CONSTRAINED CONVOLUTIONAL-RECURRENT NETWORKS TO IMPROVE SPEECH QUALITY WITH LOW IMPACT ON RECOGNITION ACCURACY

*Rasool Fakoor**

Dept. of Computer Science and Engineering
Univ. of Texas at Arlington, TX 76019
rasool.fakoor@mavs.uta.edu

Xiaodong He, Ivan Tashev, Shuayb Zarar

Microsoft Research
Redmond WA 98052
{xiaohe,ivantash,shuayb}@microsoft.com

ABSTRACT

For a speech-enhancement algorithm, it is highly desirable to simultaneously improve perceptual quality and recognition rate. Thanks to computational costs and model complexities, it is challenging to train a model that effectively optimizes both metrics at the same time. In this paper, we propose a method for speech enhancement that combines local and global contextual structures information through convolutional-recurrent neural networks that improves perceptual quality. At the same time, we introduce a new constraint on the objective function using a language model/decoder that limits the impact on recognition rate. Based on experiments conducted with real user data, we demonstrate that our new context-augmented machine-learning approach for speech enhancement improves PESQ and WER by an additional 24.5% and 51.3%, respectively, when compared to the best-performing methods in the literature.

Index Terms— Speech Enhancement, Deep Learning, Multi-task Learning, Curriculum Learning, Language Model.

1. INTRODUCTION

Recently, deep learning architectures have led to remarkable progress in problems like speech recognition [1, 2], image classification [3], machine translation [4, 5], image and video caption generation [6, 7], speech separation and enhancement [8, 9, 10, 11] and many others. Speech enhancement is the process of eliminating noise from an audio signal prior to primarily two higher-level tasks, namely recognition and playback through speaker phones [12]. Because traditional analytical processing methods have a limited capacity to capture complex signal and noise statistics, data-driven approaches are becoming increasingly popular to enhance speech [13, 14, 15, 16]. These learning based approaches typically aim to optimize a particular criterion during training (i.e. the signal mean-squared error (MSE)), while the performance of speech enhancement is usually evaluated from different aspects by multiple metrics (e.g. WER, PESQ) during test and inference time. Therefore, there is a metric discrepancy between training and evaluation, which leads to suboptimal performance.

Jointly training the speech enhancement and recognition systems (i.e. ASR) to simultaneously improve MSE and WER could potentially alleviate this problem. Unfortunately, not only optimizing such a model can be extremely challenging namely due to the complexity of ASR models but also it can be computationally very expensive, raising the need for careful modeling and training. Motivated by these observations, in this paper, we propose a model that not only effectively combines global and local contextual knowledge to enhance speech but also learns how to regularize the speech enhancement and denoising model such that the metric discrepancy could be mitigated. Specifically, to achieve good enhancement performance, our proposed model consists of convolutional layers coupled with recurrent cells. Further, we constrain this model by including a language model/decoder in the optimization objective function. Thus, our network tries to limit the impact on recognition rate, while improving speech quality. To effectively train our model, we also adapt a curriculum-learning-based [17] training paradigm.

In contrast to our approach, existing methods for speech enhancement utilize a single, unconstrained signal-quality criterion such as the MSE for optimization [13, 14, 15, 16]. Thus, although these algorithms improve speech quality, they degrade the recognition rate (measured by the WER metric). In this paper, we aim to overcome this limitation. The following are the specific contributions that we make:

- We propose a contextually-aware neural-network architecture for speech enhancement that is constrained with a language-decoder model to limit the impact on WER.
- We demonstrate a methodology to train such a network based on curriculum learning for multi-task regression.
- Through extensive experimentation and analysis, we show simultaneous improvements of 24.5% and 51.3% in PESQ and WER over existing methods in the literature that only optimize signal quality.

*Work was done as an intern at Microsoft Research Redmond.

2. PROPOSED APPROACH

One of the main challenges in speech enhancement when using deep neural networks is to effectively combine the local and global structure of input frames. For example, the model not only needs to learn how to denoise an audio frame based on the signal in that frame but also needs to take into account the temporal structure of the entire sequence of frames over a short span of time. The recurrent neural network (RNN) is a good fit for the speech enhancement problem given its capability to model the temporal structure of speech data. As we will show in section 3, although the RNN provides a useful structure for this problem, it is insufficient to achieve good performance. This is because input speech segments are very long, usually containing thousands of frames, making it difficult for the RNN to catch both local and global contextual information for speech enhancement.

