IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 October 2022, accepted 6 November 2022, date of publication 10 November 2022, date of current version 16 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3221440

RESEARCH ARTICLE

Speech Enhancement Using MLP-Based Architecture With Convolutional Token Mixing Module and Squeeze-and-Excitation Network

HYUNGCHAN SONG, (Graduate Student Member, IEEE), MINSEUNG KIM, (Graduate Student Member, IEEE), AND JONG WON SHIN, (Member, IEEE)

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea

Corresponding author: Jong Won Shin (jwshin@gist.ac.kr)

This work was supported in part by the National Research Foundation of Korea under Grant NRF-2019R1A2C2089324; and in part by the Ministry of Science and ICT (MSIT), South Korea, under the Information Technology Research Center (ITRC) Support Program supervised by the Institute of Information and Communications Technology Planning and Evaluation (IITP), under Grant IITP-2021-0-01835.

ABSTRACT The Conformer has shown impressive performance for speech enhancement by exploiting the local and global contextual information, although it requires high computational complexity and many parameters. Recently, multi-layer perceptron (MLP)-based models such as MLP-mixer and gMLP have demonstrated comparable performances with much less computational complexity in the computer vision area. These models showed that all-MLP architectures may perform as good as more advanced structures, but the nature of the MLP limits the application of these architectures to the input with a variable length such as speech and audio. In this paper, we propose the cgMLP-SE model, which is a gMLP-based architecture with convolutional token mixing modules and squeeze-and-excitation network to utilize both local and global contextual information as in the Conformer. Specifically, the token-mixing modules in gMLP are replaced by convolutional layers, squeeze-and-excitation network-based gating is applied on top of the convolutional gating module, and additional feed-forward layers are added to make the cgMLP-SE module a macaron-like structure sandwiched by feed-forward layers like a Conformer block. Experimental results on the TIMIT-DNS noise dataset and the Voice Bank-DEMAND dataset showed that the proposed method exhibited similar speech quality and intelligibility to the Conformer with a smaller model size and less computational complexity.

INDEX TERMS Speech enhancement, local and global information, low computational complexity.

I. INTRODUCTION

Speech enhancement is one of the essential tasks in many applications such as voice communication [1], speech recognition [2], and speech emotion recognition [3]. The goal of speech enhancement is to reduce the effect of background noises in noisy speech signals while preserving speech quality and intelligibility. Recently, deep learning-based speech enhancement approaches have brought about significant performance improvement by exploiting complex temporal-spectral dependencies in a data-driven way [4], [5], [6]. One of the most successful approaches is the Conformer [7], which consists of a multi-head self-attention (MHSA) module [8] followed by a convolution module sandwiched by feed-forward networks. By employing both the MHSA which is good at capturing long-term dependencies and the convolution module which effectively considers local contextual information, the Conformer has exhibited impressive performance in speech recognition [7], speaker verification [9], speech enhancement [10], and separation [11] while requiring many parameters and high computational complexity.

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Kamrul Hasan.

Recently, multi-layer perceptron (MLP)-based models such as the MLP-mixer [12] and gMLP [13] have been proposed and showed performances comparable to the Transformer-based model utilizing MHSA modules in image classification task [14] with less computational complexity. Although these all-MLP structures showed the potential to replace more advanced structures requiring heavy computation, the nature of the MLP may limit the application of these models to the data with a variable length such as speech signals. A block in the MLP-mixer [12] consists of the token-mixing module aggregating information in the tokens in each channel, which are frames when applied to speech processing, and the channel-mixing module applying MLPs across channels for each token. The gMLP block [13] can also be interpreted as a combination of channel-mixing and token-mixing modules in which the token-mixing module works as a gating unit. In [15], four variations of the gMLP are proposed to deal with data of variable length for the speech recognition task. Two of the variations that showed better performances were utilizing convolutional layers instead of linear layers in the token-mixing module, which exhibited similar performance to Transformer-encoder [8]. While the convolutional layers can process variable length sequences and capture short-term correlations in speech, long-term contextual information may not easily be handled by convolutional layers compared with MLPs. In [16] and [17], MLP-based models were applied to speech or audio signals of fixed maximum length. A keyword spotting method based on a structure similar to the MLP-mixer employing the dynamic convolution [18] and the squeeze-and-excitation network (SENet) [19] is proposed in [20]. However, it can only be applied to the input with a fixed length as the 1×1 convolution and SENet are applied to both the channel dimension and the frame dimension.

