Incorporating Intra-Spectral Dependencies With A Recurrent Output Layer For Improved Speech Enhancement

Khandokar Md. Nayem and Donald S. Williamson
Department of Computer Science, Indiana University, USA

Speech Enhancement

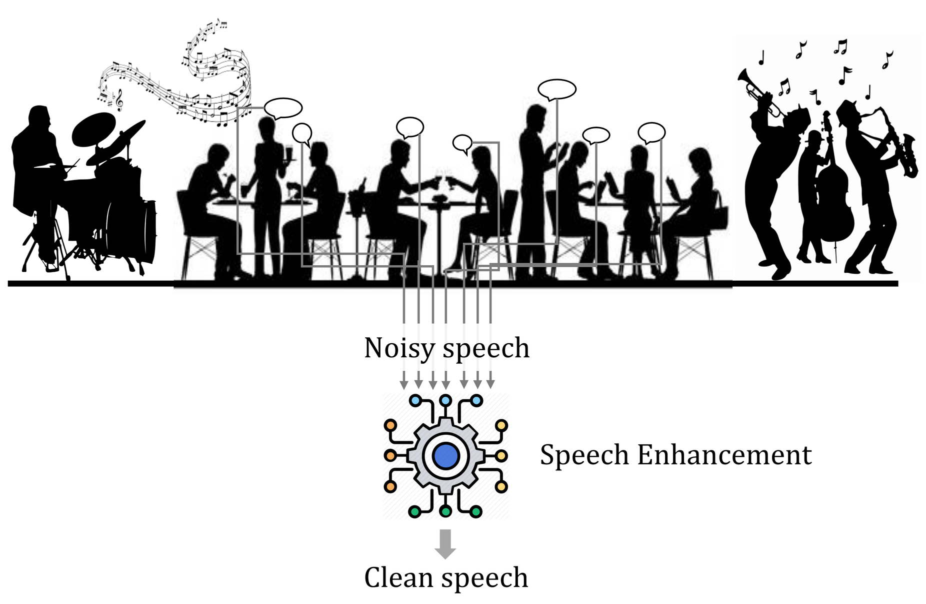


Image source, https://clipground.com/image-post/83621-black-people-restaurant-clipart-19.jpg.html#overlayGallery_post_83621_black-people-restaurant-clipart-19.jpg

Goal:
- Effectively remove background noise.
- Improve the quality and intelligibility of speech.
- Incorporate spectral-level dependencies within a single time frame.

Motivation

Related Work:
- Speech has spectral dependencies along the frequency axis [1].
- Current approaches use dedicated long-short term memory (LSTM) recurrent neural network (RNN) modules to learn spectral dependencies, at the sub-band frequency level or overall time [2, 3].
- Current approaches do not consider local spectral dependencies at adjacent or nearby frequencies over short-time instances.

Proposed Work:
- Develop a recurrent layer that captures frequency dependencies within each time frame.
- Capture temporal dependencies with an LSTM RNN.
- Conduct experiments to determine system robustness.

Notation

- In the time domain,

$$m_t = s_t + n_t \ ;$$

where $m_t \rightarrow$ noisy speech, $s_t \rightarrow$ clean speech, $n_t \rightarrow$ noise, $t \rightarrow$ time index

- In the time-frequency (T-F) domain,

$$M_{t,k} = |M_{t,k}| e^{i\theta_{M_{t,k}}} \ ;$$

where $M_{t,k} \rightarrow$ STFT of noisy speech, $|M_{t,k}| \rightarrow$ magnitude response, $\theta_{M_{t,k}} \rightarrow$ phase response, $k \rightarrow$ frequency index

- Estimation of the clean speech $\hat{S}_{t,k}$ can be predicted by,

$$|\hat{S}_{t,k}| = \boldsymbol{F}_{\boldsymbol{\phi}}(|M_{t,k}|)$$

$$\hat{S}_{t,k} = |\hat{S}_{t,k}| e^{i\theta_{M_{t,k}}}$$

where $|S_{t,k}| \rightarrow$ estimated clean magnitude, $F_{\phi}() \rightarrow$ estimation function with parameters ϕ

