprint arXiv:2110.00570*, 2021.

[45] Z. Tu, J. Zhang *et al.*, "A Two-Stage End-to-End System for Speech-in-Noise Hearing Aid Processing," in *Proc. Clarity*, 2021, pp. 3–5.

[46] E. Fonseca, X. Favory, J. Pons, F. Font *et al.*, "FSD50K: An Open Dataset of Human-Labeled Sound Events," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2021.