To deal with speech signals of variable length while capturing both global and local contextual information as in the Conformer, we propose the cgMLP-SE, a gMLP-based architecture with three modifications. Firstly, convolutional token-mixing modules are employed as in [15] to process variable-length data. Secondly, SENet [19] is applied on top of the convolutional gating unit in the token-mixing module to apply weights to channels considering the whole utterance, which was shown to be effective in many researches [21], [22], [23]. Lastly, an additional feed-forward network is added as the first module of the proposed cgMLP-SE block to make it similar to the macaron-like structure of the Conformer to boost performance. As a result, the proposed cgMLP-SE model considers local and global information like the Conformer [7] with low computational complexity and small model size thanks to the gMLP-like structure. We have conducted experiments to compare the speech enhancement performance of the proposed model with the models that can deal with the sequence of an arbitrary length [11], [24], [25], [26] and the variations of the MLP-based models [12], [13] with the convolutional token-mixing modules. Experimental results on the TIMIT-DNS noise dataset and the Voice Bank-DEMAND dataset [27] showed that the proposed cgMLP-SE model exhibited comparable or better performance than the previously proposed approaches in terms of the objective speech quality and intelligibility measures with relatively low computational complexity.

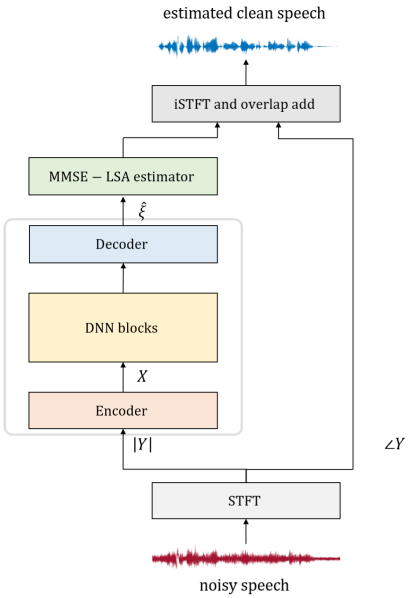

FIGURE 1. Block diagram of the DeepMMSE framework for speech enhancement.

II. SPEECH ENHANCEMENT AND DeepMMSE FRAMEWORK

Let $Y(l,k)$, $S(l,k)$, and $N(l,k)$ denote the short-time Fourier transform (STFT) of the noisy speech, clean speech, and noise for the frame l and the frequency k, respectively. Under the additive noise assumption, they can be related as

$$Y(l,k) = S(l,k) + N(l,k). \tag{1}$$

Speech enhancement is a task to estimate $S(l,k)$ based on the noisy observation $Y(l,k)$. One of the recent speech enhancement approaches is the DeepMMSE [24], [28], which incorporates a deep learning network into the statistical model-based speech enhancement framework. It basically follows the minimum mean square error log-spectral amplitude (MMSE-LSA) estimator [29] framework except for the way to obtain the key parameter, *a priori* signal-to-noise ratio (SNR), which is estimated by deep neural networks (DNNs). The *a priori* SNR ξ is defined as

$$\xi(l,k) = \frac{\Phi_S(l,k)}{\Phi_N(l,k)}, \tag{2}$$

in which Φ_S and Φ_N are the power spectral densities of speech and noise, respectively. In [28] and [24], the instantaneous estimate of ξ is used as the training target, which is given by

$$\tilde{\xi}(l,k) = \frac{|S(l,k)|^2}{|N(l,k)|^2}. \tag{3}$$

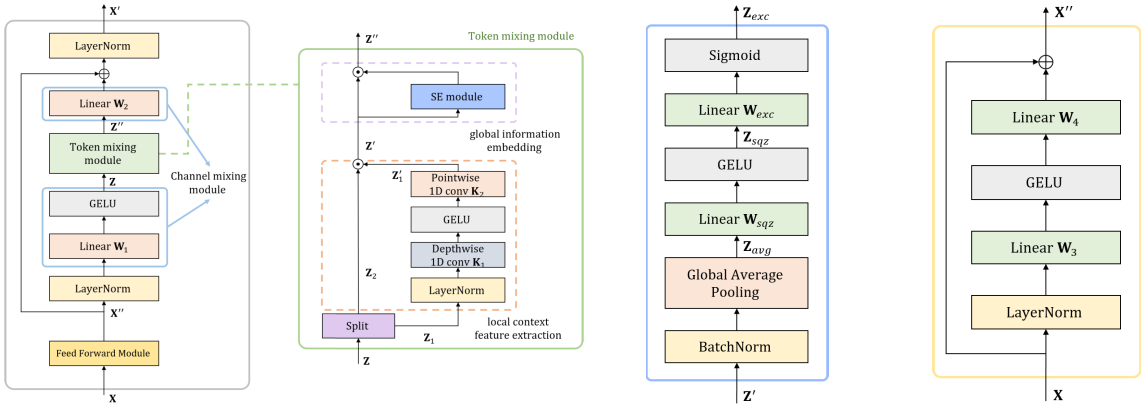

FIGURE 2. Illustrations of the proposed cgMLP-SE blocks for speech enhancement: (a) the proposed cgMLP-SE block and the token-mixing module, (b) the Squeeze-and-Excitation (SE) module, and (c) the feed-forward module.

