Monaural Speech Enhancement Using Intra-Spectral Recurrent Layers In The Magnitude And Phase Responses

Khandokar Md. Nayem and Donald S. Williamson

Department of Computer Science, Indiana University, IN, USA

LUDDY
SCHOOL OF INFORMATICS,
COMPUTING, AND ENGINEERING

45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2020)

Motivation

Monaural speech enhancement is a challenging task.

The best performing deep architectures use LSTM recurrent neural networks (RNNs).

Underutilize or ignore spectral-level dependencies.

A deep learning architecture that leverages both temporal and spectral dependencies within the magnitude and phase responses.

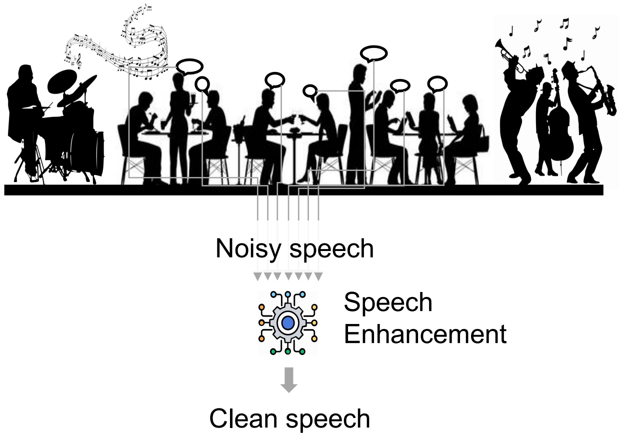


Image source, https://clipground.com/

Related work

Signal approximation
Weninger et al. 2014, Pascual et al. 2017 (Segan)

End-to-end waveform enhancement
Fu et al. 2018 utterance level optimization

Dedicated frequency LSTM modules
Li et al. 2015 in speech recognition, Deng et al. 2019 in audio restoration

Incorporating spectral-level dependencies
Nayem and Williamson 2019 for magnitude

Proposed approach

We propose an intra-spectral (e.g. across-frequency) recurrent layer as output layer that captures frequency dependencies within each time frame of a speech signal.

We train a base LSTM network to predict both the spectral-magnitude response and group delay, where the LSTM model captures temporal correlations.

We introduce Markovian recurrent connections in the output layers to capture spectral dependencies within the magnitude and phase responses.

Background & Notation

In the time domain, $m_t = s_t + n_t$

m_t → noisy speech
s_t → clean speech
n_t → noise
t → time index

In the time-frequency (T-F) domain, $M_{t,k} = \left|M_{t,k}\right| e^{i\theta^M_{t,k}}$

$M_{t,k}$ → T-F noisy speech
$|M_{t,k}|$ → magnitude response
$\theta^M_{t,k}$ → phase response
k → frequency index

$$S_{t,k} = |S_{t,k}| e^{i\theta^S_{t,k}}$$

$|S_{t,k}|$ → clean speech magnitude
$\theta^S_{t,k}$ → clean speech phase

$$N_{t,k} = |N_{t,k}| e^{i\theta^N_{t,k}}$$

$|N_{t,k}|$ → noise magnitude
$\theta^N_{t,k}$ → noise phase

Estimate clean speech, $\hat{S}_{t,k} = F_\phi(M_{t,k}) = F_\phi(|M_{t,k}|, \theta^M_{t,k}) \approx F_\phi(\left|S_{t,k}\right|, \left|N_{t,k}\right|, \theta^S_{t,k}, \theta^N_{t,k})$

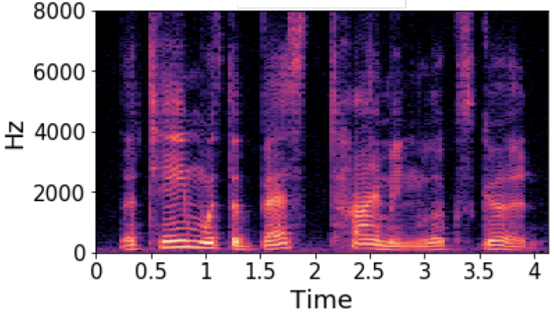
Background & Notation

Group delay (GD) of signal $S_{t,k}$, $GD^{S}_{t,k} = \angle e^{i(\theta^{S}_{t,k+1} - \theta^{S}_{t,k})}$

