Dereverberation of Speech Using Autoregressive Models of Sub-band Envelopes

Anurenjan P. R.

Advisor : Sriram Ganapathy
Learning and Extraction of Acoustic Patterns lab
Electrical Engineering Department
Indian Institute of Science, Bangalore.

29th September 2023

Outline

Introduction

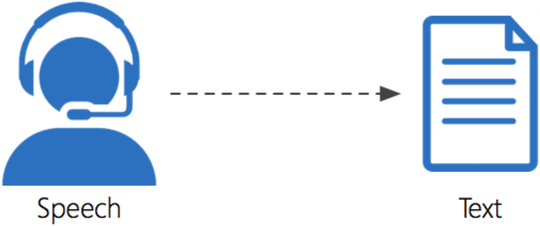


Figure: Automatic Speech Recognition (ASR) is the process of deriving the transcription (word sequence) of an utterance, given the speech waveform.

Introduction

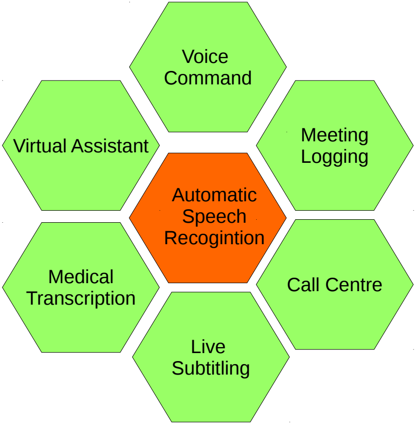


Figure: Application areas of ASR.

Far-field Speech Recognition

Figure: Talker and the microphone are far apart.

Far-field Speech Recognition - Real world Issues...

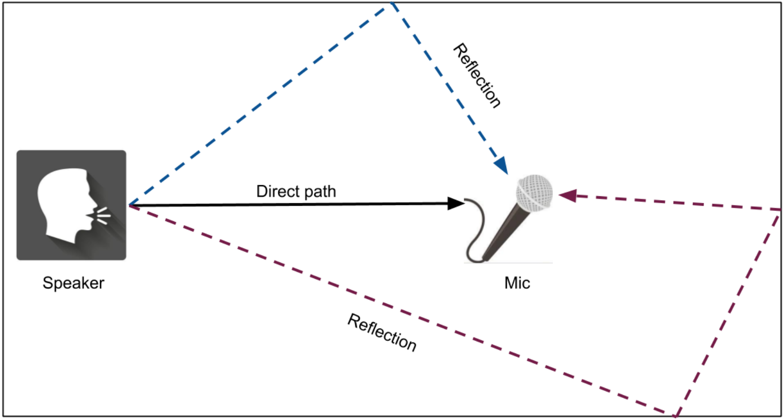


Figure: Degradation due to multipath signals - Reverberation.

Reverberation model

- Signal received at the microphone is modelled as follows,

$$r(t) = x(t) * h(t)$$

where $x(t)$, $h(t)$ and $r(t)$ denote the clean speech signal, the room impulse response and the reverberant speech, respectively.

- The room response function $h(t)$ can further be split into,

$$h(t) = h_e(t) + h_l(t)$$

where $h_e(t)$ and $h_l(t)$ represent the early and late reflection components.

Reverberation - Impact on applications

- The degradation of the automatic speech recognition (ASR) systems in presence of noise and reverberation is a challenging problem due to the low signal to noise ratio.
- Application like speaker recognition is also affected by reverberation.
- Can severely degrade speech intelligibility for human listeners, especially for hearing-impaired listeners.

Speech enhancement for far-field ASR

- In most of the present day ASR systems the multi-channel far-field recording will be going through a set of pre-processing/enhancement steps before feature extraction.
- The usual pipe line is to do a dereverberation on all the channels using a method like weighted prediction error (WPE).
- To do a beamforming step where all the available channels are combined to form a single channel.

Contributions of the work

- This work is focused on developing frequency domain linear prediction (FDLP) based systems for dereverberation and enhancement of speech signals.
- This work pursues two broad directions for addressing issues in far-field speech.
- In the first part of the talk, two methods for addressing reverberation is discussed.
- In the second part of the talk, we discuss a speech enhancement model using temporal envelopes and corresponding carriers.