Moreover, state-of-the-art speech enhancement and denoising models are usually trained on a particular criterion, while the performance is evaluated from different aspects by multiple metrics. For example, most of the models for speech enhancement are formulated as a regression problem [14, 15] and use MSE as the loss function during training. However, during the evaluation, PESQ, WER, or sentence error rate (SER) are used to assess the performance of the trained model. There is a significant metric discrepancy between training and testing, *e.g.*, a model that is trained to achieve the lowest MSE during training does not necessarily give an improvement in WER or SER at test time.

To address these problems, we first propose a convolution-recurrent neural network (CRNN) that can efficiently model local and global structure of the speech data. Moreover, we also propose a multi-task learning approach that addresses the metric discrepancy problem and leads to a more robust performance on the speech enhancement and denoising task.

2.1. Combining Local and Global Contexts

One way to capture temporal structure of the data is to use RNNs to model this relationship. However, simple RNNs do not have the adequate capacity to model both the long-term dependencies and local contextual information among different frames [18, 19]. However, for good performance, the model needs to capture the local context among neighboring frames as well as the global context. This is important because the denoising networks not only need to use the surrounding frames to denoise the current frame but also higher-level relationships to build a more effective model.

Motivated by these observations, we propose the CRNN, which models long-term dependencies between frames by the recurrent structure in the network and the local context by applying a convolution network over a local context window of neighboring frames. In this model, at every time step, t, our model first utilizes eight neighboring frames as the input to a three-layer convolution network that models the local structure of the input frame (f_t). The output of this network will

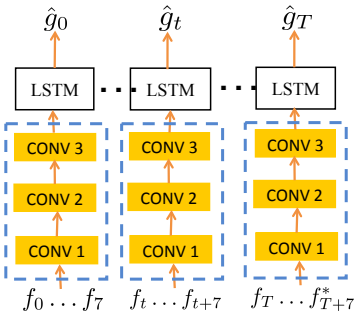

Fig. 1. Proposed architecture to combine local and global context of input frames.

be an input to an LSTM [19] unit at time t. The recurrent unit uses the current noise frame as well as previous hidden states to reconstruct an enhanced frame. To be specific, our network at time step t utilizes eight neighboring frames and h_{t-1} to reconstruct a single denoised frame $\hat{g}_t$. Our proposed model is shown in Figure1.

The objective function for this model minimizes the error between the enhanced frame (or denoised frame for short) and clean frame which can formally be defined as follow:

$$L_{re}(F;\omega) = min_{\omega} \sum_{t} \parallel \hat{g}_t - g_t \parallel_2^2 \qquad (1)$$

where ω is a network parameter, $\hat{g}_t$ is a denoised frame, and g_t is the clean frame that we use to train the parameters of this model. As we show in the experimental results section, this model outperforms other networks that either only model the local structure or the global structure of the data.

2.2. Multi-task Learning

The metric discrepancy is also the main challenge that most of the speech enhancement and denoising models face [14, 15, 16]. For example, while most of the models use MSE as the training metric, the performance is evaluated on other criteria such as PESQ, SER, and WER. One possible solution is to train the models to directly optimize PESQ or WER. Although this is plausible, there are a couple of problems. First, these metrics are very expensive to calculate and often it is not practical to use them directly during training. Moreover, these metrics (i.e PESQ) are usually discrete and non-differentiable, making the optimization (Eq. 1) very difficult. The REINFORCE algorithm [20] can be used in certain situation, but gradient estimation using REINFORCE would be non-trivial in this setup as it deals with a continuous number for sampling and gradient calculation (i.e. at each time step, network outputs a denoised frame which is a continuous vector in the speech space, making the policy-gradient estimating a hard problem [21]).