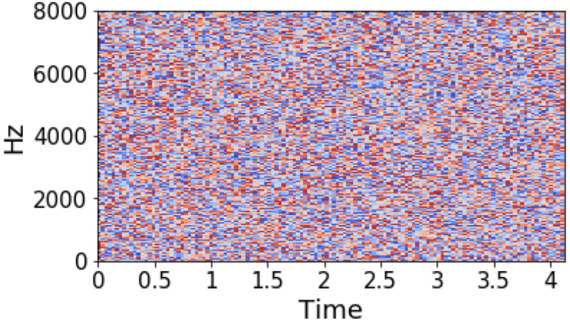
Unlike magnitude response, the phase of a speech does not show a clear structure.

Group delay of a speech shows a learn-able pattern in log-magnitude formulation.

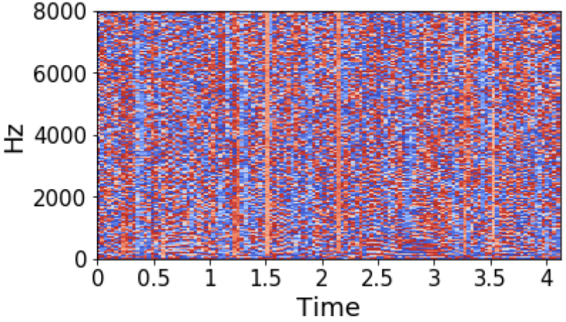
Clean magnitude



Clean phase



Clean group delay



Background & Notation

Optimal estimated magnitude loss function,

$$\mathcal{L}_{mag} = \Sigma_{t,k}\left(\left|\hat{S}_{t,k}\right| - \left|S_{t,k}\right|\right)^2 + \left(\left|\widehat{N}_{t,k}\right| - \left|N_{t,k}\right|\right)^2$$

Optimal estimated group delay loss function,

$$\mathcal{L}_{gd} = \sum_{\mathcal{X} \in \{S,N\}} \sum_{t,k} \left|\mathcal{X}_{t,k+1}\right| \frac{\left(1 - \cos\left(\widehat{GD}^{\mathcal{X}}{}_{t,k} - GD^{\mathcal{X}}_{t,k}\right)\right)}{2}$$

Optimal estimated magnitude and group delay loss function,

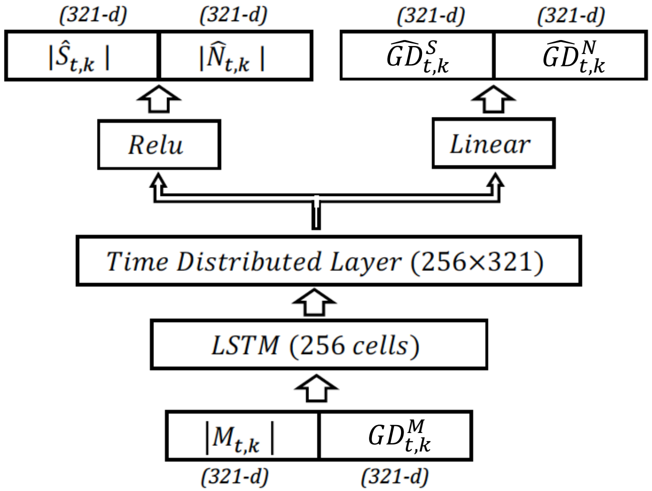
$$\mathcal{L}_{mag+gd} = \lambda\mathcal{L}_{mag} + (1-\lambda)\mathcal{L}_{gd}$$

Baseline LSTM Model

Clean speech $S_{t,k}$ and noise $N_{t,k}$ are considered as 2 separate sound sources $\mathcal{X}$.

LSTM model takes magnitude of the mixture $|M_{t,k}|$ and the group delay of the mixture $GD^{M}_{t,k}$ as inputs.

The output layer is branched in two ways, one is for magnitude approximation $|\widehat{\mathcal{X}}_{t,k}|$, and another is for GD approximation $\widehat{GD}^{\mathcal{X}}_{t,k}$ of both speech and noise.