Motivation

- In the traditional setting, the first step in the analysis of a signal is the short-term Fourier transform (STFT).
- The key assumptions about the convolution model of reverberation artifacts, is applicable for a long-analysis window in the time domain, or using convolutional transfer function with cross-band filters in the STFT domain.
- In our case, we use the former approach of long analysis window and explore dereverberation in the sub-band envelope domain.
- As the reverberation is a long-term convolution effect, we highlight that room impulse response (typically with a T60 $> 400ms$) can be absorbed as a multiplication in the frequency domain, as well as a convolution in the sub-band envelope domain.

BACKGROUND

Frequency Domain Linear Prediction (FDLP)

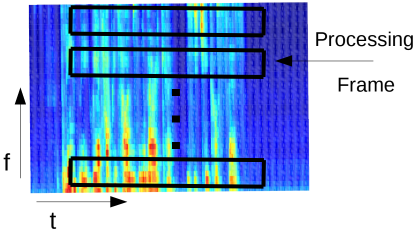
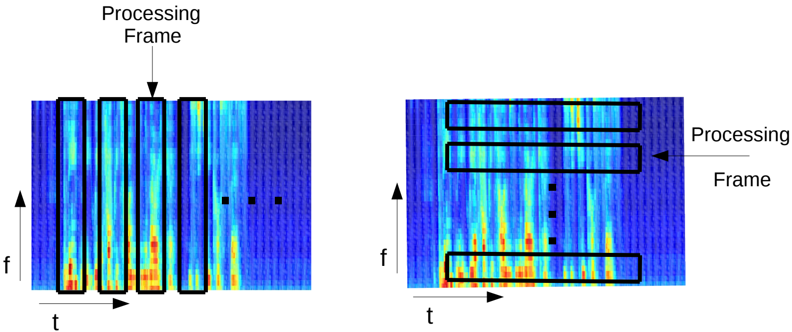


Figure: Processing happens in sub-band signals of longer duration, typically in seconds.

- FDLP is the frequency domain dual of Time Domain Linear Prediction (TDLP).
- Just as TDLP estimates the spectral envelope of a signal, FDLP estimates the temporal envelope of the signal.[1]

[1] Samuel Thomas, Sriram Ganapathy, and Hynek Hermansky. "Recognition of reverberant speech using frequency domain linear prediction". In: *IEEE Signal Processing Letters* 15 (2008), pp. 681–684.

Frequency Domain Linear Prediction (FDLP)



- In TDLP, the processing happens in small 20-30ms duration of time windows.
- In FDLP, long temporal duration of sub-band signals, typically 1-2 sec. duration are analysed.

Frequency Domain Linear Prediction

- FDLP estimates the temporal envelope of the signal, i.e. square of its Hilbert envelope
- Temporal envelope is given by the inverse Fourier transform of the auto-correlation function of DCT,

$$e(t) = F^{-1}\left\{Autocorr(y[k])\right\}$$

- where $y[k]$ is the DCT of a signal $x[n]$ having N- points.

Frequency Domain Linear Prediction

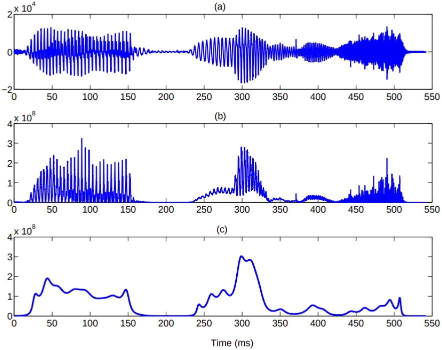


Figure: Illustration of the AR modeling property of FDLP. (a) a portion of speech signal, (b) its Hilbert envelope and (c) envelope obtained from all pole modeling in frequency domain.

Figure Credit:[2]

[2] Sriram Ganapathy. "Signal analysis using autoregressive models of amplitude modulation". PhD thesis. Johns Hopkins University, 2012.

Feature Extraction - Frequency Domain Linear Prediction based...

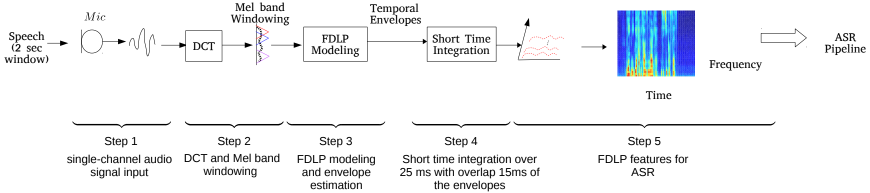


Figure: Typical process flow in FDLP based feature extraction for ASR.