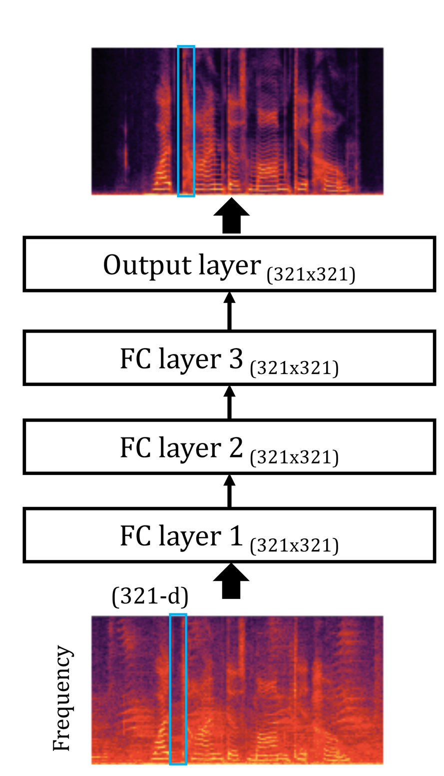
Baseline Deep Networks

Baseline DNN architecture for comparison:
- Each time frame of $|M_{t,k}|$ is the input and estimated $|\hat{S}_{t,k}|$ is the output of the network.
- Output of each layer a_t^l is computed by,

$$a_t^l = \sigma(V^l a_t^{l-1} + z^l)$$

where $l \rightarrow$ layer index, $\sigma \rightarrow$ activation function, V^l and z^l are weight and bias matrices, respectively.

- Uncorrelated (across time and frequency) outputs.
- Spectral output at each neuron does not depend on spectral outputs from other output-layer neurons.
- ReLU activation function, Adam optimizer, early stopping by validation set. [3]
- No pre-training and fine tuning steps.

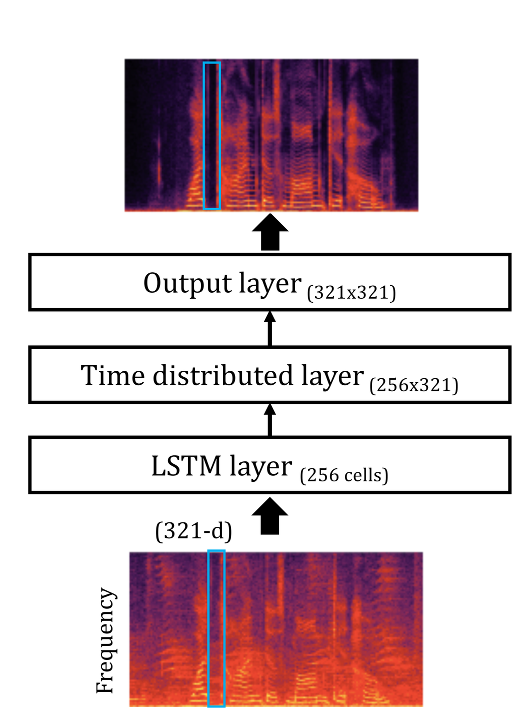


DNN structure

Baseline LSTM architecture for the proposed approach:
- Input and output are magnitude of the spectrogram same as baseline DNN.
- Output of each layer a_t^l is computed by,

$$f_t^l = \sigma_g(W_f^l a_t^{l-1} + U_f^l h_{t-1}^l + b_f^l)$$
$$i_t^l = \sigma_g(W_i^l a_t^{l-1} + U_i^l h_{t-1}^l + b_i^l)$$
$$o_t^l = \sigma_g(W_o^l a_t^{l-1} + U_o^l h_{t-1}^l + b_o^l)$$
$$c_t^l = f_t^l \circ c_{t-1}^l + i_t^l \circ \sigma_c(W_c^l a_t^{l-1} + U_c^l h_{t-1}^l + b_c^l)$$
$$h_t^l = o_t^l \circ \sigma_h(c_t^l)$$
$$a_t^l = \sigma_a(W_a^l h_t^l + b_a^l) \qquad \text{, where } l \in [1, L]$$