Group Delay to Phase

Estimate phase difference between the enhanced speech and noise,

$$\hat{\delta}^{\mathcal{X}}_{t,k} = \left|\angle e^{i(\hat{\theta}^{\mathcal{X}}_{t,k} - \theta^{\neg\mathcal{X}}_{t,k})}\right| = \arccos\left(\mathcal{T}\left(\frac{|M_{t,k}|^2 + |\mathcal{X}_{t,k}|^2 - |\neg\mathcal{X}_{t,k}|^2}{2|M_{t,k}|\otimes|\mathcal{X}_{t,k}|}\right)\right)$$

where, $\mathcal{X} \in \{S, N\}$,
$\mathcal{T}(\cdot)$ → clip values to $[-1, 1]$
$\otimes$ → element-wise multiplication.

Sign of each T-F unit, $\hat{g}_{t,k} \in \{-1,1\}$ is calculated by,

$$\hat{g}_{t,1}, \dots, \hat{g}_{t,K} = \underset{g_{t,1},\dots,g_{t,K}}{\text{argmax}} \sum_{k} \sum_{\mathcal{X}\in\{S,N\}} cos\left(\hat{\theta}^{\mathcal{X}}_{t,k+1}(g_{t,k+1}) - \hat{\theta}^{\mathcal{X}}_{t,k}(g_{t,k}) - \widehat{GD}^{\mathcal{X}}_{t,k}\right) \tag{1}$$

By the formulation of trigonometric property of group delay,

$$\hat{\theta}^{\mathcal{X}}_{t,k}(g_{t,k}) = \theta^{M}_{t,k} + \gamma g_{t,k} \hat{\delta}^{\mathcal{X}}_{t,k} \tag{2}$$

$\gamma = 1$ when $\mathcal{X} = S$, and
$\gamma = -1$ when $\mathcal{X} = N$

Using dynamic programming within each t-f frame, we solve equations (1) and (2).

Recurrent Network limitation

Time unrolled recurrent network predicts i^{th} time frame conditioned on the frequency components of the $(i-1)^{th}$ time frame.

Goal is therefore to capture temporal influence, not spectral influence.

Spectral influence can be captured using a frequency unrolled recurrent network.

Localization of frequency perception suggests that a frequency component of i^{th} time depends on its neighbor frequency components of i^{th} time.

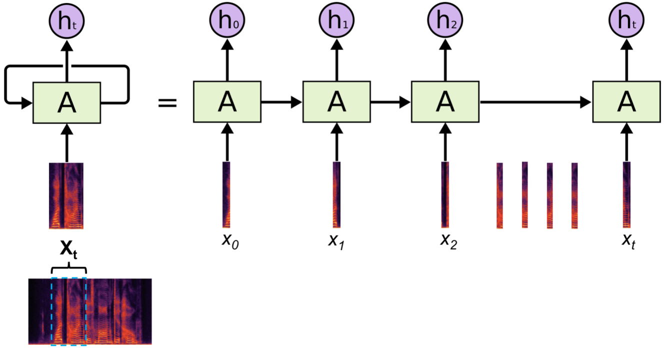
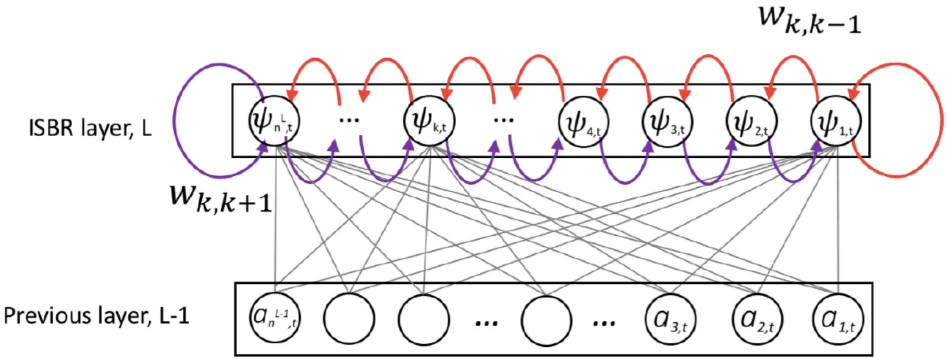