FDLP based dereverberation (Part - 1)

Feature Extraction using Spactio-Spectral Autoregressive Modeling (SSAR)

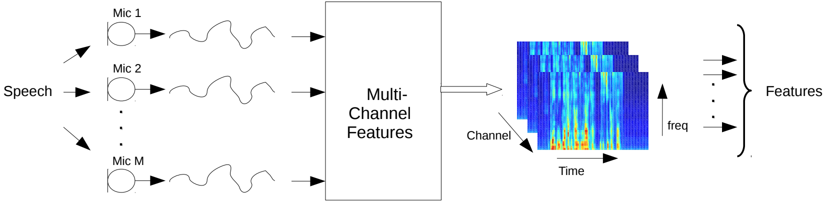


Figure: Features are extracted from all the channels **jointly** without beamforming step.

Feature Extraction using Spatio-Spectral Autoregressive Modeling (SSAR)

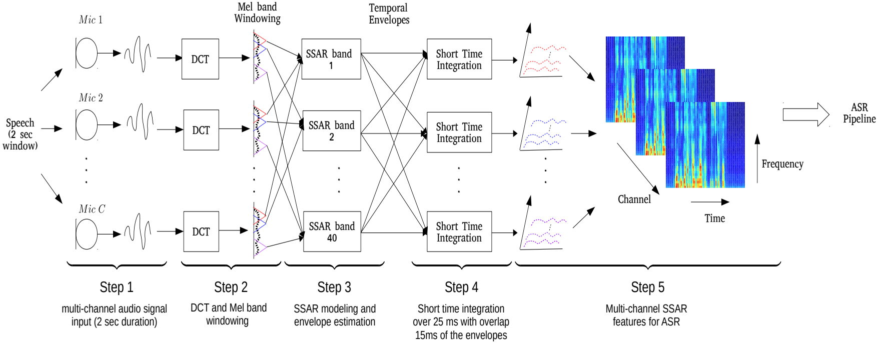


Figure: Typical process flow in SSAR modeling based feature extraction for ASR.

SSAR - Comparision of Spectrograms...

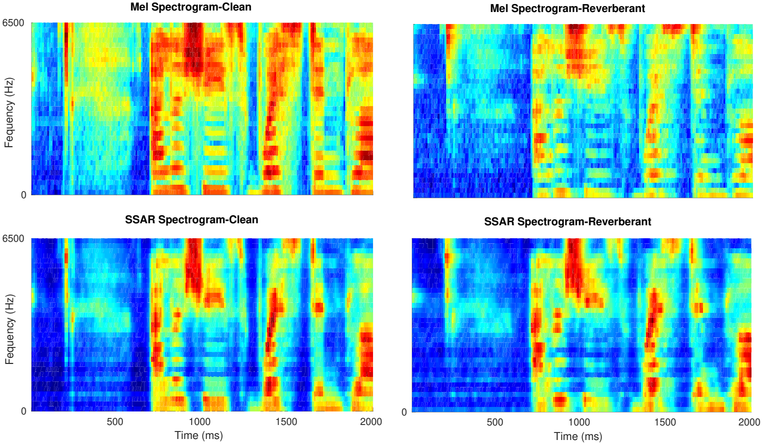


Figure: Comparison of spectrogram estimation using SSAR modeling with conventional mel spectrogram for clean (near-room) and reverberant speech (far-room) recordings from the REVERB Challenge dataset.

Experiments and results

Baseline ASR setup

- Classical delay sum beamforming with filter bank features (BF-FBANK) was used as the features.
- A 2-D acoustic model (AM)[3] was used to perform ASR.

[3] Anurenjan Purushothaman, Anirudh Sreeram, and Sriram Ganapathy. "3-d acoustic modeling for far-field multi-channel speech recognition". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 6964–6968.

Experiments and results

Dataset

- REVERB challenge dataset
  - Consists of 8 channel recordings with British English.
  - The training set consists of 17 hours of audio.
  - The Dev and Eval datasets consists of roughly 3 hours and 6 hours of audio recordings, respectively.

Experiments and results

Table: Word Error Rate (%) in REVERB dataset.

Experiments	Dev			Eval		
	Real	Simu	Avg	Real	Simu	Avg
BF-FBANK (2-D AM)	19.7	6.2	12.9	22.2	6.5	14.4
MC-FBANK (3-D AM)	20.4	6.7	13.5	21.2	6.6	13.9
SSAR (3-D AM)	18.6	6.4	**12.5**	20.5	6.8	**13.6**

Dereverberation of temporal envelopes for far-field ASR

- Dereverberation of the autoregressive estimates of the sub-band envelope using a CLSTM model followed by feature extraction for ASR.

