An Integrated CNN-GRU Framework for Complex Ratio Mask Estimation in Speech Enhancement

Mojtaba Hasannezhad*, Zhiheng Ouyang*, Wei-Ping Zhu*, and Benoit Champagne†
* Concordia University, Montreal, Canada
E-mail: m_hasann/z_ouyan@encs.concardia.ca, weiping@ece.concordia.ca
† McGill University, Montreal, Canad
E-mail: benoit.champagne@mcgill.ca

***Abstract*—In this paper, we propose a novel neural network-based speech enhancement approach, where a convolutional neural network (CNN) and a gated recurrent unit (GRU) are integrated to estimate a modified complex ratio mask (MCRM.) The new CNN structure comprised of frequency dilated convolution layers is employed to extract speech features while benefiting from the global contextual information of input speech. The CNN incorporates the skip connection and residual learning techniques to facilitate the training and accelerate the convergence. The GRU network is exploited to map the CNN-extracted features to the MCRM, which is used to enhance both magnitude and phase of the input speech. We compare the enhancement performance of the proposed method using features extracted by CNN with that of the GRU network using some conventional acoustic features, showing the advantage of the proposed CNN-GRU model. We also demonstrate that the GRU outperforms other recurrent neural network variations within the proposed model for mask estimation in terms of separated speech quality, memory footprint, and the number of model parameters in the presence of highly non-stationary noises.**

Index Terms: speech enhancement, masking-based techniques, GRU, frequency dilated CNN.

I. Introduction

One of the most challenging topics in the speech processing area is that of speech enhancement which aims to reconstruct a clean speech from measurements that were corrupted by ambient noise. Speech enhancement is of crucial importance to speech recognition based applications such as smart home devices like Alexa and Google Home [1].

The advent of deep learning has very much advanced the evolutionary progress in data-driven approaches for speech enhancement. Xu *et al.* [2] introduced a deep neural network (DNN) based method where DNN learns to map the noisy speech log power spectrum to the clean one, and reported promising results in terms of speech quality and intelligibility. This method requires a large training set to form an accurate mapping [3]. Another DNN-based approach was proposed in [4] for ideal binary mask (IBM) estimation, resulting in significant improvement in speech enhancement. Wang *et al.* [5] performed a comparative study of different targets for a fully-connected (FC) neural network that is trained to map a complementary set of acoustic features to different targets. They employed the neural network to estimate both a spectral mask like ideal ratio mask (IRM) and the short-time Fourier transform (STFT) spectral magnitude, and concluded that estimating a spectral mask leads to better results in terms of objective intelligibility and quality metrics. Since phase processing is usually ignored in the aforementioned methods, while it directly impacts quality of the results, as demonstrated in [6], William *et al.* [7] estimate a complex ideal ratio mask (cIRM) by an FC network to enhance the speech phase alongside the magnitude. Although the clean speech spectrogram can be synthesized accurately using cIRM, neural network surprisingly fails to estimate the imaginary part of the mask [8].

Most of the DNN-based speech enhancement methods utilize an FC network which, in spite of its large number of parameters, processes input samples independently. However, better performance in speech enhancement can be obtained if the strong temporal dependencies of speech are considered. Hence, an recurrent neural network (RNN) using feed-back connections between hidden layers, which treats the input signal as a temporal sequence has emerged as a natural choice to model dynamics of speech [1]. As such, an long short-term memory (LSTM) network, which avoids vanishing and exploding gradient problems of RNN, is employed in [9] to estimate IRM considering the temporal contextual information of speech.