Motivated by these problems, we propose a multi-task learning framework during training that uses a language model (i.e. language decoder) to regularize the training and

improve the performance of the denoising network with respect to PESQ, WER, or SER. That is, we first run the model to perform denoising, and once the denoising is done, i.e., reaches the last frame of the input speech segment, our approach uses the last hidden unit representation of the CRNN as the input to a RNN based language model, in which the language model is trained to generate the text transcript of the input speech segment. This is like imposing another task of sequence-to-sequence multimodal translation, which encodes the sequence of the denoised speech signal into one vector representation and then translates it into the sequence of words in the transcription. In order to build this language model, we add the following loss function:

$$L_{lm}(S, H_t; \theta) = -\sum_{j'}^{T}\sum_{i}^{|V|} s_i^{j'} \log(\hat{s}_i^{j'}) + \lambda \parallel \Theta \parallel_2^2 \quad (2)$$

where H_T is the last hidden unit of the denoising model, S is a transcript for a given file, and θ is the parameter of this network. The Eq. 2 is cross-entropy loss function that tries to minimize the word prediction error. Combing this loss function with Eq. 1 helps the network to constraint and regularize the denoising network such that it will have better performance regarding WER, SER, and PESQ during test time:

$$L(S, F, H_t; \mathbf{\Theta}) = L_{re} + \lambda_1 L_{lm} + \lambda_2 \parallel \Theta \parallel_2^2 \quad (3)$$

This architecture is shown in Figure 2 in which the language model is shown in dotted box (b) and denoising model is shown in dotted box (a). It is worth noting that the intuition behind this model is that the original denoising model does unconstrained optimization[1] as the result, the denoising model only minimizes the MSE as much as it can without considering PESQ, WER, etc. This causes the model to sometimes overfit on the MSE metric and shows worst performance on other metrics. However, by adding the language model, the model is not only focused on minimizing the MSE, but also tries to denoise in a way such that the denoised speech signal can lead to better word prediction decoded by the language model. Therefore, adding the language modeling task effectively regularize the training of the denoising model and will lead to more robust performance as reflected in the improvements in terms of WER and PESQ too. As the results show, this approach is very effective, outperforms other methods significantly on a range of evaluation metrics.

2.3. Curriculum Learning

The language model and the denoising model operate very differently, while the language model catches the dependency at the word level, the denoising model works at the lower speech frame level. If we train them together from scratch, the model has hard time to converge. Specifically, the denoising model needs hundreds of epochs to converge to a stable

[1]By unconstrained optimization, we are referring to the fact that we did not explicitly impose any constraints on Eq. 1, i.e. bound the model outputs

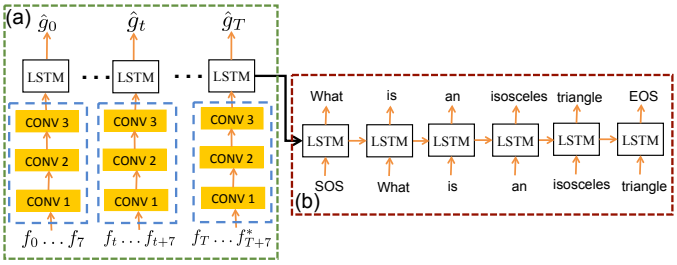

Fig. 2. Our multi-tasks based learning architecture.

model given the level of difficulty and complexity in this denoising problem. On the other hand, since there is one transcript per file and usually it is short (i.e around 60 words per files), the language model only needs a few epochs to converge. To deal with this problem, we design a curriculum learning paradigm [17] to train this model. We first train the denoising model for few hundred epochs until it stops improving, i.e. only train it with Eq.1. At this stage, we introduce Eq. 3 to the model as the new objective functions. Note that our proposed curriculum learning is different from traditional one [17] such that in the traditional curriculum learning, it first starts with the simpler problem then goes to the harder problem. However, in our proposed approach, we first start with the core task, then we combine it with another task to further regularize the training. As shown in the next section, the proposed method is very effective for the challenging denoising task.