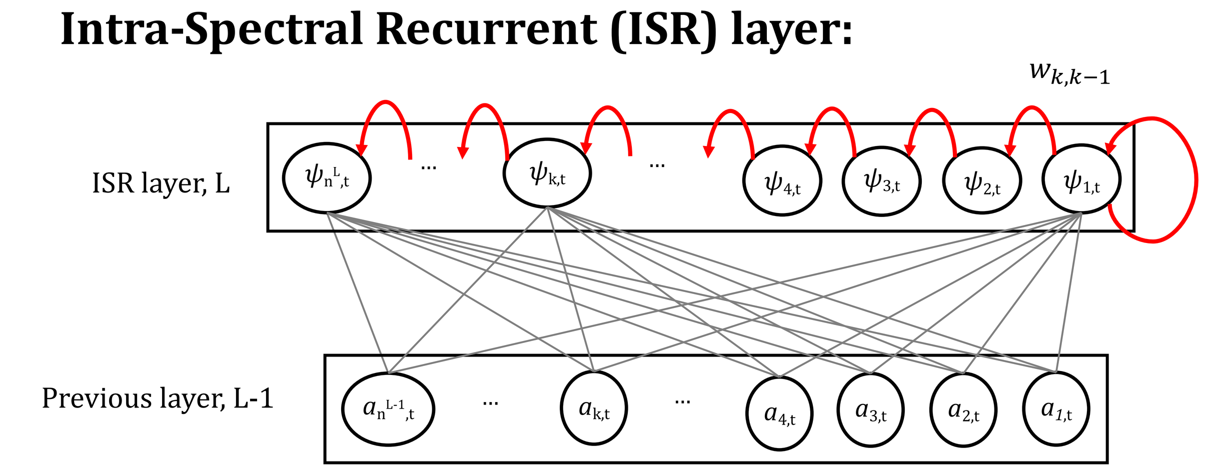
- Relationships across time frames are learned.
- Does not learn spectral relationships within the frequency axis.



LSTM structure

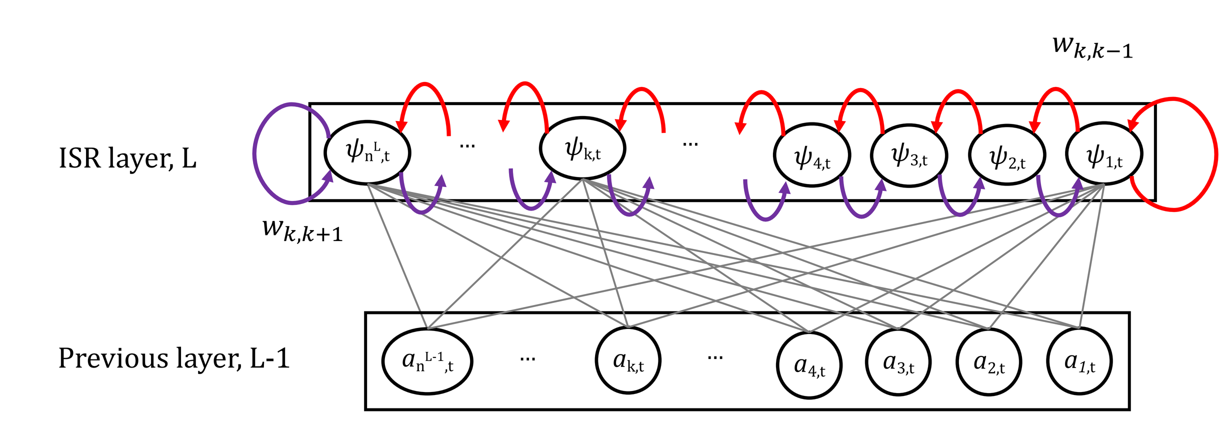
Proposed Intra–Spectral Output Layers

Intra-Spectral Recurrent (ISR) layer:



$$\Delta = \sigma(R^L a_t^{L-1} + \beta^L)$$
$$\psi_{1,t} = \Delta_1 + \sigma_\psi(w_{1,1} \times \psi_{1,t-1})$$
$$\psi_{k,t} = \Delta_k + \sigma_\psi(w_{k,k-1} \times \psi_{k-1,t})$$
$$\text{, where } k \in [2, n^L]$$

Intra-Spectral Bi-directional Recurrent (ISBR) layer:



$$\psi_{1,t} = \Delta_1 + \sigma_\psi(w_{1,2} \times \psi_{2,t}) + \sigma_\psi(w_{1,1} \times \psi_{1,t-1})$$
$$\psi_{n^L,t} = \Delta_{n^L} + \sigma_\psi(w_{n^L,n^L} \times \psi_{n^L,t-1}) + \sigma_\psi(w_{n^L,n^L-1} \times \psi_{n^L-1,t})$$
$$\psi_{k,t} = \Delta_k + \sigma_\psi(w_{k,k+1} \times \psi_{k+1,t}) + \sigma_\psi(w_{k,k-1} \times \psi_{k-1,t})$$
$$\text{,where } k \in [2, n^L - 1]$$

- Each neuron in the output layer corresponds to a frequency bin.
- Incorporate a first-order Markov assumption to learn spectral dependencies across frequencies (ISR, ISBR).
- A traditional LSTM network is first pre-trained, then a ISR/ISBR output layer replaces the original output layer.
- LSTM network learns the temporal dependencies and ISR/ISBR learns spectral dependencies.

Experiments and Results

- IEEE speech corpus consists of 720 utterances.
- Noise types: speech-shaped noise (SSN), cafeteria, factory, and babble.
- Trained in 3 SNR levels (-3, 0, 3 dB), tested in additional 2 SNR levels (-6 and 6 dB).
- Total training signals ~50 hrs, total validation signals ~11 hrs, total testing signals ~18.3 hrs.

Table: Average scores of the different models for seen SNRs (e.g. -3, 0, and 3 dB). Best results are shown in **bold**.

	PESQ				STOI				SI-SDR			
	SSN	Cafe	Factory	Babble	SSN	Cafe	Factory	Babble	SSN	Cafe	Factory	Babble
Mixture	1.95	1.86	1.83	1.77	0.71	0.62	0.65	0.59	-0.51	-2.06	-0.96	-1.97
DNN [4]	2.04	1.89	2.02	1.89	0.75	0.63	0.72	0.56	-1.75	-1.1	-1.4	-1.39
LSTM	2.12	1.97	2.05	1.95	0.77	0.64	0.76	0.62	-0.96	-1.35	-0.15	-0.44
D-ISR	2.24	2.08	2.26	2.08	0.85	**0.76**	0.86	**0.76**	-1.49	-2.91	-2.75	-3.48
L-ISR	2.27	2.21	2.29	2.11	0.82	0.68	0.84	0.72	0.06	-1.34	**0.17**	-1.3
L-ISBR	**2.3**	**2.24**	**2.31**	**2.13**	**0.88**	0.74	**0.87**	0.73	**2.35**	**-0.12**	-0.94	**-0.01**
L-FT[2]	2.12	2.01	2.07	2.04	0.82	0.74	0.82	0.66	1.04	-1.16	-0.88	-0.1