Besides, the network inputs play a key role in deep learning-based speech enhancement. Conventional acoustic-phonetic features as the network input are appropriate only for specific applications and come with their own limitations. In contrast, the convolutional neural network (CNN) automatically extract the sophisticated features of the input signal. It has already been successfully employed for feature extraction in speech recognition [10] and acoustic scene classification [11]. However, a traditional CNN comprising pairs of convolution and sub-sampling (pooling) layers has its own limitations. In particular, the pooling layer reduces the resolution and sensitivity to local variations due to the sub-sampling [1]. So-called max pooling only keeps rough information of the input spectrogram, while average pooling neglects important local structures by attenuating the contribution of the individual units in the local region [12]. Moreover, as the speech spectrogram dimension along the frequency axis in on the order of a few hundreds, the limited receptive field of the convolution layers results tends to destroy the global contextual information of speech [13]. To expand the receptive

field of a convolution layer, it is a common practice to enlarge the size of CNN kernels which, however, inevitably leads to higher complexity and reduces processing speed. An alternative approach is provided by stride convolution, which shrinks the size of kernels and introduces translation invariance; however, it overly flattens the prediction of mask units and lessens the model accuracy, subsequently leading to a loss of frequency resolution reduction [14], [15]. To tackle these problems, a new CNN structure with one-dimensional (1D) convolution and frequency-dilated 2D convolution was introduced in our previous work [13] to enlarge the receptive field while keeping the kernel size small.

In this paper, a novel integrated CNN-GRU framework is proposed for a modified complex ratio mask estimation (MCRM). We employ a new low-complexity CNN structure that exploits speech spectrogram global correlations to extract the most suitable features from the input speech. In our CNN, the receptive field is expanded without enlarging the kernel size by stacking frequency dilated convolution layers. The residual learning and skip-connection techniques are also employed to ease training and accelerate convergence. The CNN-extracted features are then forwarded to the GRU network to predict MCRM so as to enhance both the magnitude and phase of the speech. It is worth mentioning that this model takes advantage of both the spectral and temporal contextual information of speech due to the combined use of CNN and GRU. Through extensive experiments, we demonstrate that our proposed model not only yields better enhancement performance but also enjoys a lower complexity as compared to other DNN-based methods.

II. Proposed Framework

The proposed model comprises two main stages as shown in Fig. 1. Firstly, the short-time Fourier transform (STFT) of the input speech signal is computed and input to the CNN with frequency dilated convolution to extract features of the input signal. Then, the GRU network, focusing on the temporal information of speech, maps these features to real and imaginary components of the MCRM. The estimated mask is then used to refine the STFT of the input noisy speech.

A. Complex Spectrogram and Modified Complex Ratio Mask

In the time domain, the noisy speech signal $y(t)$ is modeled as a summation of a clean speech $x(t)$ and noise signal $n(t)$, where t denotes the discrete-time index. Let $X(k,l)$ denote the complex spectrogram of the clean speech, as computed via the STFT of the framed signal with k and l denoting time frame and frequency bin indices, respectively. The complex spectrogram can be expressed in terms of its real and imaginary components, $X(k,l) = X_r(k,l) + iX_i(k,l)$, where the subscripts r and i stands for real and imaginary components, respectively.

Since these components together bear the information of both speech magnitude and phase, they can be considered as the training target together. A key benefit of these components is that due to their similarity, they allow employing a single neural network to predict both components together [13]. Besides, since estimating a mask is more efficient than directly estimating a spectrogram, a complex IRM can be used to calculate the complex spectrogram of the clean speech from the noisy one, i.e., $X = M \times Y$, neglecting (k,l) for brevity. Thus, the real and imaginary components of the mask can be derived as follows,

$$M = \frac{Y_r X_r + Y_i X_i}{{Y_i}^2 + {Y_r}^2} + i\frac{Y_r X_i - Y_i X_r}{{Y_i}^2 + {Y_r}^2} \tag{1}$$

Therefore, this complex ratio mask can be considered as the network training target [7].

