Dereverberation of Speech Using Autoregressive Models of Sub-band Envelopes

A THESIS

SUBMITTED FOR THE DEGREE OF

Doctor of Philosophy

IN THE

Faculty of Engineering

BY

Anurenjan P. R.

Electrical Engineering
Indian Institute of Science
Bangalore – 560 012 (INDIA)

October, 2023

Declaration of Originality

I, **Anurenjan P. R.**, with SR No. **04-03-00-12-12-17-1-15198** hereby declare that the material presented in the thesis titled

Dereverberation of Speech Using Autoregressive Models of Sub-band Envelopes

represents original work carried out by me in the **Department of Electrical Engineering** at **Indian Institute of Science** during the years **2017 -2023**.

With my signature, I certify that:

- I have not manipulated any of the data or results.
- I have not committed any plagiarism of intellectual property. I have clearly indicated and referenced the contributions of others.
- I have explicitly acknowledged all collaborative research and discussions.
- I have understood that any false claim will result in severe disciplinary action.
- I have understood that the work may be screened for any form of academic misconduct.

Date: 4/3/2023

Student Signature

In my capacity as supervisor of the above-mentioned work, I certify that the above statements are true to the best of my knowledge, and I have carried out due diligence to ensure the originality of the report.

Advisor Name: Sriram Ganapathy

Advisor Signature

DEDICATED TO

My Father, Mother, Wife & Daughter

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.

I would like to thank the members of LEAP laboratory for their useful comments and suggestions on my work. The help rendered by Anirudh Sreeram and Rohit Kumar were invaluable.

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.

Abstract

Automatic speech recognition (ASR) based technologies are radically changing the way we interact with digital services and information. Most of these application leverage on hands-free speech, where talkers are able to speak at a distance from the microphones without the nuance of handheld or body-worn device. The applications like, meeting annotations, speech to text transcription in teleconferencing, hands-free interfaces for controlling consumer-products, like interactive TV, virtual assistants in mobile phones, smart speakers etc, will benefit from distant talking mode of operation. The main issues in distant talking speech recognition is the corruption of speech signals by noise and the reverberation. This thesis is focused on developing dereverberation methods for speech processing using sub-band temporal envelopes.

This thesis pursues two broad directions for addressing issues in far-field ASR. In the first part of the thesis, two methods for dereverberation are proposed. In the second part of the thesis, we develop a speech enhancement model, where the audio signal is re-synthesized using dereverberated temporal envelopes and corresponding carrier components.

In the first part of the thesis, two methods to address reverberation is developed. The first method deals with developing a 3-D Acoustic modeling framework for far-field ASR (Automatic Speech Recognition), where spatio-spectral features from all the available channels are extracted. The features that are input to the 3-D CNN are extracted by modeling the signal peaks in the spatio-spectral domain using a multi-variate autoregressive (AR) modeling approach. This AR model is efficient in capturing the channel correlations in the frequency domain of the multi-channel signal. In the second method, a neural model for speech dereverberation using the long-term sub-band envelopes of speech is developed. The neural dereverberation model estimates the envelope gain, which when applied to reverberant signals, suppresses the late reflection components in the far-field signal. The dereverberated envelopes are used for feature extraction in speech recognition.

The second part of the thesis deals with envelope-carrier based speech enhancement. Here, we investigate the effect of far-field artifacts on temporal envelopes and the corresponding carrier components. A dual path recurrent neural model is used to parallelly learn the mapping for the reverberant envelopes and the carrier signals. Further, joint learning of the speech enhancement model with the end-to-end ASR model a single neural model is proposed.

Both parts of the thesis use the frequency domain linear prediction (FDLP) based model for extracting the envelopes of the sub-band signals in long analysis windows. We show several ASR and speech quality experiments to highlight the benefits of the proposed techniques.

Publications based on this Thesis

1. **3-D Feature and Acoustic Modeling for Far-Field Speech Recognition**

   A. Purushothaman A. Sreeram and S. Ganapathy

   Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, **ICASSP 2020**.

2. **Deep Learning Based Dereverberation of Temporal Envelopes for Robust Speech Recognition**

   A. Purushothaman A. Sreeram Rohit Kumar and S. Ganapathy

   Proceedings of 21st Annual Conference of International Speech Communication Association, **INTERSPEECH 2020**.

3. **SRIB-LEAP lab submission to Far-field Multi-Channel Speech Enhancement Challenge for Video Conferencing**

   P. R. Gudepu, R. Kumar, M. K. Jayesh, **A. Purushothaman**, S. Ganapathy and M. A. Basha

   Proceedings of 22nd Annual Conference of International Speech Communication Association, **INTERSPEECH 2021**.

4. **Dereverberation of Autoregressive Envelopes for Far-field Speech Recognition**

   A. Purushothaman A. Sreeram, R. Kumar and S. Ganapathy

   Elsevier Journal on Computer, Speech and Language, **2021**.

5. **End-to-end speech recognition with joint dereverberation of sub-band autoregressive envelopes**

   R. Kumar, **A. Purushothaman**, A. Sreeram and S. Ganapathy

   Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing, **ICASSP 2022**.