Image source, https://colah.github.io/

Intra-Spectral Bi-directional Recurrent (ISBR) layer



$$
\begin{aligned}
\boldsymbol{\Delta} &= \sigma(\boldsymbol{R}^L \boldsymbol{a}_t^{L-1} + \boldsymbol{\beta}^L) \\
\psi_{1,t} &= \Delta_1 + \sigma_\psi(w_{1,2} \times \psi_{2,t}) \\
&\quad + \sigma_\psi(w_{1,1} \times \psi_{1,t-1}) \\
\psi_{n^L,t} &= \Delta_{n^L} + \sigma_\psi(w_{n^L,n^L} \times \psi_{n^L,t-1}) \\
&\quad + \sigma_\psi(w_{n^L,n^L-1} \times \psi_{n^L-1,t}) \\
\psi_{k,t} &= \Delta_k + \sigma_\psi(w_{k,k+1} \times \psi_{k+1,t}) \\
&\quad + \sigma_\psi(w_{k,k-1} \times \psi_{k-1,t}), \quad k \in [2, n^L - 1]
\end{aligned}
$$

Each neuron of the ISBR layer represents a frequency bin of the signal.

Recurrent neurons from low to high frequencies and from high to low frequencies.

Spectral dependencies across both (increasing & decreasing) directions along frequency axis.

Intra-Spectral Bi-directional Recurrent (ISBR) layer

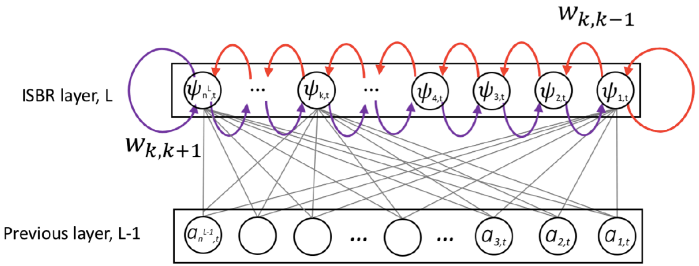
$\mathbf{\Delta}$ is the vector of activations, $\{\Delta_1, \ldots, \Delta_{n^L}\}$, based on inputs from the prior LSTM layer and $R^L$is the weight matrix, β^L is the bias vector.

$w_{k,k-1}$is the weight from $(k-1)^{st}$ to k^{th} frequency node.

σ and σ_ψ are the activation functions for the feed-forward and recurrent paths.

A LSTM network is first pre-trained, then an ISBR output layer replaces the original output layer.

LSTM network learns the temporal dependencies and ISBR learns spectral dependencies.



$$
\begin{aligned}
\mathbf{\Delta} &= \sigma(\boldsymbol{R}^L \boldsymbol{a}_t^{L-1} + \boldsymbol{\beta}^L) \\
\psi_{1,t} &= \Delta_1 + \sigma_\psi(w_{1,2} \times \psi_{2,t}) \\
&\quad + \sigma_\psi(w_{1,1} \times \psi_{1,t-1}) \\
\psi_{n^L,t} &= \Delta_{n^L} + \sigma_\psi(w_{n^L,n^L} \times \psi_{n^L,t-1}) \\
&\quad + \sigma_\psi(w_{n^L,n^L-1} \times \psi_{n^L-1,t}) \\
\psi_{k,t} &= \Delta_k + \sigma_\psi(w_{k,k+1} \times \psi_{k+1,t}) \\
&\quad + \sigma_\psi(w_{k,k-1} \times \psi_{k-1,t}), \quad k \in [2, n^L - 1]
\end{aligned}
$$

Dataset

Train model with IEEE male (single speaker) corpus.

Evaluate using IEEE male (720 utterances) and TIMIT (multiple speaker, 6300 utterances) speech corpus.

4 Noise types- speech-shaped noise (SSN), cafeteria, factory, and babble.

Trained and validated in 3 SNR levels (-3, 0, 3 dB).
Total training signals, 60000 (500 utterances) ~ 50 hrs.
Total validation signals, 13200 (110 utterances) ~ 11 hrs.