However, DNN cannot properly estimate the imaginary part of this mask as shown in [8]. Interestingly, it is found from our extensive simulations that the energy of the imaginary part of the mask is consistently lower than that of the real part, which makes DNN biased to the real part, i.e., the real part of the complex mask is emphasized. In order to compensate this imbalance, we boost the imaginary part, i.e. $M = M_r + j(k + M_i)$. We refer to the resulting mask as the Modified Complex Ratio Mask (MCRM.) Moreover, since the values of the real and imaginary components have a wide range, they should be compressed for application to neural networks. We perform this compression with a hyperbolic tangent which gives the best results in comparison with other compression methods. After this compression, k will be added to the imaginary part of the mask. Different values for k are evaluated and $k = 0.5$ performs best empirically. It is worth mentioning that k will be deducted from the imaginary part of the estimated mask in the testing stage.

B. Feature Extraction by Frequency Dilated Convolution CNN

As shown in Fig. 1, features extraction from the STFT of input noisy speech is carried out using a CNN with frequency dilated convolution, wherein a few input values are skipped at each step. This CNN is made up of four stacked frequency dilated convolution layers with increasing dilation rates of $1, 2, 4,$ and 8 to exponentially expand the receptive field of convolutional layers while keeping the kernel size small. The number of channels of these stacked layers is $16, 32, 16,$ and 8, respectively, and the ReLU activation function is employed.

Since our goal is to enlarge the receptive field along the frequency axis, 1D convolutional layers with a kernel size of 1×7 are used. This CNN structure benefits from both skip connection and residual learning, which facilitate the learning and speed up the convergence. To implement the residual learning, the input passes through both a 1D convolutional layer and an identity mapping layer composed of 1×1 kernels, with their sum input to the next layer. Subsequently, the sum of the outputs of the frequency dilated convolutional layers is adjusted by an identity mapping layer, and then reshaped to be forwarded to the GRU stage for the final regression. It is worth mentioning that, instead of summation, we also tried to stack the outputs; however, the results were not as good as the summation.

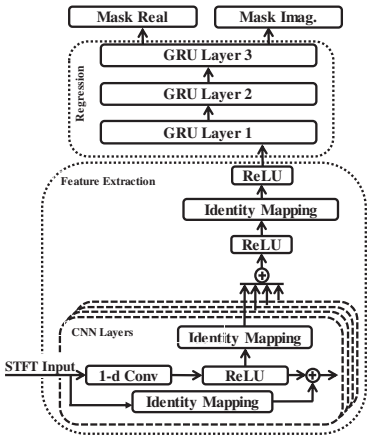

Fig. 1: Proposed integrated CNN-GRU framework

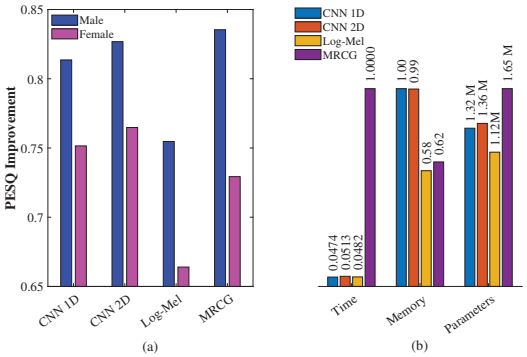

Fig. 2: Feature comparison: (a) Average PESQ score improvement; (b) Comparison of computational time, memory, and number of parameters in million.

C. GRU Network for Regression

GRU is considered as an efficient implementation of LSTM since it uses merely two gates instead of three gates and a memory cell. Despite the simplified implementation of GRU, its performance is sometimes better than LSTM even with less training data [16]. Here, we employ a GRU network with three hidden layers for the sake of both reducing network complexity and enhancing performance, with each layer comprising 256 units to model temporal dynamics of speech. To avoid over-fitting and make the network robust to unseen acoustic scenarios, dropout at the rate of 0.3 is adopted. We point out that since the volume of training data is very large in comparison with the number of model parameters, the network only learns the fundamental information of training data and therefore never encounters over-fitting. Finally, an affine fully-connected layer transfers the GRU layer outputs to the real and imaginary components of the ratio mask.