Dereverberation of temporal envelopes for far-field ASR

- Dereverberation of the autoregressive estimates of the sub-band envelope using a CLSTM model followed by feature extraction for ASR.
- Deriving a signal model for reverberation effects on sub-band speech envelopes and posing the dereverberation problem as a gain estimation problem.

Dereverberation of temporal envelopes for far-field ASR

- Dereverberation of the autoregressive estimates of the sub-band envelope using a CLSTM model followed by feature extraction for ASR.
- Deriving a signal model for reverberation effects on sub-band speech envelopes and posing the dereverberation problem as a gain estimation problem.
- Joint learning of the dereverberation model parameters and the acoustic model for ASR in a single neural pipeline.

Signal model

- The speech data recorded by a far-field microphone is modeled as,

$$r(n) = x(n) * h(n)$$

where $x(n)$, $h(n)$ and $r(n)$ are the source speech signal, the room impulse response function and the far-field speech, respectively.

Signal model

- The speech data recorded by a far-field microphone is modeled as,

$$r(n) = x(n) * h(n)$$

  where $x(n)$, $h(n)$ and $r(n)$ are the source speech signal, the room impulse response function and the far-field speech, respectively.
- The room response function can be further expanded as,

$$h(n) = h_e(n) + h_l(n)$$

  where $h_e(n)$ and $h_l(n)$ represent the early and late reflection components.

Signal model

- Let $x_q(n)$, $h_q(n)$ and $r_q(n)$ denote the sub-band clean speech, room-response and the reverberant speech, respectively for the q^{th} sub-band.

Signal model

- Let $x_q(n)$, $h_q(n)$ and $r_q(n)$ denote the sub-band clean speech, room-response and the reverberant speech, respectively for the q^{th} sub-band.
- The sub-band envelopes of far-field speech $m_{rq}(n)$, extracted using frequency domain linear prediction (FDLP), can be approximated as, [4]

$$m_{rq}(n) \approx \frac{1}{2} m_{xq}(n) * m_{hq}(n)$$

where, $m_{xq}(n)$, $m_{hq}(n)$ denote the sub-band envelope of the clean source signal and the room impulse response function, respectively.

[4] Anurenjan Purushothaman et. al, "Dereverberation of autoregressive envelopes for far-field speech recognition", in Journal of Computer Speech and Language 2022.

Signal model

- Given this envelope convolution model, we can further split the far-field speech envelope into early and late reflection components.

$$m_{rq}(n) = m_{rqe}(n) + m_{rql}(n)$$

where $m_{rqe}(n)$ and $m_{rql}(n)$ denote the sub-band envelopes of early and late reflection parts.

Envelope dereverberation model

- The envelope dereverberation model tries to subtract the late reflection components $m_{rql}(n)$ from reverberant sub-band temporal envelope $m_{rq}(n)$.

Envelope dereverberation model

- The envelope dereverberation model tries to subtract the late reflection components $m_{rql}(n)$ from reverberant sub-band temporal envelope $m_{rq}(n)$.
- The sub-band envelope residual (targets for the neural model) is the log-difference of the sub-band envelope for the direct components and the sub-band envelope of the reverberant sub-band signal.

Envelope dereverberation model

- The envelope dereverberation model tries to subtract the late reflection components $m_{rql}(n)$ from reverberant sub-band temporal envelope $m_{rq}(n)$.
- The sub-band envelope residual (targets for the neural model) is the log-difference of the sub-band envelope for the direct components and the sub-band envelope of the reverberant sub-band signal.
- The neural model is trained with reverberant sub-band envelopes ($\log(m_{rq}(n))$ as input and model outputs the gain (in the log domain this is $\log\frac{m_{xq}(n)}{m_{rq}(n)}$), which when multiplied with the reverberant envelopes (additive in log domain), generates the estimate of source signal envelope ($\log(\hat{m}_{xq}(n))$.

Envelope dereverberation model

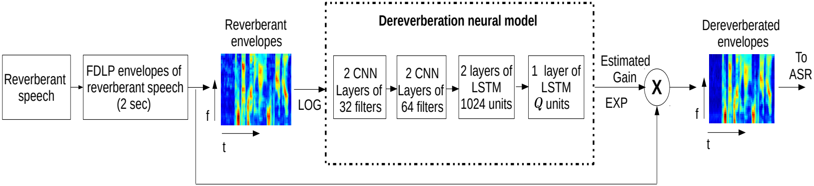


Figure: Block schematic of the proposed envelope dereverberation model[5].

