PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Towards Neural-Based Single Channel Speech Enhancement for Hearing Aids

Muhammed Shifas PV[(1)], Claudio Santelli[(2)], Yannis Stylianou[(1)]

[(1)]Speech Signal Processing Lab (SSPL), University of Crete, Greece, shifaspv@csd.uoc.gr
[(2)]Sonova AG, Staefa, Switzerland

Abstract

Advancements in machine learning techniques have promoted the use of deep neural networks (DNNs) for supervised speech enhancement. However, the DNN's benefits of non-explicit noise statistics and nonlinear modeling capacity come at the expense of increased computational complexity for training and inference which is an issue for real-time restricted applications, like hearing aids. Contrary to the conventional approach which separately models the feature extraction and temporal dependency through a sequence of convolutional layers followed by a fully-connected recurrent layer, this work promotes the use of convolutional recurrent network layers for single-channel speech enhancement. Thereby, temporal correlations among inherently extracted spectral feature vectors are exploited, while further reducing the parameter set to be estimated relative to the conventional method. The proposed method is compared to a recent low algorithmic delay architecture. The models were trained in a speaker independent fashion on the NSDTSEA data set composed of different environmental noises. While objective speech quality and intelligibility measures of the two architectures are similar, the number of network parameters in the suggested enhancement method being reduced by 66%. This reduction is highly beneficial for storage and computation constraint applications.

Keywords: speech enhancement, deep neural networks, convolutional recurrent network

1 INTRODUCTION

Speech enhancement (SE) technique aims to suppress the background noise in a noisy communication scenarios. Classical speech enhancement methods like spectral subtraction and Wiener filtering partly addressed this challenge by modeling first and second order noise statistics parametrically [1]. However, the assumption of an explicit noise model and its slowly varying noise parameters are limiting factors of these models.

Recent advancements in neural networks (NN) and their non parametric noise modeling capacity have led to improvements in the field of single channel speech enhancement. Convolutional layers capture locally varying pattern of the input while considerably reducing the number of model parameters to be estimated. For time sequence modeling, the temporal context window size in which temporal correlation are exploited highly depends on the kernel dimensions and convolutional layer stack size. However, for SE tasks, modeling of successive frame dependency over a long time window is essential to capture noise and speech characteristics.

Recurrent NN(RNN) explicitly exploit temporal correlations in time sequences [2] . A recent low-latency source separation approach proposed an architecture with a convolutional feature extraction stack followed by recurrent processing thereof [3]. The recurrent cells were composed of fully-connected sub-layers. However, these fully connected sub-layers come at the expense of a large memory footprint potentially limiting their applications into computation constraint applications.

Convolutional RNN are composed of convolutional linear operators, and thus, significantly reduce the number of weights to be estimated [4]. Promising results have been shown in image classification tasks [5]. In this work, we propose a recurrent feature extraction scheme based on ConvLSTM's for single-channel SE (ConvLSTM-SE). Its performance is compared to a low-algorithmic latency approach separating feature extraction and temporal modeling by a cascade of convolutional and long short-term memory (LSTM) layers.

2 PROPOSED CONVOLUTIONAL LSTM SPEECH ENHANCEMENT (ConvLSTM-SE) MODEL

The idea of recurrent feature extraction was initially proposed by Xingjian Shi et.al [4]. The current feature estimate H_t, C_t depends on the previously extracted feature H_{t-1}, C_{t-1} and the current input instance X_t (figure 1). This dependency is mathematically formulated as:

$$\begin{aligned} i_t &= \Phi(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} o C_{t-1} + b_i) \\ f_t &= \Phi(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} o C_{t-1} + b_i) \\ C_t &= f_t \; o \; C_{t-1} + i_t \; o \; tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \\ O_t &= \Phi(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} o C_{t-1} + b_i) \\ H_t &= O_t \; o \; tanh(C_t) \end{aligned} \tag{1}$$

where, the symbols $*$ indicate the convolution operation and o for element wise matrix multiplication. The parameter W and function Φ are the layer weights and non-linear activation function respectively.

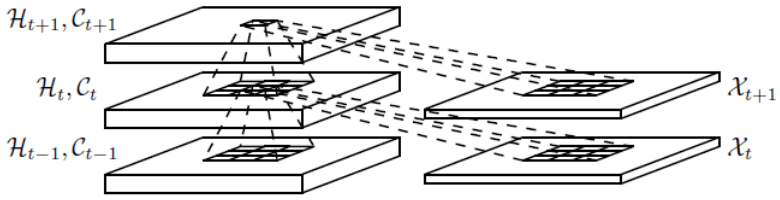


Figure 1. Convolutional recurrency (ConvLSTM) structure

The proposed model has employed the ConvLSTM-SE layers as its main module, with a structure shown in Figure 2b. A 5ms speech segment is processed at every time instance, and half magnitude spectra of 160 point short time Fourier transform (STFT) are fed into the network ($F = 81$). The model is trained to estimate the clean output STFT magnitude corresponding to the noisy input. The noisy phase is then used to reconstruct the enhanced output waveform. For this study, the ConvLSTM filters of size [1X3] with $D = 256$ channels are considered. The final FC matrix has a dimension of [256*81,81].