III. Experiments

A. Experimental Setup

In order to evaluate the performance of the proposed model in comparison with some of the existing methods, the TIMIT dataset [17] is used, which comprises 6300 utterances spoken by 630 male and female speakers with different dialects, each speaker uttering 10 phonetically-rich sentences. Four highly non-stationary noises from the NOISEX-92 dataset [18], namely, babble, restaurant, factory, and street, are selected and each is divided into two parts for training and testing stage, so that the noise is unseen during the evaluation stage. Random sections from the first part of each noise signal are mixed with the clean speech utterances at different SNR levels of -5, 0, 5, and 10 dB. In total, the network is trained with more than 10^5 mixtures (i.e. 6300 utterances $\times 4$ SNR levels $\times 4$ noise types). In the testing stage, 60 unmatched clean utterances are mixed with random sections from the second part of each noise at unmatched SNR levels of -6, 0, 6, and 12 dB which will yield a total of 960 mixtures. The input signals are sectioned into 20 ms frames using the Hanning window with 10 ms (50%) overlap. The sampling rate is set at $16kHz$ and a 320-point DFT is computed where each frame consists of 160 samples. As the cost function for training the CNN-GRU network, the minimum mean square error (MMSE) is used with Adam optimizer, which provides an extension to stochastic gradient descent [19]. Two common performance metrics, i.e., perceptual evaluation of speech quality (PESQ) and segmental signal-to-noise ratio (SSNR) [20], are used to evaluate the enhanced speech.

B. Feature Extraction

To compare the performance of the new model, which integrates CNN-extracted features with a GRU network using fixed feature sets, we use Mel-filterbank energy (Log-Mel) features as spectrum-based features and multi-resolution cochleagram (MRCG) [21] as high-quality Gammatone-domain features. These static features are concatenated with their delta and acceleration, resulting in a total dimension of 78 (i.e., 26 static + 2×26 dynamic) for the Log-Mel and 768 (256 static + 2×256 dynamic parameters) for the MRCG. The feature vectors are normalized to zero mean and unit variance to bring the dynamic and static parts to the same range [21]. Besides, the CNN using 1D and 2D convolutions with kernel sizes of 1×7 and 3×7, respectively, is employed to extract the most appropriate features of the noisy speech signal at the input.

The results of the comparison are shown in Fig. 2. It is seen that log-Mel results in the lowest PESQ score while achieving the lowest cost in terms of computational time, memory, and number of parameters. MRCG leads to good results in terms of speech quality at the price of higher computational time and number of parameters. While the MRCG features benefit from both local and contextual information of cochleagrams, the good results are also attributed to the feature set dimension, which is about ten times that of log-Mel. Notably, the use of 1D and 2D CNN for feature extraction results in very good speech quality, with a lower complexity than MRCG. Between 1D and 2D CNN, the obtained PESQ scores are

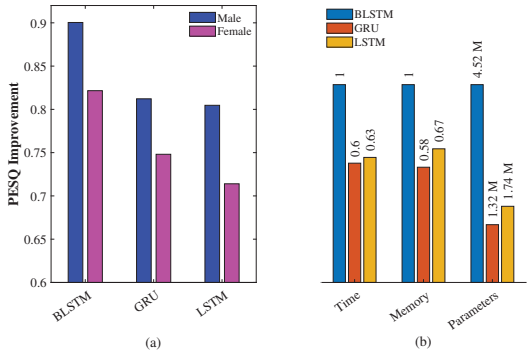

Fig. 3: Comparison among different RNNs: (a) Average PESQ score improvement; (b) Computational time, memory, and number of parameters in million.

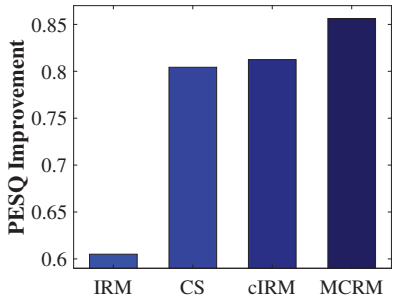

Fig. 4: Evaluation of different training targets with CNN-GRU model.

comparable while 1D CNN is more efficient in terms in terms of computational time and the number of parameters. Hence, 1D CNN offers a good trade-off when considering the performance and complexity together.