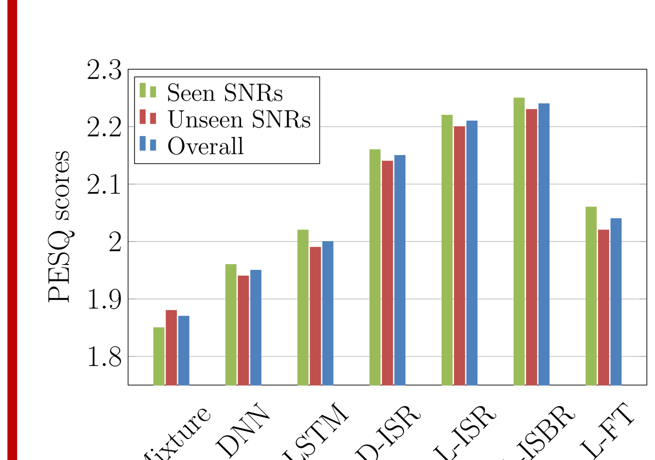


Fig: PESQ scores for seen and unseen SNR conditions.

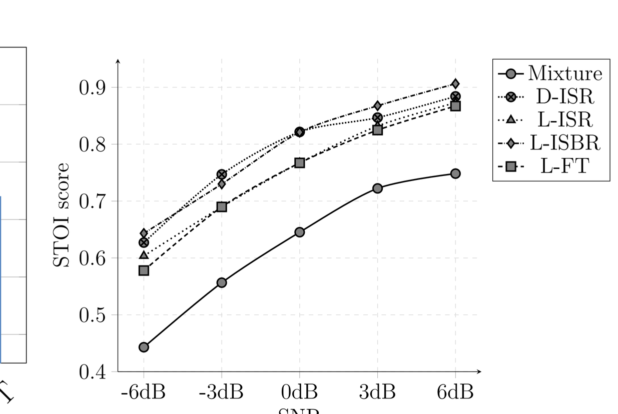


Fig: Average performance at each SNR for STOI.

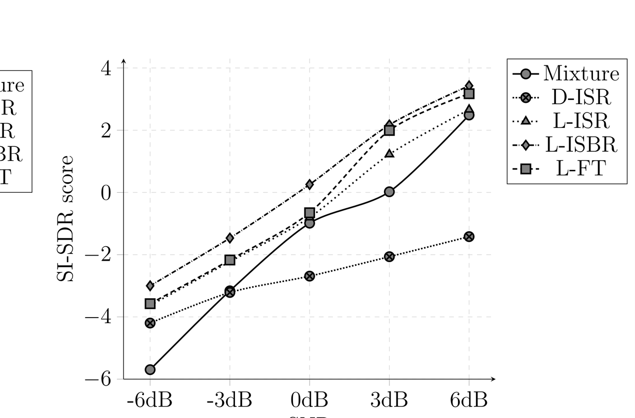


Fig: Average performance at each SNR for SI-SDR.

Conclusion and Future Works

- Improvements in a variety of noises and SNR values prove that the proposed ISR/ISBR layer along with a base LSTM network successfully captures both temporal and spectral correlations.
- Overall performance of LSTM network with ISR/ISBR layer (L-ISR/L-ISBR) shows the correlation between adjacent frequencies are important in the estimation of clean speech.
- Currently, mixture phase is used with enhanced speech magnitude.
- Phase-level dependencies will be intergraded in the future work.
- Spectral dependencies greater than first-order markov should be explored.

References

[1] T. F. Quatieri, Discrete-time speech signal processing: principles and practice. Upper Saddle River, NJ: Prentice Hall, 1st ed., 2002.

[2] J. Li, A. Mohamed, G. Zweig, and Y. Gong, "LSTM time and frequency recurrence for automatic speech recognition," in Proc. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 187–191, 2015.

[3] J. Deng, B. Schuller, F. Eyben, D. Schuller, Z. Zhang, H. Francois, and E. Oh, "Exploiting time-frequency patterns with lstm-rnns for low-bitrate audio restoration," Neural Computing and Applications, pp. 1–13, 2019.

[4] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," in Proc. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, pp 7-19, 2015.