However, as the dynamic range of $\tilde{\xi}$ is too large to effectively estimate by DNNs, $\tilde{\xi}$ is mapped into the value in [0, 1] using the cumulative distribution function (CDF) of it modeled as a Gaussian CDF as follows:

$$\bar{\xi}(l,k) = \frac{1}{2}\left[1 + erf\left(\frac{10\log_{10}(\tilde{\xi}(l,k)) - \mu(k)}{\sigma(k)\sqrt{2}}\right)\right], \quad (4)$$

where $erf(\cdot)$ is the error function, and the parameter $\mu(k)$ and $\sigma(k)$ are the mean and the standard deviation computed for the training data, respectively.

Fig. 1 shows the block diagram of the DeepMMSE framework for speech enhancement. The input of the network is the magnitude spectrogram, $|Y(l,k)|$ for all K frequency bins and L frames, which is denoted as $|\mathbf{Y}| \in \mathbb{R}^{K\times L}$. The magnitude spectrogram is encoded by a 1D convolution layer into $\mathbf{X} \in \mathbb{R}^{D\times L}$. The decoder transforms the output of the DNN blocks, $\mathbf{X}' \in \mathbb{R}^{D\times L}$, into the estimates of the mapped *a priori* SNRs $\hat{\bar{\xi}}(l,k)$, in a form $\hat{\bar{\boldsymbol{\xi}}} \in \mathbb{R}^{K\times L}$. $\hat{\bar{\boldsymbol{\xi}}}$ is converted to the estimate of the *a priori* SNR $\hat{\xi}(l,k)$ by using the inverse function of (4) as

$$\hat{\xi}(l,k) = 10^{\{\sigma(k)\sqrt{2}erf^{-1}(2\hat{\bar{\xi}}(l,k)-1)+\mu(k)\}/10}. \quad (5)$$

In the MMSE-LSA estimator framework [29], the estimated speech magnitude $|\hat{S}(l,k)|$ can be expressed as

$$|\hat{S}(l,k)| = G(l,k)\cdot|Y(l,k)|, \quad (6)$$

in which $G(l,k)$ is the gain function

$$G(l,k) = \frac{\hat{\xi}(l,k)}{\hat{\xi}(l,k)+1}\exp\left\{\frac{1}{2}\int_{v(l,k)}^{\infty}\frac{e^{-t}}{t}dt\right\}, \quad (7)$$

$$v(l,k) = \frac{\hat{\xi}(l,k)}{\hat{\xi}(l,k)+1}\hat{\gamma}(l,k), \quad (8)$$

where $\gamma(l,k)$ is the *a posteriori* SNR computed as $\gamma(l,k) = |Y(l,k)|^2/\Phi_S(l,k)$ in [29], but is simply estimated as $\hat{\gamma}(l,k) = \hat{\xi}(l,k)+1$ in [28] and [24].

III. PROPOSED cgMLP-SE MODEL

In this paper, we propose the B repetition of the cgMLP-SE model as the DNN blocks in Fig. 1, which is a gMLP-based structure with three modifications: (i) the convolutional token-mixing module, (ii) the SENet in the token-mixing module, and (iii) the additional feed-forward network. Fig. 2 shows the block diagrams for the proposed cgMLP-SE block and sub-modules. Detailed explanations of each of the modifications are given in the following subsections.

A. CONVOLUTIONAL TOKEN-MIXING MODULE FOR DATA OF VARIOUS LENGTHS

To deal with the input of a variable length, we propose to use the convolutional token-mixing modules in the MLP-mixer [12] and the gMLP [13], which are illustrated in Fig. 3.

1) cMLP-MIXER BLOCK

Figure 3 (a) shows the cMLP-mixer block, a modification of the MLP-mixer block with a convolutional token-mixing module. The input matrix $\mathbf{X} \in \mathbb{R}^{D\times L}$ is first processed by the token-mixing module:

$$\mathbf{Z} = \mathbf{X} + \mathbf{K}_2 \odot GELU(\mathbf{K}_1 \otimes LN(\mathbf{X})) \in \mathbb{R}^{D\times L}, \quad (9)$$

where $LN(\cdot)$ and $GELU(\cdot)$ denote the layer normalization and Gaussian error linear unit (GELU) activation operation [30], respectively. $\otimes$ denotes the depth-wise convolution operator which applies D single convolution filters with a kernel size P to each input channel, while $\odot$ indicates the point-wise convolution operator. The output of the token-mixing module, $\mathbf{Z} \in \mathbb{R}^{D\times L}$, is then processed by the channel-mixing module:

$$\mathbf{X}' = \mathbf{Z} + \mathbf{W}_2 GELU(\mathbf{W}_1 LN(\mathbf{Z})) \in \mathbb{R}^{D\times L}, \quad (10)$$

where $\mathbf{W}_1 \in \mathbb{R}^{H_1\times D}$ and $\mathbf{W}_2 \in \mathbb{R}^{D\times H_1}$ are linear projection matrices with the expanded hidden dimension H_1, and $\mathbf{X}'$ is the output of the cMLP-mixer block.

