Speech Separation Using a Composite Model for Complex Mask Estimation

Mojtaba Hasannezhad*, Zhiheng Ouyang*, Wei-Ping Zhu*, and Benoit Champagne†
Email: {m_hasann/z_ouyan}@encs.concordia.ca, weiping@ece.concordia.ca, benoit.champagne@mcgill.ca
*Department of Electrical and Computer Engineering, Concordia University, Canada
†Department of Electrical and Computer Engineering, McGill University, Canada

***Abstract*—Speech spectrograms exhibit strong contextual dependencies along both time and frequency dimensions. In this paper, a novel composite model integrating a long short-term memory (LSTM) and convolutional neural network (CNN) to exploit temporal and spectral contextual speech information, respectively, is proposed for speech separation. LSTM and CNN operate in a parallel fashion to speed up the process and independently extract a complementary set of speech features. A fully-connected network then maps these features to the real and imaginary components of a ratio mask to enhance the magnitude and phase of the corrupted speech simultaneously. In the CNN path, a new delicately designed CNN with frequency dilated one-dimension (1D) convolutional layers is employed to expand the receptive field of CNN kernels without increasing the complexity. Furthermore, this CNN benefits from residual learning and skip connections to facilitate training and accelerate convergence. In spite of different neural networks included in the composite model, the proposed separation system not only has a low computational complexity, but also significantly outperforms some other deep learning-based methods.**

I. Introduction

Speech separation aims at separating a desired speech signal from its noisy background, consisting of ambient noise and interference. It has been a challenging topic in the speech processing area and found many important applications, especially in speech recognition-related services and products.

Thanks to growing computing resources and widely available training datasets, significant advances have been made in data-driven approaches for speech separation in recent years. These approaches model speech separation as a supervised learning problem and help resolve some issues of traditional unsupervised methods, like musical noise and speech distortion [1][2]. In particular, deep learning as a promising alternative to statistical solutions has been extensively used to develop supervised methods.

Xu *et al.* [3] employed a fully-connected (FC) network to directly map the log-power spectrum (LPS) of noisy speech to that of the clean one, and reported a significant improvement on speech quality and intelligibility in comparison with traditional speech separation methods. Many similar methods were introduced, as in e.g. [4] and [5]. Although direct mapping is straightforward, it requires a large training dataset to accurately learn the mapping between input and output. On contrast, deep learning-based methods were proposed in [6] and [7], where the network target is one of the common spectral masks, such as ideal binary mask (IBM) and ideal ratio mask (IRM). In these methods, the masks are applied to a subset of spectrogram time-frequency (TF) cells of the noisy speech. These methods yield notable improvements in speech separation results. The authors in [7] also compared the speech separation performance resulting from different mask types with direct mapping, and showed that masking-based methods generally give a better separation performance. Besides, to enhance speech phase alongside magnitude, a complex IRM (cIRM) was presented in [8], where an FC network was employed to predict the spectral mask.

Most of the above methods use an FC network to estimate a desired target, while they neglect the strong temporal dependencies of speech. Even though some of these studies adopt a concatenation of several consecutive speech frames as input to the network to make use of the temporal information, this not only causes additional network complexity, but also processing delays, especially when a large window size is chosen. Moreover the information outside the window is always ignored regardless of the window length. To resolve these issues, Jitong *et al.* [9] proposed an LSTM network for IRM estimation which shows substantial performance improvements as well as better speaker generalization over the FC network. Although LSTM exploits the temporal contextual information of speech, it does not consider the spectral dependencies in speech spectrogram.