[5] Anurenjan Purushothaman et al. "Deep Learning Based Dereverberation of Temporal Envelopes for Robust Speech Recognition". In: *Proc. Interspeech 2020*. 2020, pp. 1688–1692. DOI: 10.21437/Interspeech.2020-2283. URL: http://dx.doi.org/10.21437/Interspeech.2020-2283.

Experiments and results

Table: Word Error Rate (%) in REVERB dataset for different features and proposed dereverberation method.

Model Features	Dev			Eval		
	Real	**Simu**	**Avg**	**Real**	**Simu**	**Avg**
WPE-BF-FBANK	19.1	6.1	12.6	14.7	**6.5**	10.6
WPE-BF-FDLP	17.8	6.8	12.3	14.0	7.0	10.5
WPE-BF-FDLP + derevb.	**16.3**	**5.6**	**10.9**	**13.4**	7.1	**10.2**

Joint Learning

- Encouraged by the results obtained by the envelop dereverberation, we further wanted to improve the dereverberation model with the ASR cost.
- By jointly training the pre-trained dereverberation and pre-trained ASR with the ASR cost function, we further find improvements.
- The Joint training is performed for 3 iterations on the given data.

Joint Learning

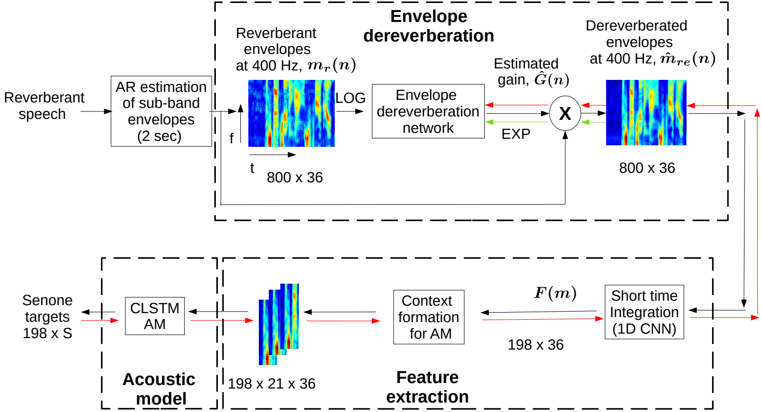


Figure: Block schematic of joint learning framework, the entire model can be constructed as an end-to-end neural framework. The black arrows denote the forward pass, the red arrows represent backward propagation with ASR loss, and green arrows denote the backward propagation with mean square error loss.

Experiments and results

Table: Word Error Rate (%) in REVERB dataset for different features and proposed dereverberation method.

Model Features	Dev			Eval		
	Real	**Simu**	**Avg**	**Real**	**Simu**	**Avg**
WPE-BF-FBANK	19.1	6.1	12.6	14.7	**6.5**	10.6
WPE-BF-FDLP	17.8	6.8	12.3	14.0	7.0	10.5
WPE-BF-FBANK + derevb.	17.3	5.5	11.4	13.1	6.9	10.0
WPE-BF-FBANK + derevb.[6]	15.8	**5.2**	10.5	12.8	6.7	9.8
WPE-BF-FBANK + derevb.[7]	19.6	6.9	13.3	17.5	9.0	13.3
WPE-BF-FDLP + derevb.	16.3	5.6	10.9	13.4	7.1	10.2
WPE-BF-FDLP + derevb. + joint.	**15.2**	5.6	**10.4**	**12.1**	7.1	**9.6**

[6] K. Han etal, "Learning Spectral Mapping for Speech Dereverberation and Denoising", IEEE/ACM TASLP 2015

[7] J. F. Santos etal, "Speech Dereverberation With Context-Aware Recurrent Neural Networks", IEEE/ACM TASLP 2018

Dereverberation of temporal envelopes - extension to E2E ASR

- Here, we extend the work done on HMM-DNN hybrid ASR to E2E ASR systems.

[8] Rohit Kumar et al. "End-To-End Speech Recognition with Joint Dereverberation of Sub-Band Autoregressive Envelopes". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 6057–6061.

Dereverberation of temporal envelopes - extension to E2E ASR

- Here, we extend the work done on HMM-DNN hybrid ASR to E2E ASR systems.
- We also investigate the effect to new regularizing loss functions.