3. EXPERIMENTS

Dataset

We evaluate the performance of our methodology with single-channel recordings based on real user queries to the Microsoft Windows Cortana Voice Assistant. We split studio-level clean recordings into training, validation and test sets comprising 7500, 1500 and 1500 queries, respectively. Further, we mixed these clean recordings with noise data (collected from 25 different real-world environments), while accounting for distortions due to room characteristics and distances from the microphone. Thus, we convolved the resulting noisy recordings with specific room-impulse responses and scaled them to achieve a wide input SNR range of 0-30 dB. Each (clean and noisy) query had more than 4500 audio frames of spectral amplitudes, each lasting 16 ms. We applied a smoothing function based on a Hann window to the frames allowing accurate reconstruction with a 50% overlap. These audio frames in the spectral domain formed the features for our algorithm. Since we utilized a 512-point short-time Fourier Transform (STFT), each feature vector was a positive real number of a dimensionality of 256.

Hyperparameter Optimization

We use random search [22] on the validation set to select hyperparameters for this dataset. A stack of two LSTMs are

Method	SNR(DB)	LSD	MSE	SIR	SDR	SAR	WER	SER	PESQ
Noisy data	15.18	23.07	0.04399	39.1	−0.67	−0.66	15.4	25.07	2.26
Existing model [16]	41.08	17.49	0.03533	8.58	−0.58	1.68	44.93	66.60	2.19
Existing model [14]	40.70	20.09	0.03485	7.47	−1.28	1.32	54.92	75.87	2.17
Existing model [14]	44.51	19.89	**0.03436**	7.84	−1.04	1.48	55.38	74.93	2.20
Existing model [15]	27.03	20.84	0.03711	5.82	−2.36	0.80	60.72	79.87	1.93
Our model (NOC)	40.23	17.78	0.03544	6.84	−1.50	1.23	45.19	66.40	2.23
Our model (CRNN)	41.26	15.90	0.03480	9.52	0.19	2.12	22.73	38.93	2.70
Our model (CRNN + LM)	44.22	15.94	0.03462	9.48	0.18	2.14	23.79	40.13	2.69
Our model (CRNN + LM +CL)	40.38	16.14	0.03457	**9.76**	**0.36**	**2.23**	**21.90**	**37.27**	**2.74**
Clean data	57.31	1.01	0.0	79.02	57.05	58.36	2.19	7.4	4.48

Table 1. Performance comparison for speech enhancement tasks.

used in our best model (**CRNN + LM** and **CRNN**). These models both have 1072 hidden units. The weight decays are $2.8951e^{-5}$ and $3.6998e^{-5}$. In addition, the **LM** uses 857 words in its vocabulary and all transcript are capped to have 60 words maximum. In order to optimize our network parameters, we use Adam [23] with learning rates of $6.4710e^{-5}$ and set $\beta 1$, $\beta 2$ to 0.8 and 0.999, respectively. The convolution layers in our models (yellow boxes in Fig. 1 and 2) have the following specifications: 1) Conv 1 has 16 filters with the kernel size of $(7, 5)$, stride size of $(3, 1)$, 2) Conv 2 has 32 filters with the kernel size of $(5, 3)$, stride size of $(3, 1)$, and 3) Conv 3 has 64 filters with the kernel size of $(5, 1)$, stride size of $(3, 1)$. In addition, all convolution layers use $(2, 1)$ dilation [24] as well.

Performance Comparison

We carry out an extensive evaluation to evaluate the proposed models. In the evaluation, we compare the proposed model with state-of-the-art baselines for the speech enhancement and denoising task. We summarize the results of these experiments in Table 1. We compare our models to recent deep neural network based approaches which are strong baselines, including [15], [14], and [16]. We first build a model (**NOC**) that does not consider the global context of the data (i.e. no RNN) and only considers local context. Then we extend these models to our denoising model **CRNN**. As the results show, our proposed model **CRNN** outperforms the baselines on the key metrics of PESQ, WER, SER, and others.

In addition, Table 1 shows that our proposed multi-task model (**CRNN + LM**) outperforms other models and furthers improve PESQ, WER, and SER. In addition, we show in Figure 3 the improvement in PESQ scores by using our model.

Curriculum learning

We also studied the impact of the proposed curriculum learning procedure. Given the large gap between denoising model which operate at the lower speech frame level and the language model which operate at the higher word level, it is important to use the proposed curriculum learning based training method (section 2.3) to train our model. As results in Table show 2, when we train both denoising network and language model together from the beginning, the performance is quite bad, compared to the performance using the proposed curriculum learning.