Several neural network architectures based on a combination of LSTM and CNN including CLDNNs [10], multi-LCNN [11], and hybrid LSTM-CNN [6] were suggested to extract the temporal and spectral contextual information of an input voice for acoustic scene classification. Apart from the fact that these models were not tested for speech separation, due to the restricted receptive field of the CNN filters, they encounter limitations in the CNN path. Since the frequency dimension of speech spectrogram is on the order of a few hundred, there is a need for a large receptive field of CNN kernels to maintain the contextual information along the frequency axis. To overcome this issue, researchers have suggested using large CNN kernels and stride convolution. However, a larger CNN kernel increases the complexity while the stride convolution overly smooths the TF cell prediction. Furthermore, the pooling layer in the traditional CNN structure only retains the rough information of the receptive field. Hence, a new CNN structure with 1D convolution in frequency and 2D dilated convolution in the TF plane was proposed to overcome these

978-1-7281-8058-8/20/$31.00 ©2020 IEEE

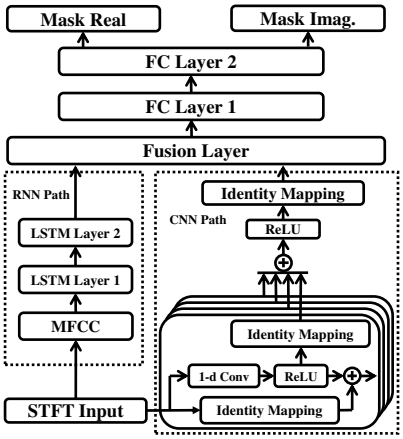

Fig. 1: Proposed composite model with two parallel paths.

problems in our previous work [12].

In this paper, we propose a composite model using a light LSTM and a new low-complexity CNN to simultaneously extract a complementary set of spectral and temporal speech features. The LSTM and CNN operate independently in a parallel fashion, where the former extracts temporal information in the RNN path, while the latter simultaneously exploits spectral information in the CNN path. Unlike the serial models such as CLDNN and LCNN, where the performance of each step depends on the previous step, the components in our composite model function independently which can speed up the process. The outputs of both paths are then fused and input to an FC network to be mapped to a complex ratio mask for enhancing both speech magnitude and phase. It is shown that the new composite model not only gives a superior speech separation performance, but also has a lower complexity compared to some existing deep learning-based methods.

II. Proposed Composite Model

The proposed model integrates an RNN, CNN, and FC network as shown in Fig. 1. For the CNN path, the short-time Fourier transform (STFT) of the input speech is fed to the CNN to exploit the spectral contextual information of the speech. Meanwhile, the Mel frequency cepstral coefficients (MFCCs) are computed within the RNN path and then input to the LSTM network to extract the speech temporal information. The fusion layer combines the outputs from the CNN and RNN paths and delivers them to an FC network for the final regression where the objective is to estimate the real and imaginary components of a complex ratio mask. The main components of the network are described in the following.

A. Complex Spectrogram and Ratio Mask

Application of STFT to the clean speech yields its complex spectrogram as $S(k,l) = S_r(k,l) + jS_i(k,l)$, where S_r and S_i are the real and imaginary components of S, and k and l denote time frame and frequency bin, respectively. In the sequel, the (k,l) argument will be omitted for brevity. Since S_r and S_i bear both speech phase and magnitude information, they can be considered as the training target of the network to enhance both the magnitude and phase of the noisy speech simultaneously. Due to similar structures, a single network can be employed to predict both S_r and S_i at the same time [13]. However, estimating a TF mask is more efficient than a direct spectrogram [8]. Hence, considering that $S = M \circ Y$, where M and Y respectively denote the STFT of the spectral mask and noisy signal and $\circ$ denotes element-wise multiplication, we can express mask M in terms of its real and imaginary components as follows,

$$M = \frac{Y_r S_r + Y_i S_i}{{Y_i}^2 + {Y_r}^2} + i\frac{Y_r S_i - Y_i S_r}{{Y_i}^2 + {Y_r}^2} \quad (1)$$

It is worth mentioning that these components are then compressed by tangent hyperbolic, since they are originally of a wide dynamic rage not suitable for neural network.