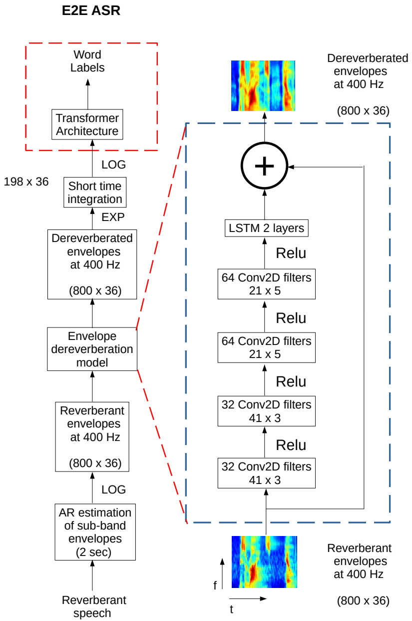
[8] Rohit Kumar et al. "End-To-End Speech Recognition with Joint Dereverberation of Sub-Band Autoregressive Envelopes". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 6057–6061.

Dereverberation of temporal envelopes - extension to E2E ASR

- Here, we extend the work done on HMM-DNN hybrid ASR to E2E ASR systems.
- We also investigate the effect to new regularizing loss functions.
- We propose a far-field E2E ASR system, where a joint learning of enhancement model and the E2E ASR model is done[8].

[8] Rohit Kumar et al. "End-To-End Speech Recognition with Joint Dereverberation of Sub-Band Autoregressive Envelopes". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 6057–6061.

Dereverberation of temporal envelopes for E2E ASR



Experiments and results

Table: WER (%) in REVERB dataset for separate learning of the dereverberation and E2E models as well as the joint learning.

Model Config.	Dev			Eval		
	Real	Sim	Avg	Real	Sim	Avg
WPE-BF-FBANK (baseline)	15.3	10.5	12.9	11.5	9.2	10.4
WPE-BF-FDLP	14.1	6.7	10.4	10.1	6.5	8.3
- + derevb. [MSE]	11.4	7.7	9.5	9.5	7.0	8.2
- + derevb. [MSE+BEGAN]	11.3	7.5	9.4	8.7	6.6	7.6
- + joint. [MSE]	10.3	6.3	8.3	**7.1**	**5.6**	**6.3**
- + joint. [MSE+BEGAN]	9.3	6.1	**7.7**	7.7	5.9	6.8

Summary - Part 1

- In this work, we propose a new framework of multi-channel features using SSAR modeling and a 3-D acoustic model for neural beamforming.

Summary - Part 1

- In this work, we propose a new framework of multi-channel features using SSAR modeling and a 3-D acoustic model for neural beamforming.
- We propose a new neural model for dereverberation of temporal envelopes.

Summary - Part 1

- In this work, we propose a new framework of multi-channel features using SSAR modeling and a 3-D acoustic model for neural beamforming.
- We propose a new neural model for dereverberation of temporal envelopes.
- Further the neural model was jointly trained with the acoustic model to improve the ASR cost.

Summary - Part 1

- In this work, we propose a new framework of multi-channel features using SSAR modeling and a 3-D acoustic model for neural beamforming.
- We propose a new neural model for dereverberation of temporal envelopes.
- Further the neural model was jointly trained with the acoustic model to improve the ASR cost.
- We extended the previously proposed dereverberation method to E2E ASR framework.

Speech enhancement with FDLP (Part - 2)

Speech enhancement with FDLP

- We decompose the sub-band speech signal into the constituent envelope and carrier part.
- A dereverberation neural model is designed that attempts to enhance the envelope and carrier signals jointly.
- Further, joint learning of the speech enhancement model with the end-to-end ASR model is proposed with a single neural framework.

Envelope-carrier decomposition

- We use a uniform 64-band Quadrature Mirror Filter bank (QMF) for decomposing the input signal, $x[n]$ into 64 uniformly spaced frequency bands.
- This is a perfect reconstruction analysis/synthesis filter bank.
- We use FDLP to estimate the envelope of the sub-band signal $x_q[n]$.

Envelope-carrier decomposition cont...

- Let $e_q[n]$ represent the sub-band temporal envelope found out by FDLP.
- The corresponding carrier (remaining residual signal), $c_q[n]$ is found by point wise division of the signal $x_q[n]$ by the estimated envelope $e_q[n]$,

$$c_q[n] = x_q[n]/e_q[n], \qquad \forall\ n \in \mathbb{Z} \tag{1}$$

Enhancement with envelope-carrier

- Apart from independently enhancing either the envelope or the carrier, we can learn the mapping between clean and reverberant versions of both the envelope and the carrier in parallel.
- By feeding a neural model with the reverberant envelope concatenated with the corresponding carrier, the network is trained to output the late reflection components.
- The joint learning of the envelope-carrier enhancement module and the E2E ASR architecture is done by combining the two separate models and training it jointly.