C. Comparison of RNN Types

To overcome the expanding and vanishing problems of RNN, LSTM is introduced with a memory cell and gates to facilitate information flow over time. Afterward, GRU as a more efficient form of LSTM is introduced wherein two LSTM gates are combined with no memory cell, which means that it exposes full hidden content without any control [3]. As such, GRU is more computationally efficient than LSTM. To take full advantage of input information, bidirectional RNNs are then introduced where the backward and forward hidden states are concatenated and connected to each output node to predict the output sequence using the preceding and following information.

Enhancement performances using GRU, LSTM, and BLSTM networks as variations of RNN for mask estimation are compared in terms of the quality of results, computational time, memory, and the number of parameters. All the networks are trained and tested with the same configuration and consist of three hidden layers each having 256 units. The results of these comparisons for males and females are shown in Fig. 3 where computational memory and time are normalized to 1 for the sake of simplicity. As shown in the figure, BLSTM outperforms LSTM and GRU as expected. The reason is that BLSTM takes advantage of past and future information of the input speech signal. Also, the high quality of the BLSTM results is associated with the high number of parameters which enables BLSTM to learn more detailed information form training data. However, BLSTM requires almost twice as much computational time and memory to achieve such a better quality of results. We also note that, GRU slightly outperforms LSTM terms of the quality of the results while requiring less computational time and complexity.

D. Comparison of Training Targets

As already mentioned, various training targets have been introduced in the literature, most of them aiming only at magnitude enhancement. To process phase alongside magnitude enhancement, on the one hand, direct estimation of the complex spectrogram (CS) is proposed in [13], but estimating the real and imaginary parts of a spectrogram is challenging as the network has to predict every single element of the spectrogram over the time-frequency (TF) domains. On the other hand, it is possible for the network to estimate real and imaginary parts of the cIRM [7], which deals with a subset of TF cells as described in Section II-A, but this type of approach fails to estimate the imaginary part. Hence, to overcome these limitations, we propose MCRM where the cumulative energy level of the real and imaginary parts of cIRM are equalized. Fig. 4 shows the average PESQ score improvement where the aforementioned training targets are estimated with the CNN-GRU model. We also evaluated IRM to show the advantage of phase besides magnitude enhancement. As seen, training targets that consider phase enhancement significantly outperform IRM which shows the importance of phase processing along with magnitude enhancement. Moreover, the estimated MCRM by the CNN-GRU model leads to the best PESQ score among other aforementioned training targets.

TABLE I: Average PESQ and SSNR scores of different models

Method	PESQ				SSNR				No. of Parameters
	-6	0	6	12	-6	0	6	12	
Unprocessed	1.21	1.63	2.10	2.56	-9.74	-5.06	0.49	6.33	-
	0.91	1.37	1.87	2.36	-9.17	-4.86	0.51	6.27	
FFT-Mag	1.73	2.18	2.51	2.73	0.45	1.96	3.13	3.99	2.66M
	1.26	1.69	1.99	2.16	0.57	1.79	2.94	3.64	
TMS	1.70	2.14	2.55	2.89	0.69	2.39	4.18	5.69	12.35M
	1.46	1.90	2.34	2.70	1.18	2.83	4.49	5.77	
IRM	1.83	2.38	2.92	3.39	-0.75	3.24	**7.04**	**10.06**	2.66M
	1.42	1.99	2.58	3.13	-0.46	3.53	**7.55**	**10.81**	
SMM	1.82	2.25	2.65	3.05	-1.22	1.83	5.10	9.09	2.66M
	1.42	1.89	2.34	2.77	-0.63	2.36	5.76	9.44	
cIRM	1.94	2.43	2.92	3.34	1.06	3.70	6.41	9.09	2.82M
	1.61	2.13	2.61	3.05	1.13	3.67	6.43	8.97	
Proposed	**2.01**	**2.50**	**2.99**	**3.44**	**1.44**	**4.06**	6.73	9.59	**1.32M**
	1.83	**2.29**	**2.77**	**3.20**	**1.96**	**4.35**	6.93	9.59	