B. CNN Path: Dilated 1D Frequency Convolution

The purpose of the CNN path is to exploit spectral contextual information of speech. The real and imaginary components of the input STFT are fed to the CNN as two channels. To exponentially expand the receptive field of CNN kernels, four dilated 1D frequency convolution layers are stacked with increasing dilation rates of 1, 2, 4, and 8. The 1D convolution is chosen, since the goal of this path is to exploit the contextual information alongside the frequency axis and reduce the computational burden of CNN. To maintain the network symmetry, the number of channels for the four layers is respectively set to 16, 32, 16, and 8 with ReLU activation function. The output of each layer is input to an identity mapping layer to adjust the number of channels and then added to the outputs of other dilated 1D frequency layers. By the identity mapping layer, we mean a layer with a kernel size of 1×1 which just changes the number of channels. To adopt residual learning, the input of each layer is taken as the sum of the output and the bypassed input of the previous layer. Finally, the number of channels of CNN output is shrunk by the last identity mapping layer, and then the flattened output is delivered to the fusion layer.

C. RNN Path

In the composite model, LSTM is employed to extract temporal contextual information of the input speech. In the RNN path, two layers are stacked each having 128 LSTM units. The dropout technique at the rate of 0.3 is adopted to avoid over-fitting and improve generalization. We point out that, unlike CNN, feeding LSTM with input STFT will not lead to good results. Hence, MFCCs concatenated with their deltas and acceleration are selected as the LSTM input, as further explained in Section III-C.

D. Fusion and Regression

The flattened output of the CNN and RNN paths are combined by an FC fusion layer. Afterward, the enriched complementary set of features bearing temporal and spectral information of the input speech is mapped to the real and imaginary components of a complex ratio mask using a 2-layer FC network, where each layer consists of 512 nodes

with ReLU activation function being utilized. Furthermore, the dropout technique is again utilized here at the rate of 0.3. It is worth mentioning that the composite model never encounters an over-fitting problem since it is trained with a huge amount of data while the number of parameters is relatively small. As such, the network learns just fundamental information from the training dataset but not the details.

III. Experiments

A. Experimental Setup

The performance of the composite model is evaluated with TIMIT [14] dataset consisting of 6300 utterances spoken by 630 males and females. These utterances are mixed with random cuts of non-stationary noises, namely babble, restaurant, street, and factory, from NOISEX-92 [15] at SNR levels of -5, 0, 5, and 10 dB. In total, more than $100,000$ (6300×4 noises $\times 4$ SNR levels) mixtures form the training dataset. The sampling rate is set to 16 kHz and the input utterances are divided into frames of 320 samples using a Hanning window with 160 samples overlap, equivalent to 20 ms frame length and 10 ms frameshift, based on which a 320-point DFT is then computed for each frame. Testing is performed using 60 unmatched utterances mixed with unseen cuts of the aforementioned noises and unmatched SNR levels of -6, 0, 6, and 12 dB, i.e. 960 (6300×4 noises $\times 4$ SNR levels) mixtures. Adam optimizer is used to minimize the MSE cost function defined between the ground truth and the network-estimated real and imaginary components of the ratio mask with size 322 (161×2 for real and imaginary). Finally, evaluation of the enhancement results is carried out by means of PESQ and segmental signal-to-noise ratio (SSNR) objective measures.

B. Comparison of RNN types for Composite Model

LSTM consists of three control gates and one memory cell. An efficient implementation of LSTM is GRU which consists of just two gates and no memory cell. These processing units can be also used in a bidirectional fashion. Here, we test bidirectional LSTM (BLSTM) in our comparison study [16].

To conduct the comparison, two hidden layers of the aforementioned RNN types, each having 128 units, are used to build the RNN path of the composite model. Figure 2(a) illustrates the average PESQ score improvement of the composite model using BLSTM, LSTM, and GRU, on all the mentioned noises and SNR levels. Figure 2(b) shows the comparison of computational time, memory, and the number of parameters of the model using different RNN variations. Clearly, the model using BLSTM outperforms that using LSTM or GRU for males, but the number of parameters of the model using BLSTM jumps by around 33% compared to LSTM and GRU. Consequently, more computational time and memory are exhausted by the model using BLSTM in the RNN path. However, though the computational cost of using GRU or LSTM in the RNN path is roughly the same, the performance of the model using LSTM is better than GRU for males and better than both BLSTM and GRU for females. Hence, LSTM offers the best trade-off for the composite model to exploit the temporal contextual information of the input speech.