Enhancement with envelope-carrier

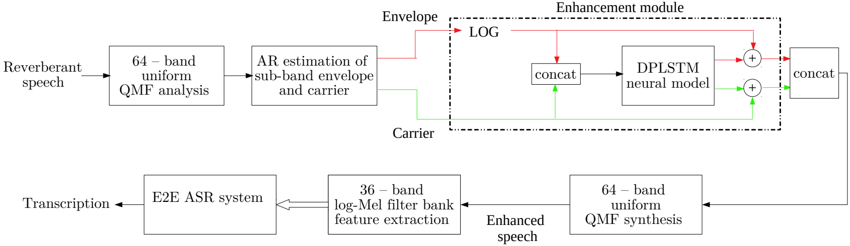


Figure: Block schematic of speech enhancement model, the feature extraction module and the E2E ASR model. Red arrows denote the envelopes, $e[n]$, the green arrows represent the carrier, $c[n]$. The entire model can be constructed as an end-to-end neural framework.

Dual path LSTM (DPLSTM)

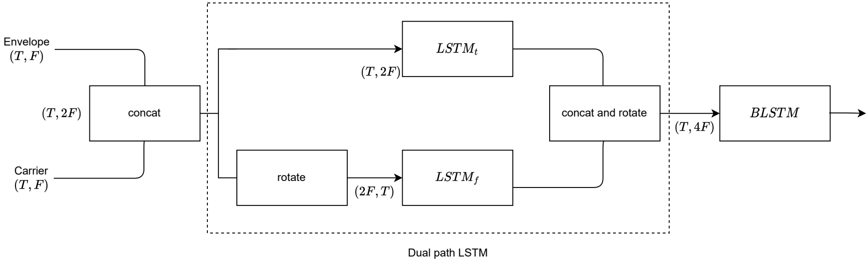


Figure: The dual path LSTM model architecture for speech enhancement.

Experiments and results

Table: WER (%) in REVERB dataset for separate learning of the dereverberation and E2E models as well as the joint learning.

Model Config.	**Dev**			**Eval**		
	Real	**Sim**	**Avg**	**Real**	**Sim**	**Avg**
WPE-BF-FBANK (baseline)	12.8	8.7	10.8	11.9	7.9	9.9
WPE-BF-FBANK- + env. derevb.	12.7	8.5	10.6	10.1	7.8	9
WPE-BF-FBANK- + crr. derevb.	11.2	8.3	9.8	10.8	7.6	9.2
WPE-BF-FBANK- + env. & crr. derevb.	10.6	7.6	9.1	9.1	6.9	8
- + joint.	9.4	6.4	**7.9**	7.3	5.7	**6.5**

Experiments and results

Table: SRMR and PESQ values (%) in REVERB dataset for envelope and carrier based enhancements.

	SRMR					PESQ
	Dev. Real	Dev. Simu	Eval. Real	Eval. Simu	REVERB Tr_cut	REVERB Tr_cut
WPE-BF-FBANK (baseline)	5.35	4.2	4.61	4.75	4.48	3.01
WPE-BF-FBANK + env. derevb.	4.62	3.83	4.12	4.25	4.11	2.89
WPE-BF-FBANK + crr. derevb.	5.52	4.46	4.69	5.27	4.77	3.01
WPE-BF-FBANK +env. + crr. derevb.	**5.52**	**4.47**	**4.69**	**5.27**	**4.77**	3.01

Experiments and results

Table: MOS values (%) in REVERB dataset for envelope and carrier based enhancements.

	Eval Real - near	**Eval Real - far**	**Eval Simu - near**	**Eval Simu - far**
Baseline - WPE - GEV	3.78	3.65	3.74	4.12
+ FDLP enhancement	3.98	3.67	4.01	4.4

Five audio files were randomly selected from Eval Real and Eval Simu, both near and far room settings. Participants were asked to rate the audio quality on a scale from 1 (Poor) to 5 (Excellent). 20 subjects participated in the study.