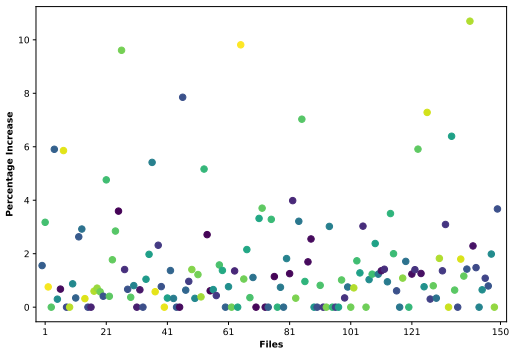

Fig. 3. The positive percentage increase for PESQ score in the test files.

Method	WER	SER	PESQ
Our model (CRNN)	22.73	38.93	2.70
Our model (CRNN + LM)	23.79	40.13	2.69
Our model (CRNN + LM + CL)	**21.90**	**37.27**	**2.74**

Table 2. Results on effects of Curriculum learning (CL).

4. CONCLUSION

In this paper, we propose a model that combines both local and global contextual information for speech enhancement. We show that our approach leads to better enhancement performance compared to existing baselines. Furthermore, we propose multi-task learning with curriculum learning, which regularizes the training process of the speech-enhancement model through a language model/decoder. Thus, we limit the impact of speech enhancement on recognition accuracy.

5. REFERENCES

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[2] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, Jan 2012.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *ICLR*, 2015.

[5] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks," in *NIPS*, 2014, pp. 3104–3112.

[6] Rasool Fakoor, Abdel-rahman Mohamed, Margaret Mitchell, Sing Bing Kang, and Pushmeet Kohli, "Memory-augmented attention modelling for videos," *CoRR*, vol. abs/1611.02261, 2016.

[7] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[8] Suman Samui, Indrajit Chakrabarti, and Soumya K Ghosh, "Deep recurrent neural network based monaural speech separation using recurrent temporal restricted boltzmann machines," in *INTERSPEECH*, 2017.

[9] Y. Wang, J. Du, L. R. Dai, and C. H. Lee, "A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1535–1546, July 2017.

[10] M. Kim, "Collaborative deep learning for speech enhancement: A run-time model selection method using autoencoders," in *ICASSP*, March 2017, pp. 76–80.

[11] Rasool Fakoor, Xiaodong He, Ivan Tashev, and Shuayb Zarar, "Reinforcement learning to adapt speech enhancement to instantaneous input signal quality," in *NIPS, Machine Learning for Audio Signal Processing Workshop*, 2017.

[12] I. Tashev, A. Lovitt, and A. Acero, "Unified framework for single channel speech enhancement," in *2009 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, Aug 2009, pp. 883–888.

[13] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, Jan 2014.

[14] Seyedmahdad Mirsamadi and Ivan Tashev, "Causal speech enhancement combining data-driven learning and suppression rule estimation," in *INTERSPEECH*, 2016.

[15] Yong Xu, Jun Du, L.-R Dai, and C.-H Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *INTERSPEECH*, 01 2014, pp. 2670–2674.

[16] Andrew Maas, Quoc V. Le, Tyler M. ONeil, Oriol Vinyals, Patrick Nguyen, and Andrew Y. Ng, "Recurrent neural networks for noise reduction in robust asr," in *INTERSPEECH*, 2012.

[17] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston, "Curriculum learning," in *ICML*. 2009, pp. 41–48, ACM.

[18] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio, "Understanding the exploding gradient problem," *CoRR*, vol. abs/1211.5063, 2012.

[19] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computing*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[20] Ronald J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Journal of Machine Learning*, vol. 8, no. 3-4, pp. 229–256, May 1992.

[21] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller, "Deterministic policy gradient algorithms," in *ICML*. 2014, pp. 387–395, JMLR Workshop and Conference Proceedings.

[22] James Bergstra and Yoshua Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, Feb. 2012.

[23] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[24] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," *ICLR*, 2016.