2) cgMLP BLOCK

Fig. 3 (b) shows the architecture of the cgMLP block, which is a modification of the gMLP block with a convolutional token-

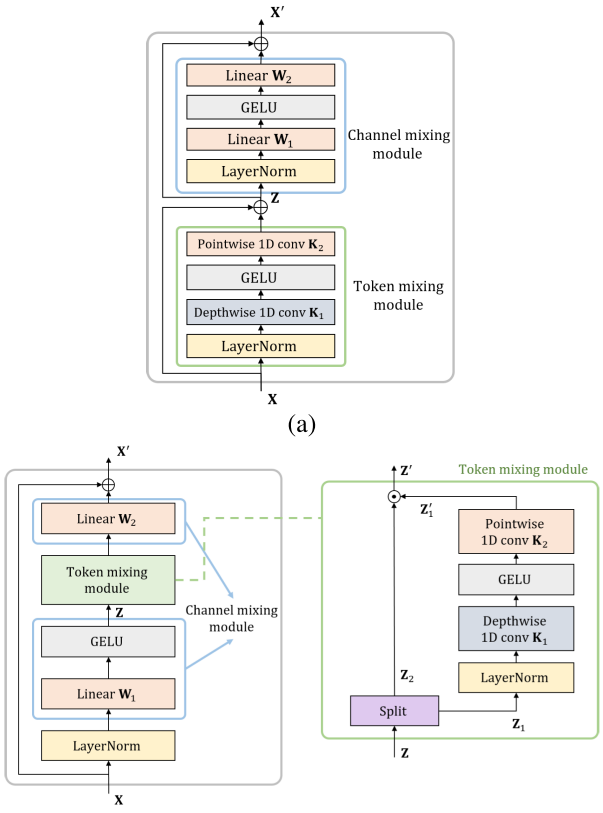

FIGURE 3. MLP-based blocks with the convolutional token-mixing modules: (a) the cMLP-mixer block and (b) the cgMLP block.

mixing module. Unlike the MLP-mixer, the token-mixing module of the gMLP, which is called the spatial gating unit, is located in the middle of the channel-mixing module. The whole process is as follows:

$$\mathbf{Z} = GELU(\mathbf{W}_1 LN(\mathbf{X})) \in \mathbb{R}^{H_1 \times L}, \tag{11}$$

$$\mathbf{Z}' = SGU(\mathbf{Z}) \in \mathbb{R}^{\frac{H_1}{2} \times L}, \tag{12}$$

$$\mathbf{X}' = \mathbf{X} + \mathbf{W}_2 \mathbf{Z}' \in \mathbb{R}^{D \times L}, \tag{13}$$

in which the $SGU(\cdot)$ is the spatial gating unit of the gMLP, $\mathbf{W}_1 \in \mathbb{R}^{H_1 \times D}$ and $\mathbf{W}_2 \in \mathbb{R}^{D \times \frac{H_1}{2}}$ are linear projection matrices and $\mathbf{X}'$ is the output of the cgMLP block. In the spatial gating unit, the input $\mathbf{Z} \in \mathbb{R}^{H_1 \times L}$ is first split into $\mathbf{Z}_1 \in \mathbb{R}^{\frac{H_1}{2} \times L}$ and $\mathbf{Z}_2 \in \mathbb{R}^{\frac{H_1}{2} \times L}$. Then, they are processed as:

$$\mathbf{Z}'_1 = \mathbf{K}_2 \odot GELU(\mathbf{K}_1 \otimes LN(\mathbf{Z}_1)) \in \mathbb{R}^{\frac{H_1}{2} \times L}, \tag{14}$$

$$\mathbf{Z}' = \mathbf{Z}'_1 \circ \mathbf{Z}_2 \in \mathbb{R}^{\frac{H_1}{2} \times L}, \tag{15}$$

where the size of the kernel $\mathbf{K}_1$ applying a depth-wise convolution to each of $\frac{H_1}{2}$ channels is P, and $\circ$ denotes the element-wise multiplication.

Unlike the original MLP-mixer and gMLP, the cMLP-mixer and cgMLP can capture only local contextual information but cannot accommodate long-term contextual information due to the kernel size P.