Tested in additional 2 SNR levels (-6 and 6 dB).
Total testing signals, 22000 (110 utterances) ~ 18.3 hrs.

Experimental Setup

LSTM model

Single LSTM layer with 256 units and a time-distributed layer with 321 units.

4 separate output layers (one for each target) in parallel.

ReLU activation function is used for the two output layers that predict magnitude spectrograms.

Linear activation function is used in the dense layer and for predicting group-delay.

Adam optimizer, early stopping by validation set.

LSTM-ISBR (ISBR) model

Trained LSTM network.

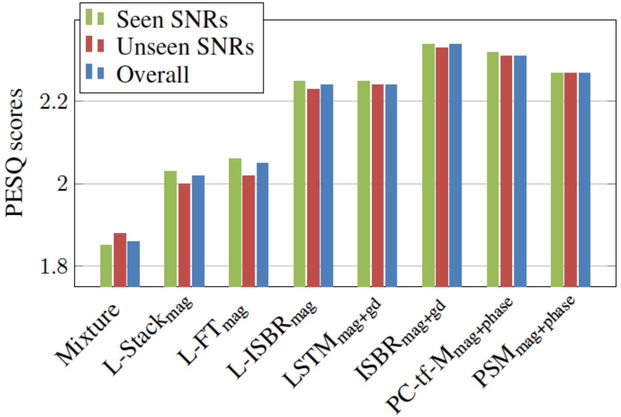
Output layer of the LSTM network is replaced by ISBR layer, and the model is retrained.

Results

Table: Average scores for each approach. Best results are shown in bold.

	IEEE corpus			TIMIT corpus		
	PESQ	STOI	SI-SDR	PESQ	STOI	SI-SDR
Mixture	1.86	0.62	-1.47	1.58	0.51	-2.33
L-Stack$_{\text{mag}}$ [1]	2.02	0.59	-0.59	1.82	0.5	-0.84
L-FT$_{\text{mag}}$ [2]	2.05	0.6	-0.2	1.88	0.52	-0.26
L-ISBR$_{\text{mag}}$ [3]	2.24	0.64	0.22	1.93	0.52	-0.03
LSTM$_{\text{mag+gd}}$	2.24	0.64	0.12	1.97	0.53	-0.1
ISBR$_{\text{mag+gd}}$	**2.34**	**0.67**	**0.92**	**2.04**	**0.58**	**0.84**
PC-tf-M$_{\text{mag+phase}}$ [4]	2.31	**0.67**	0.85	**2.04**	**0.58**	0.72
PSM$_{\text{mag+phase}}$ [5]	2.27	0.65	0.4	2	0.56	0.32

Fig: PESQ scores for seen, unseen and overall SNR conditions for the IEEE corpus.



[1] J. Deng et al., “Exploiting time-frequency patterns with lstm-rnns for low-bitrate audio restoration,” Neural Computing and Applications, pp. 1–13, 2019.
[2] J. Li, A. Mohamed, G. Zweig, and Y. Gong, “LSTM time and frequency recurrence for automatic speech recognition,” in Proc. ASRU, pp. 187–191, 2015.
[3] K. M. Nayem and D. S. Williamson, “Incorporating intra-spectral dependencies with a recurrent output layer for improved speech enhancement,” in Proc. MLSP, 2019.
[4] J. Lee and H.-G. Kang, “A joint learning algorithm for complex-valued tf masks in deep learning-based single channel speech enhancement systems,” IEEE/ACM TASLP, vol. 27, pp. 1098–1109, 2019.
[5] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks,” in Proc. ICASSP, pp. 708–712, 2015.

Conclusion

Our approach outperforms the T-F masking approaches, which indicates that incorporating spectral-level magnitude and phase dependencies are beneficial.

Proposed ISBR layer can used as output layer on top of any state-of-the-art model.

Only the first-order Markovian assumption considered in ISBR layer.

Explore higher-order spectral dependencies along with sub-band spectral dependencies in a single time frame.

Thank You

Khandokar Md. Nayem

knayem@iu.edu

Donald S. Williamson

williads@indiana.edu

ASPIRE Research Group, https://aspire.sice.indiana.edu/