E. Comparison with Existing DNN-Based Methods

The proposed CNN-GRU framework is also compared with some other DNN-based methods. The FFT-MAG and target magnitude spectrum (TMS) methods are introduced in [2] and [5], respectively. Both are implemented with an FC network made up of three hidden layers each having 1024 and 2048 units per layer for TMS and FFT-MAG, respectively. The former concatenates a complementary set of features vectors from 5 signal frames as network input, while the latter utilizes the log-power spectral magnitudes of 11 frames. Both techniques use the noisy phase to restore the clean speech. Besides, spectral magnitude mask (SMM), IRM [5], and cIRM [7] are used to estimate a spectral mask, where each method employs a three-hidden layer FC networks. Each layer comprises 1024 units, and the concatenation of features from 5 contiguous frames is used to capture the contextual information of the input signal. The comparison results are shown in Table I, wherein the upper and lower numbers in each cell are for males and females, respectively. In terms of PESQ, the proposed framework outperforms the other methods at all SNR levels, while in terms of SSNR, it gives better results at lower SNR values of -6 and 0 dB while IRM apparently works better at 6 and 12 dB SNR. Nevertheless, the number of model parameters in the proposed framework is less than than for the other methods, as seen from the rightmost column of the table.

IV. Conclusion

In this paper, we have proposed a joint CNN and GRU network for MCRM estimation in which a new low-complexity CNN structure with frequency dilated convolution layers adopting skip connection and residual learning has been designed to automatically learn the features of the input noisy speech. Through extensive experimental and comparison studies, we found that GRU offered the best trade-off among RNN variations to obtain the mask estimate. In particular, we showed that the CNN-extracted features surpass manual acoustic features for speech enhancement. We also showed that the use of the proposed MCRM to handle both phase and magnitude of the noisy speech has can significantly improved the quality of the separated speech. Overall, our results showed that the proposed CNN-GRU model for can outperform competing approaches for mask estimation terms of enhanced speech quality under adverse noise conditions as well as implementation complexity, i.e., memory footprint, number of model parameters and training time.

References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Tran. on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Tran. on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[3] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[4] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE/ACM Tran. on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[5] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Tran. on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[6] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Tran. on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1931–1940, 2014.

[7] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Tran. on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2015.

[8] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "PHASEN: A phase-and-harmonics-aware speech enhancement network." in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020, pp. 9458–9465.

[9] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

[10] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *Int. Conf. on Machine Learning*, 2016, pp. 173–182.

[11] J. Guo, N. Xu, L.-J. Li, and A. Alwan, "Attention based CLDNNs for short-duration acoustic scene classification." in *INTERSPEECH*, 2017, pp. 469–473.

[12] H. Zhang and J. Ma, "Hartley spectral pooling for deep learning," *arXiv preprint arXiv:1810.04028*, 2018.

[13] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, "A fully convolutional neural network for complex spectrogram processing in speech enhancement," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5756–5760.

[14] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[15] J. Lei, C. Wang, B. Zhu, Q. Lv, Z. Huang, and Y. Peng, "Multi-LCNN: a hybrid neural network based on integrated time-frequency characteristics for acoustic scene classification," in *Int. Conf. on Tools with Artificial Intelligence*, 2018, pp. 52–59.

[16] R. Dey and F. M. Salemt, "Gate-variants of gated recurrent unit (GRU) neural networks," in *IEEE Int. Midwest Symposium on Circuits and Systems (MWSCAS)*, 2017, pp. 1597–1600.

[17] 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," *NASA STI/Recon technical report n*, vol. 93, 1993.

[18] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] J. H. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Fifth Int. Conf. on spoken language processing*, 1998.

[21] M. Delfarah and D. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Tran. on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1085–1094, 2017.