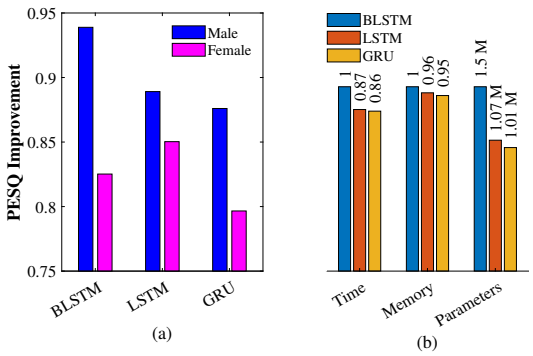

Fig. 2: Comparison of different units, (a) Average PESQ score improvement, (b) Comparison of computational time, memory, and number of parameters (in million).

C. Comparison of Different LSTM Network Inputs

Here, we investigate the whole model performance using well-known spectrogram-based and Gammatone-domain features, namely: MFCC, log Mel-filterbank energy (Log-Mel) [17], Gammatone frequency (GF) [5], and multi-resolution cochleagram (MRCG) [18]. These features are concatenated with their delta and acceleration and then normalized to zero mean and unit variance to avoid unbiased participation of different components of the feature vector. Consequently, the feature size for MFCC, Log-Mel, GF, and MRCG is $39, 78, 64,$, and 768, respectively. The comparison results are shown in Fig. 3 in terms of PESQ performance, computational time, memory, and the number of parameters. As shown, MRCG yields the best model performance which results not only from the high quality of these features, as they use local and contextual information of speech cochleagram, but also from the high number of network parameters when employing these features, which is about 27% more than others. Similarly, GF features lead to relatively better results as they are also defined in gammatone-domain. However, extracting these features is quite costly in terms of computational time in comparison with Log-Mel and MFCC, which are extracted roughly 20 times faster, while the model performance using them is comparable with GF. Between Log-Mel and MFCC which result in equal model performance, MFCC is preferable because of its smaller dimension. In conclusion, the MFCC features are used as the input of the LSTM network.

D. Comparison with other DNN-Based Methods

To show the advantage of the composite model, it is compared with some other well-known deep learning-based methods. Spectral magnitude mask (SMM), IRM [7], and cIRM are three mask types that are predicted by an FC network with three-layers each having 1024 units. Predicted SMM and IRM are applied to the magnitude spectrogram and the clean speech is then reconstructed using the input noisy phase, while cIRM enhances both magnitude and phase simultaneously. In contrast, FFT-Mag [7] directly maps input STFT magnitude to the clean speech spectral magnitude using a three-layer FC network with 1024 units per layer, and the target magnitude

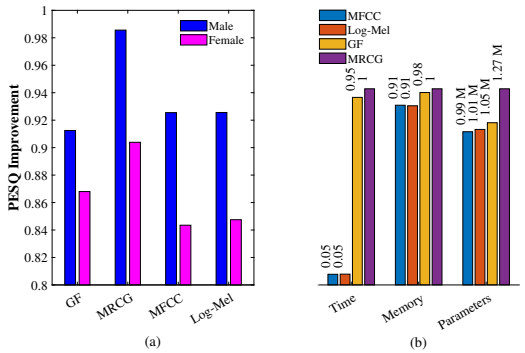

Fig. 3: Feature comparison: (a) Average PESQ score improvement; (b) Comparison of computational time, memory, and number of parameters (in Million).