B. SQUEEZE-AND-EXCITATION MODULE FOR GLOBAL CONTEXTUAL INFORMATION

Recently, the architectures based on the Conformer [7] have demonstrated that utilizing both the global and local contextual information is beneficial for various speech processing tasks [7], [9], [10], [11], [16]. The cgMLP block can only exploit short-term information, and thus we adopt the SENet [19] on top of the token-mixing module of the cgMLP to capture global information without significantly increasing the model size and computational complexity as shown in Fig. 2 (a). Fig. 2 (b) shows the structure of the SE module, which consists of a squeezing step and an excitation step. In the squeezing step, the global average pooling across frames, denoted as $GAP(\cdot)$, is conducted to extract a vector summarizing each channel, and then a linear layer compresses it into a bottleneck vector with a reduced dimension as:

$$\mathbf{Z}_{avg} = GAP(BN(\mathbf{Z}')) \in \mathbb{R}^{\frac{H_1}{2} \times 1}, \tag{16}$$

$$\mathbf{Z}_{sqz} = GELU(\mathbf{W}_{sqz} \mathbf{Z}_{avg}) \in \mathbb{R}^{\frac{H_1}{2r} \times 1}, \tag{17}$$

where $BN(\cdot)$ denotes the batch normalization and $\mathbf{W}_{sqz} \in \mathbb{R}^{\frac{H_1}{2r} \times \frac{H_1}{2}}$ denotes a linear projection matrix with a squeezing ratio r. In the excitation step, the bottleneck vector $\mathbf{Z}_{sqz}$ is expanded by a linear layer $\mathbf{W}_{exc}$ to the vector $\mathbf{Z}_{exc}$ with the original dimension D, and then it is applied in each channel as a gain:

$$\mathbf{Z}_{exc} = SIG(\mathbf{W}_{exc} \mathbf{Z}_{sqz}) \in \mathbb{R}^{\frac{H_1}{2} \times 1}, \tag{18}$$

$$\mathbf{Z}'' = \mathbf{Z}' \circ (\mathbf{Z}_{exc} \mathbf{1}_L^\intercal) \in \mathbb{R}^{\frac{H_1}{2} \times L}, \tag{19}$$

where $SIG(\cdot)$ denotes the sigmoid function, and $\mathbf{W}_{exc} \in \mathbb{R}^{\frac{H_1}{2} \times \frac{H_1}{2r}}$ denotes a linear projection matrix, and $\mathbf{1}_L$ is an all-1 L-dimensional column vector to apply the same weight $\mathbf{Z}_{exc}$ for all columns. For the causal configuration, we adopt causal SENet [31], which employs cumulative global average pooling operation and sine-based fixed positional embedding in the squeezing step.

C. ADDITIONAL FEED-FORWARD MODULE FOR BETTER PERFORMANCE

Additionally, a feed-forward module shown in Fig. 2 (c) is added as the first module of cgMLP-SE to make the cgMLP-SE block similar to the macaron-like structure of the Conformer. The input $\mathbf{X} \in \mathbb{R}^{D \times L}$ is processed by the feed-forward module to produce the input of the next module, $\mathbf{X}'' \in \mathbb{R}^{D \times L}$ as follows:

$$\mathbf{X}'' = \mathbf{X} + \mathbf{W}_4 GELU(\mathbf{W}_3 LN(\mathbf{X})) \in \mathbb{R}^{D \times L}, \tag{20}$$

where $\mathbf{W}_3 \in \mathbb{R}^{H_2 \times D}$ and $\mathbf{W}_4 \in \mathbb{R}^{D \times H_2}$ are linear projection matrices with the hidden layer dimension H_2.

IV. EXPERIMENTS

A. DATASET

In our experiments, we evaluated the performance of speech enhancement on the dataset made by mixing the DNS

TABLE 1. Comparison of the number of parameters and the computational cost of each model for the input of 4 seconds. *M* denotes a million.

Model	# Parameters (M)	MACs (M)	Causal
TCN [24]	1.9	491.3	No
BLSTM	4.4	1104.9	No
DPRNN	4.8	1225.6	No
Conformer	3.5	843.4	No
cMLP-mixer	1.3	304.3	No
cgMLP	1.0	248.4	No
cgMLP-SE	2.1	501.6	No
TCN [24]	1.9	491.3	Yes
LSTM	4.5	1119.1	Yes
DPRNN	4.9	1225.1	Yes
Conformer	3.4	831.4	Yes
cMLP-mixer	1.2	301.2	Yes
cgMLP	1.0	240.8	Yes
cgMLP-SE	2.1	530.4	Yes

TABLE 2. Speech enhancement performance using the DeepMMSE framework for various models on the TIMIT-DNS noise dataset.

Model	PESQ	STOI	CSIG	CBAK	COVL	Causal
Noisy	1.35	0.79	2.46	1.99	1.84	-
TCN	2.01	0.85	3.40	2.80	2.67	No
BLSTM	2.01	0.86	3.39	2.79	2.67	No
DPRNN	2.06	0.86	3.46	2.89	2.74	No
Conformer	**2.18**	**0.87**	**3.61**	**2.96**	**2.87**	No
cMLP-mixer	2.06	0.86	3.47	2.86	2.74	No
cgMLP	2.05	0.86	3.45	2.84	2.72	No
cgMLP-SE	2.17	**0.87**	3.59	**2.96**	2.86	No
TCN	1.87	0.83	3.22	2.66	2.50	Yes
LSTM	1.90	0.84	3.28	2.72	2.55	Yes
DPRNN	1.97	0.85	3.35	2.78	2.63	Yes
Conformer	**2.03**	**0.86**	**3.45**	**2.85**	**2.72**	Yes
cMLP-mixer	1.94	0.84	3.34	2.76	2.61	Yes
cgMLP	1.94	0.84	3.36	2.75	2.61	Yes
cgMLP-SE	**2.03**	0.85	3.44	**2.85**	2.71	Yes