Experiments and results

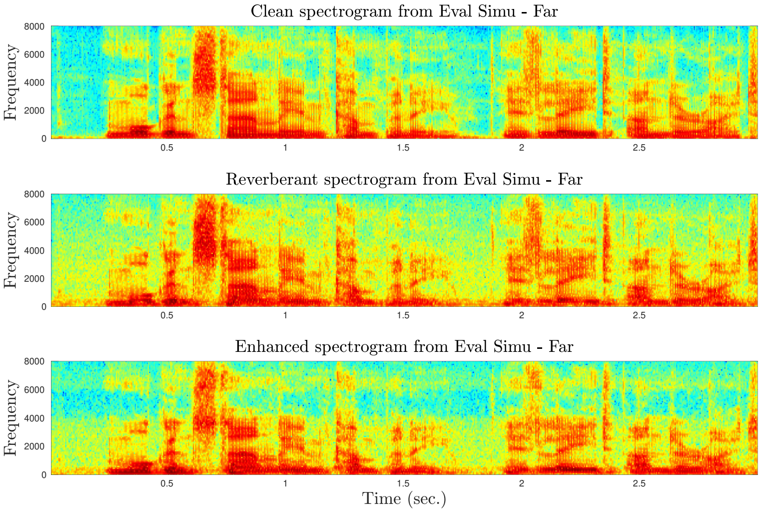


Figure: Spectrograms, after and before enhancement.

Summary - Part 2

- In this part of the work, we propose a speech enhancement model for E2E ASR systems using frequency domain linear prediction based sub-band envelopes and carrier.

Summary - Part 2

- In this part of the work, we propose a speech enhancement model for E2E ASR systems using frequency domain linear prediction based sub-band envelopes and carrier.
- Using the joint learning of the neural speech enhancement module and the E2E ASR model, we perform several speech recognition experiments on the REVERB challenge dataset as well as on the VOiCES dataset.

Summary - Part 2

- In this part of the work, we propose a speech enhancement model for E2E ASR systems using frequency domain linear prediction based sub-band envelopes and carrier.
- Using the joint learning of the neural speech enhancement module and the E2E ASR model, we perform several speech recognition experiments on the REVERB challenge dataset as well as on the VOiCES dataset.
- We also performed subjective and objective speech quality evaluation on REVERB challenge dataset.

Summary - Part 2

- In this part of the work, we propose a speech enhancement model for E2E ASR systems using frequency domain linear prediction based sub-band envelopes and carrier.
- Using the joint learning of the neural speech enhancement module and the E2E ASR model, we perform several speech recognition experiments on the REVERB challenge dataset as well as on the VOiCES dataset.
- We also performed subjective and objective speech quality evaluation on REVERB challenge dataset.
- Results show that the proposed speech enhancement method improves speech quality over the baseline.

Publications

- P. Anurenjan *et al.* "Dereverberation of autoregressive envelopes for far-field speech recognition." Computer Speech & Language 72 (2022): 101277.
- P. Anurenjan *et al.* "Speech enhancement with frequency domain auto-regressive modeling" TASLP (Transactions on Audio, Speech, and Language Processing). IEEE, 2023.

Publications

- K. Rohit, P. Anurenjan *et al.* "End-To-End Speech Recognition with Joint Dereverberation of Sub-Band Autoregressive Envelopes." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.
- P. Anurenjan *et al.* "Deep Learning Based Dereverberation of Temporal Envelopes for Robust Speech Recognition." Proc. Interspeech 2020, 1688-1692, DOI: 10.21437/Interspeech.2020-2283.
- P. Anurenjan *et al.* "3-D acoustic modeling for far-field multi-channel speech recognition." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

Acknowledgements

- This work was partly supported by project grants from Samsung Research India, Bangalore and the College of Engineering, Trivandrum, Kerala.

Acknowledgements

- I would like to express my deepest gratitude to my thesis advisor Prof. Sriram Ganapathy for his continuous support, guidance and motivation.
- I would like to thank Prof. S. Govindarajan, Prof. P. K. Ghosh, Prof. K. V. S. Hari, Prof. A. G. Ramakrishnan, Prof. A. Gopalan, Prof. C. Seelamanthula, Prof. P. S. Sastry and Prof. M. Arigovidan for their valuable lectures during my research training program.
- I would like to thank Prof. Ajayan K. R., College of Engineering, Trivandrum for giving inspiration and support to purse my higher studies.

Acknowledgements

- I would like to thank the present and past members of LEAP laboratory for their help, useful comments and suggestions on my work.
- I would like to thank my co-authors Anirudh Sreeram, Rohit Kumar and Debottam Dutta for their contribution.
- I would like to thank Rubin Jose Peter, Abhijith, Vivek R. S., Kiran R., Vinod V., Prince Philip, Jerin Thomas, Anand Mohan, Kiran Praveen, Arjun, Gokulan and Gokul for supporting me during my best and worst times at IISc.
- Finally i thank my parents, wife and daughter for their support in pursuing my higher education.

THANK YOU...