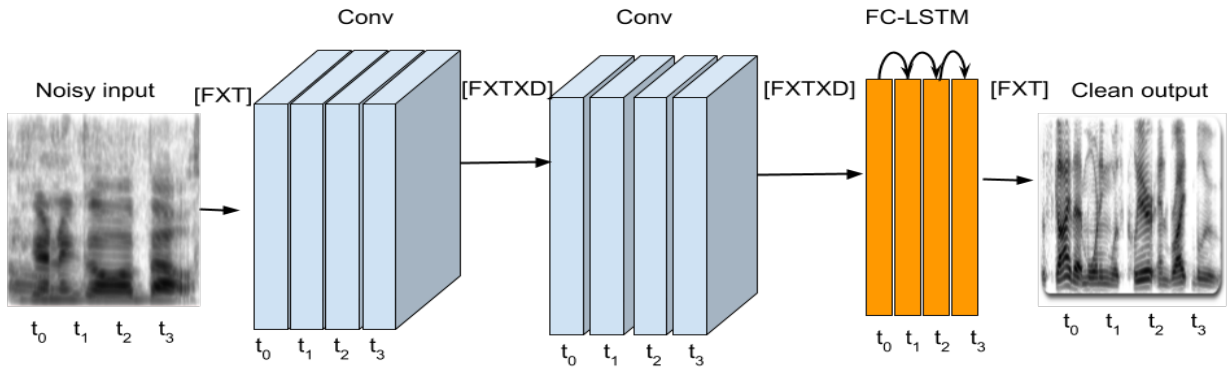
To compare with, a recent neural model is considered [3]. It has a series of convolution layers of depth $D = 256$ followed by a fully connected LSTM layer (Figure 2a). Both the models were trained in end-to-end on the NSDTSEA data set [6] containing a wide range of real-life additive noises. Two objective evaluation scores were used to evaluate the performance: perceptual evaluation of speech quality (PESQ) and short time objective intelligibility (STOI).

Table 1. Model performance comparison

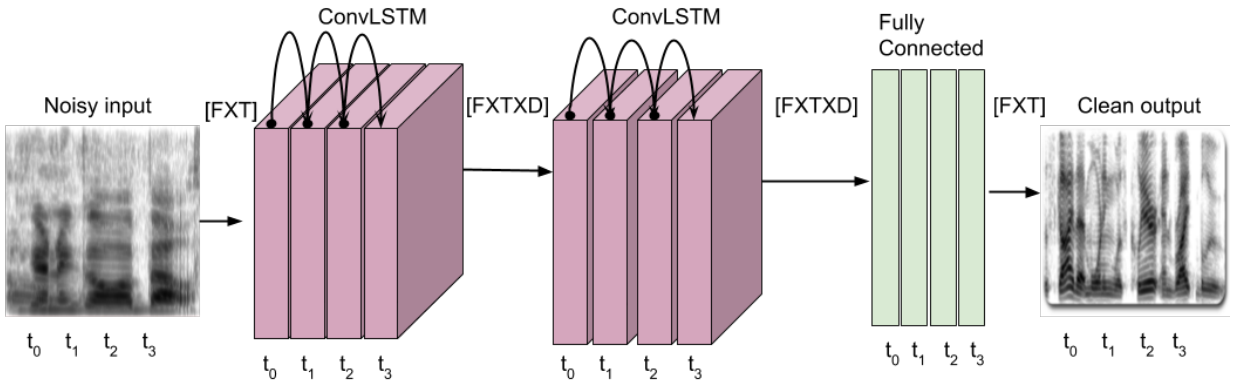
Metrics	Noisy	Naithani. et al. [3]	ConvLSTM-SE
PESQR	1.86	2.03	2.10
STOI	0.28	0.56	0.60
Parameters		12.3 Million	4.2 Million

3 DISCUSSION

The model performances of the proposed model are in Table 1. Both models show similar performance in terms of PESQ and STOI. However, the number of trained model parameters was reduced by 66%, from 12.3



(a) A low latency model, Naithani. et al. [3]



(b) Proposed ConvLSTM-SE model

Million to 4.2 Million. This is mainly attributed to the convolutional recurrency (ConvLSTM) placed at the initial layers, which had facilitated an efficient temporal modeling with minimum parameters.

4 CONCLUSION

In this work, we have proposed a neural network (NN) architecture using recurrent convolutional feature extraction, thereby facilitating efficient temporal sequence modeling for SE with fewer parameters. Relative to a low algorithmic-latency baseline method, the proposed ConvLSTM-SE has reduced the NN memory load by 66%, while maintaining the objective performance.

ACKNOWLEDGEMENTS

This work was partly funded by the E.U. Horizon2020 Grant Agreement 675324, Marie Sklodowska-Curie Innovative Training Network, ENRICH.

REFERENCES

[1] P. C. Loizou, Speech enhancement: theory and practice. press, 2007.

[2] A. Maas, Q. V. Le, T. M. Oneil, O. Vinyals, P. Nguyen, and A. Y. Ng, “Recurrent neural networks for

noise reduction in robust asr," 2012.

[3] Naithani, G., et al. "Low-latency sound source separation using convolutional recurrent deep neural networks." 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. 2017.

[4] Xingjian, S. H. I., et al. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting." Advances in neural information processing systems. 2015.

[5] Hartmann, Till S. "Seeing in the dark with recurrent convolutional neural networks." arXiv preprint arXiv:1811.08537 (2018).4

[6] Valentini-Botinhao, Cassia. "Noisy speech database for training speech enhancement algorithms and TTS models." University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR) (2017).