challenge noise set [32] with the TIMIT speech data [33]. In this dataset we call the TIMIT-DNS noise dataset, the training set was constructed dynamically for each epoch by mixing each of 58,801 noise data from the DNS challenge noise set [32] with one of the 4,620 utterances spoken by 462 speakers from the TIMIT training set at a random signal-to-noise ratio (SNR) between -10 dB to 20 dB. The validation set was composed of 1,480 utterances of speech from the TIMIT test set mixed with noises from the DNS challenge noise set which were not used for training at the SNRs from -10 dB to 20 dB with 5 dB intervals. The evaluation set was constructed by adding unused DNS challenge noise data to the rest of 100 male and 100 female utterances from the TIMIT test set at the SNRs from -5 dB to 15 dB with 5 dB intervals.

Additionally, the experiments on the Voice Bank-DEMAND dataset [27] which is a widely used benchmark were carried out. The Voice Bank-DEMAND training set consists of 11,572 clean speech recordings from 28 speakers mixed with two artificial and eight real noises. Each clean speech utterance was contaminated by a randomly selected section of one of the noises with a random SNR out of {0, 5, 10, 15} dB. The Voice Bank-DEMAND test set includes 824 clean speech recordings from 2 speakers from the Voice Bank database [34] mixed with one of the 5 types of noise from the DEMAND noise database [35] at SNR levels among {2.5, 7.5, 12.5, 17.5} dB.

TABLE 3. Speech enhancement performance using the DeepMMSE framework for various models on the Voice Bank-DEMAND noise dataset.

Model	PESQ	STOI	CSIG	CBAK	COVL	Causal
Noisy	1.97	0.82	3.36	2.44	2.64	-
TCN [24]	2.95	0.94	4.28	3.46	**3.64**	No
BLSTM	2.92	**0.95**	4.29	**3.47**	3.62	No
DPRNN	2.90	**0.95**	4.26	3.45	3.60	No
Conformer	**2.96**	0.94	**4.30**	3.43	**3.64**	No
cMLP-mixer	2.92	**0.95**	4.28	3.46	3.61	No
cgMLP	2.91	0.94	4.26	**3.47**	3.60	No
cgMLP-SE	**2.96**	**0.95**	4.28	3.46	**3.64**	No
TCN [24]	2.77	0.93	4.14	3.32	3.46	Yes
LSTM	2.83	**0.94**	4.17	3.31	3.50	Yes
DPRNN	2.81	**0.94**	4.15	3.40	3.50	Yes
Conformer	**2.84**	**0.94**	**4.21**	3.37	**3.54**	Yes
cMLP-mixer	2.81	**0.94**	4.19	3.37	3.51	Yes
cgMLP	2.80	**0.94**	4.17	3.39	3.49	Yes
cgMLP-SE	2.83	**0.94**	4.20	**3.41**	3.53	Yes

B. EXPERIMENTAL SETUP

In order to demonstrate the performance of the proposed cgMLP-SE model, we compared the performances with different configurations of the DNNs in the DeepMMSE framework. Specifically, we have compared the proposed model with the TCN [24], LSTM [25], DPRNN [26], and Conformer [7]. In the noncausal configuration, BLSTM had 3 layers with 256 hidden units, DPRNN had 2 layers with 256 hidden units, and Conformer had 4 blocks with the attention dimension of 256, attention head of 4, hidden units of 512, and kernel size of 33. In the causal configuration, LSTM had 2 layers with 512 hidden units, DPRNN had 4 layers with 256 hidden units, and Conformer had 4 blocks with the attention dimension of 256, attention head of 4, hidden units of 512, and kernel size of 17. The TCN had the same architecture as described in [24] in both noncausal and causal configurations. The parameter values of the MLP-based models were $B = 4, K = 257, D = 256, H_1 = H_2 = 512$, $P = 33$ for noncausal configuration and $P = 17$ for causal configuration, and $r = 4$. Table 1 shows the number of parameters and the number of multiply-accumulate operations (MACs) in million for compared models used as DNN blocks in Fig. 1. The computational complexity was measured by using the *ptflops counter* [36]. With the configuration parameters used in the experiments, the proposed cgMLP-SE model had the number of parameters and MACs which were twice the cgMLP and comparable to TCN. The computational complexities of the LSTM, DPRNN, and Conformer models were significantly higher than the proposed model.

The sampling rate was 16 kHz, and the window was 512-point Hann window with a 50% overlap. The 512-point STFT and inverse STFT were applied for analysis and synthesis, respectively.