TABLE I: Average SSNR and PESQ Scores of Different Methods With Unmatched SNR Levels

Method	PESQ				SSNR				No. of Parameters
	-6	0	6	12	-6	0	6	12	
Unprocessed	1.29	1.64	2.02	2.41	-9.03	-5.00	-0.50	4.31	-
	0.98	1.35	1.79	2.20	-8.54	-4.85	-0.36	4.33	
FFT-Mag	1.84	2.28	2.61	2.81	0.49	2.14	3.44	4.40	2.66M
	1.40	1.80	2.11	2.28	0.74	2.21	3.44	4.28	
TMS	1.72	2.20	2.62	2.95	0.88	2.68	4.53	6.07	12.35M
	1.49	2.00	2.45	2.82	1.45	3.26	4.98	6.29	
IRM	1.83	2.40	2.95	3.41	-0.98	3.25	7.13	**10.21**	2.66M
	1.41	2.02	2.62	3.15	-0.55	3.62	**7.73**	**11.14**	
SMM	1.85	2.30	2.71	3.10	-1.22	2.03	5.50	9.51	2.66M
	1.45	1.95	2.39	2.83	-0.56	2.70	6.07	9.97	
cIRM	1.98	2.48	2.94	3.34	1.04	3.84	6.63	9.21	2.82M
	1.65	2.17	2.63	3.07	1.22	4.00	6.59	9.15	
Proposed	**2.00**	**2.52**	**3.01**	**3.42**	**1.50**	**4.25**	**7.19**	10.16	**0.99M**
	1.74	**2.30**	**2.76**	**3.20**	**2.04**	**4.72**	7.54	10.25	

spectrum (TMS) [3] maps log-power spectral magnitude of the noisy speech to that of the clean one by the same network, except for the use of 2048 units per layer. Both FFT-Mag and TMS reconstruct the clean speech with input noisy phase.

The results of the comparisons are shown in Table I. The values in the table are the average of PESQ score and SSNR over all the noises at unmatched SNR levels of -6, 0, 6, and 12 dB. The upper and lower numbers in each table cell are for males and females, respectively. As shown, the proposed composite model outperforms all other methods in terms of the PESQ score. With respect to the SSNR, the composite model yields better results at SNR levels of -6 and 0 dB for both males and females as well as at 6 dB for males, while IRM performs slightly better for females at 6 dB SNR level and also marginally better at 12 dB for both males and females. In regards to the number of model parameters, the same table shows that the promising performance of the composite model is achieved at a lower computation cost where only 0.99 M parameters are involved.

IV. Conclusion

In this paper, a composite model has been proposed for speech separation in which a light LSTM and a new CNN structure are exploited to extract the temporal and spectral information of input speech. A complex ratio mask is considered as the network objective to simultaneously enhance both magnitude and phase of the input mixture. The performance of the composite model using different RNN variations with different inputs was then compared. Through a series of comparative experiments, the advantages of the proposed model over some known deep learning-based methods in both separation performance and computational complexity were finally demonstrated.

References

[1] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, "A deep neural network based harmonic noise model for speech enhancement." in *INTERSPEECH*, 2018, pp. 3224–3228.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2018.

[3] 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. on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[4] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245.

[5] X. Zhao, Y. Wang, and D. Wang, "Cochannel speaker identification in anechoic and reverberant conditions," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1727–1736, 2015.

[6] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," in *Proc. of the Detection and Classification of Acoustic Scenes and Events (DCASE2016)*, 2016, pp. 11–15.

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

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

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

[10] 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.

[11] 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 *IEEE Int. Conf. on Tools with Artificial Intelligence*, 2018, pp. 52–59.

[12] 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.

[13] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2017, pp. 1–6.

[14] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa TIMIT acoustic-phonetic continuous speech corpus CD-ROM. nist speech disc 1-1.1," *NASA STI/Recon Technical Report*, vol. 93, 1993.

[15] 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.

[16] K.-L. Du and M. Swamy, "Recurrent neural networks," in *Neural networks and statistical learning*. Springer, 2019, pp. 351–371.

[17] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[18] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1993–2002, 2014.