For the training step, we used the adam optimizer [37] with the learning rate $1e^{-3}$ and the binary cross entropy as a loss function. The gradients were clipped to $[-1, 1]$ range. The number of epochs was 300 and the batch size was 32.

C. EXPERIMENTAL RESULTS

The performance of the speech enhancement was measured by the ITU-T Recommendation P.862.2 wideband Perceptual Evaluation of Speech Quality (PESQ) [38] scores,

TABLE 4. Ablation study on the contribution of each module in the cgMLP-SE blocks to the performance improvement for the TIMIT-DNS noise dataset. SE module and FF module denote the squeeze-and-excitation module and the additional feed-forward module, respectively.

Model	# Parameters (M)	PESQ	STOI	CSIG	CBAK	COVL	Causal
Noisy	-	1.35	0.79	2.46	1.99	1.84	-
cgMLP	1.0	2.05	0.86	3.45	2.84	2.72	No
cgMLP with SE module	1.1	2.12	**0.87**	3.49	2.90	2.77	No
cgMLP with FF module	2.0	2.13	**0.87**	3.57	2.94	2.83	No
cgMLP-SE	2.1	**2.17**	**0.87**	**3.59**	**2.96**	**2.86**	No
cgMLP	1.0	1.94	0.84	3.36	2.75	2.61	Yes
cgMLP with SE module	1.1	1.96	**0.85**	3.38	2.75	2.62	Yes
cgMLP with FF module	2.0	1.99	**0.85**	3.40	2.82	2.67	Yes
cgMLP-SE	2.1	**2.03**	**0.85**	**3.44**	**2.85**	**2.71**	Yes

Short-Time Objective Intelligibility (STOI) [39], and composite measures for signal distortion (CSIG), residual noise (CBAK), and overall quality (COVL) [40]. Table 2 shows the performance of the DeepMMSE-based speech enhancement employing various models with causal and noncausal configurations on the TIMIT-DNS noise dataset. For both the causal and noncausal versions, TCN and (B)LSTM showed similar performance and DPRNN exhibited better performance. The Conformer showed the best performance among the existing four models, especially in the noncausal configuration with a large margin. The cMLP-mixer and cgMLP introduced in subsection III.A which adopt the convolutional token-mixing module to deal with inputs of various lengths demonstrated similar or slightly worse performance compared with the DPRNN. The proposed cgMLP-SE model showed similar performance to the Conformer in all measures, although it had 1.3M fewer parameters and 340 M less number of MACs. We may conclude that the proposed model provided a good compromise between computational complexity and performance.

Table 3 shows the results for a smaller but widely-used dataset, the Voice Bank-DEMAND dataset. The tendency was similar but slightly different in that the noncausal version of the TCN showed decent performance and LSTM performed better than DPRNN. Still, the Conformer exhibited the best performance and the proposed cgMLP-SE model showed comparable performance with a significantly smaller model size and less computation.

D. ABLATION STUDY

We performed an ablation study to investigate how much each module in the proposed cgMLP-SE model contributed to performance improvement on the TIMIT-DNS noise dataset. As the input signals had various lengths, the contribution of the convolutional token-mixing module was not evaluated but those for the SE module and the additional feed-forward module were analyzed. Table 4 shows the performance of the DeepMMSE speech enhancement using the cgMLP model with and without the SE module and the feed-forward module. We can see that almost all the measures were improved by employing the SE module or the feed-forward module, and further enhanced by adopting both of them. From the results, we can confirm that combining the SE module and the additional feed-forward module was effective to improve both the objective speech quality and intelligibility measures, and one can choose to use the cgMLP with the SE module without the feed-forward module if the model size and computational complexity need to be further reduced.

V. CONCLUSION

In this paper, we propose the cgMLP-SE block for speech enhancement, which is an efficient MLP-based architecture that can deal with the inputs of various lengths with reasonable computational complexity and model size. The cgMLP-SE block is the gMLP-based structure with three modifications: (i) a convolutional token-mixing module to process input data of various lengths, (ii) an SE module to consider global contextual information, and (iii) an additional feed-forward module to boost performance. We evaluated the DeepMMSE speech enhancement using the proposed cgMLP-SE and other architectures on the TIMIT-DNS noise dataset and the Voice Bank-DEMAND dataset in terms of speech quality and intelligibility. Experimental results showed that the proposed cgMLP-SE model demonstrated comparable PESQ scores, STOI, and composite measures to the Conformer with significantly smaller model size and computational complexity.

REFERENCES

[1] K. Tan, X. Zhang, and D. Wang, "Real-time speech enhancement for mobile communication based on dual-channel complex spectral mapping," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6134–6138.
[2] C.-Y. Li and N. T. Vu, "Improving speech recognition on noisy speech via speech enhancement with multi-discriminators CycleGAN," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2021, pp. 830–836.
[3] H. Zhou, J. Du, Y.-H. Tu, and C.-H. Lee, "Using speech enhancement preprocessing for speech emotion recognition in realistic noisy conditions," in *Proc. Interspeech*, Oct. 2020, pp. 4098–4102.
[4] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.
[5] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.
[6] H. Kim and J. W. Shin, "Target exaggeration for deep learning-based speech enhancement," *Digit. Signal Process.*, vol. 116, Sep. 2021, Art. no. 103109.
[7] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020, *arXiv:2005.08100*.
[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–15.
[9] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H.-Y. Lee, and H. Meng, "MFA-Conformer: Multi-scale feature aggregation conformer for automatic speaker verification," 2022, *arXiv:2203.15249*.

[10] E. Kim and H. Seo, "SE-Conformer: Time-domain speech enhancement using conformer," in *Proc. Interspeech*, Aug. 2021, pp. 2736–2740.
[11] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, "Continuous speech separation with conformer," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 5749–5753.
[12] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "MLP-Mixer: An all-MLP architecture for vision," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–12.
[13] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to MLPs," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 9204–9215.
[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[15] J. Sakuma, T. Komatsu, and R. Scheibler, "MLP-ASR: Sequence-length agnostic all-MLP architectures for speech recognition," 2022, *arXiv:2202.08456*.
[16] B. Han, Z. Chen, B. Liu, and Y. Qian, "MLP-SVNET: A multi-layer perceptrons based network for speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7522–7526.
[17] J. Tae, H. Kim, and Y. Lee, "MLP singer: Towards rapid parallel Korean singing voice synthesis," in *Proc. IEEE 31st Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Oct. 2021, pp. 1–6.
[18] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11030–11039.
[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[20] W. Gharbieh, J. Huang, Q. Wan, H. S. Shim, and H. C. Lee, "DyConvMixer: Dynamic convolution mixer architecture for open-vocabulary keyword spotting," in *Proc. Interspeech*, Sep. 2022, pp. 5205–5209.
[21] S. Han, J. Byun, and J. W. Shin, "Time-domain speaker verification using temporal convolutional networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6688–6692.
[22] N. Kanda, G. Ye, Y. Wu, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "Large-scale pre-training of end-to-end multi-talker ASR for meeting transcription with single distant microphone," 2021, *arXiv:2103.16776*.
[23] N. Kanda, G. Ye, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "End-to-end speaker-attributed ASR with transformer," 2021, *arXiv:2104.02128*.
[24] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1404–1415, 2020.
[25] K. Greff, R. K. Srivastava, J. Koutnìk, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space Odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.
[26] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 46–50.
[27] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proc. 9th ISCA Workshop Speech Synth. Workshop*, Sep. 2016, pp. 146–152.
[28] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Commun.*, vol. 111, pp. 44–55, Aug. 2019.
[29] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.
[30] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.
[31] D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, and B. Gong, "MoViNets: Mobile video networks for efficient video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16020–16030.
[32] C. K. A. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Interspeech 2021 deep noise suppression challenge," 2021, *arXiv:2101.01902*.
[33] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *STIN*, vol. 93, p. 27403, Feb. 1993.
[34] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. Int. Conf. Oriental COCOSDA Conf. Asian Spoken Lang. Res. Eval. (O-COCOSDA/CASLRE)*, Nov. 2013, pp. 1–4.
[35] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Meetings Acoust.*, 2013, Art. no. 035081.
[36] V. Sovrasov. *Flops Counter for Convolutional Networks in PyTorch Framework*, 2021. [Online]. Available: https://github.com/sovrasov/flops-counter.pytorch
[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[38] *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codec*, document P.862.2, ITU Recommendation, 2007.
[39] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[40] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

HYUNGCHAN SONG (Graduate Student Member, IEEE) received the B.S. degree in electronic and control engineering from the Hanbat National Institute of Technology, Daejeon, South Korea, in 2016, and the M.S. degree from the School of Electrical Engineering and Computer Science (EECS), Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 2018, where he is currently pursuing the Ph.D. degree. His research interests include speech enhancement, speech source localization, acoustic echo cancellation, speech representation, and machine learning.

MINSEUNG KIM (Graduate Student Member, IEEE) received the B.S. and M.S. degrees in electrical engineering and computer science from the Gwangju Institute of Science and Technology, Gwangju, South Korea, in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree. His research interests include speech enhancement, acoustic echo cancellation, and voice activity detection.

JONG WON SHIN (Member, IEEE) received the B.S. and Ph.D. degrees from the School of Electrical Engineering and Computer Science (EECS), Seoul National University, Seoul, South Korea, in 2002 and 2008, respectively. From 2008 to 2012, he was with Qualcomm Inc., San Diego, CA, USA. Since 2012, he has been with EECS, Gwangju Institute of Science and Technology, Gwangju, South Korea, where he is currently an Associate Professor. His research interests include vast areas of speech signal processing, such as speech enhancement, voice activity detection, source localization, acoustic echo cancellation, and speech emotion recognition.