6. **Speech Dereverberation with Frequency Domain Autoregressive Modeling**

   A. Purushothaman, D. Dutta, R. Kumar and S. Ganapathy

   Under review in IEEE Transactions on Audio, Speech and Language Processing, **IEEE TASLP**.

Acronyms

ASR	Automatic speech recognition
DPLSTM	Dual path long short time memory
FDLP	Frequency domain linear prediction
GEV	Generalized eigen value
AR	Auto-regessive
LP	Linear prediction
AM	Acoustic modeling / Acoustic models
LM	Language modeling / Language models
DCT	Discrete cosine transform
HMM	Hidden Markov model
GMM	Gaussian mixture model
DNN	Deep neural networks
LSTM	Long short term memory
MVDR	Minimum variance distortion-less response
PESQ	Perceptual evaluation of speech quality
SRMR	Speech-to-reverberation modulation energy ratio
RNN	Recurrent neural networks
WER	Word error rate
WPE	Weighted prediction error

Acronyms

Table 1: *Acronyms used in this thesis*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Automatic speech recognition (ASR) based technologies are radically changing the way we interact with digital services and information. Most of these application leverage on far-field speech, where the users are able to speak at a distance from the microphones without the nuance of a handheld or body-worn device [2]. The applications like, meeting logging, speech to text transcription in teleconferencing, hands-free interfaces for controlling consumer-products, like interactive TV, virtual assistants in mobile phones etc. allow distant talking mode of operation. The main issues in far-field speech recognition is the corruption of speech signals by noise and reverberation. Usually, the effect of noise is short-term, where as reverberation is a long term effect. Further, noise component can be assumed to be statistically independent of the signal component. However, reverberant components are dependent on the signal and hence require dedicated algorithms.

1.1 The problem of reverberation

When speech is captured by a microphone at a distance, a number of reflections from various different surfaces in the recording environment reach the capturing device. There will be a direct path component and multiple reflected components reaching the mic as seen in Figure 1.1. These reflected path signals will hinder with the direct path component, leading to reduced

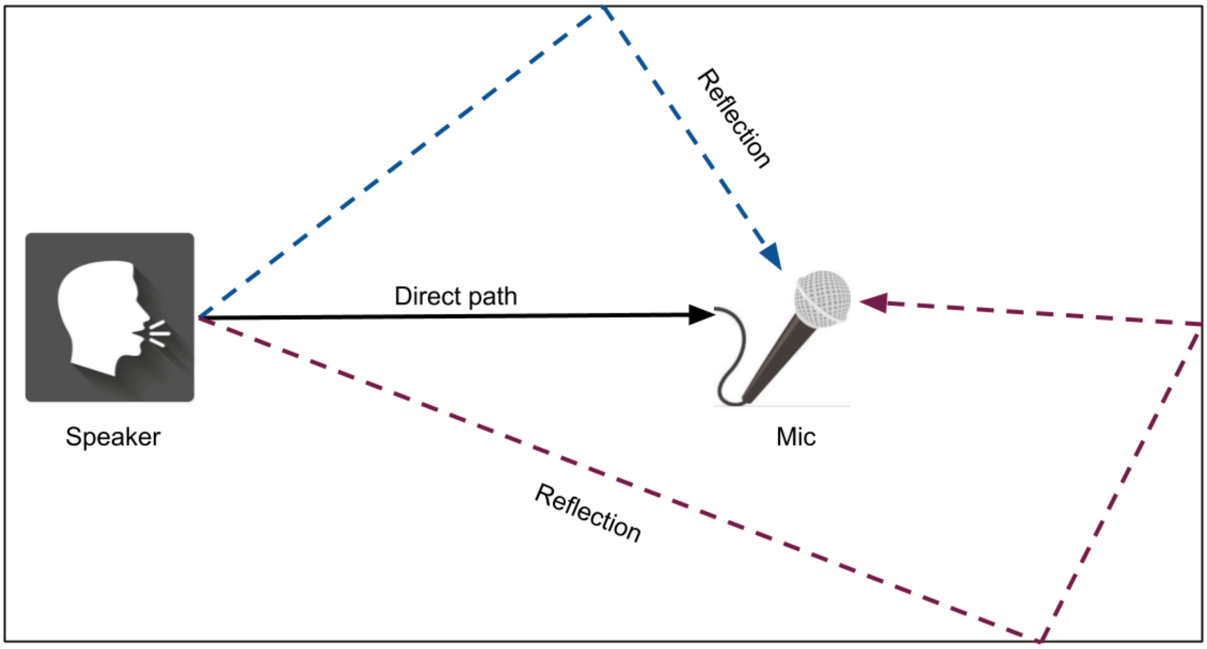


Figure 1.1: *Degradation due to multipath signals - Reverberation.*

quality and even incomprehensibility [3].

The performance of systems like, automatic speech recognition and speaker recognition degrades in noisy and reverberant conditions [4, 5, 6, 7, 8]. Reverberation alters the acoustic characteristic of the original signal. This deterioration is due to smearing of temporal envelopes caused by reverberation [9].

1.1.1 Mathematical model

When speech is recorded in far-field reverberant environment, the data collected in the microphone is modeled as

$$r(t) = x(t) * h(t), \tag{1.1}$$

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 $h(t) = h_e(t) + h_l(t)$, where $h_e(t)$ and $h_l(t)$, represent the early and late reflection components.

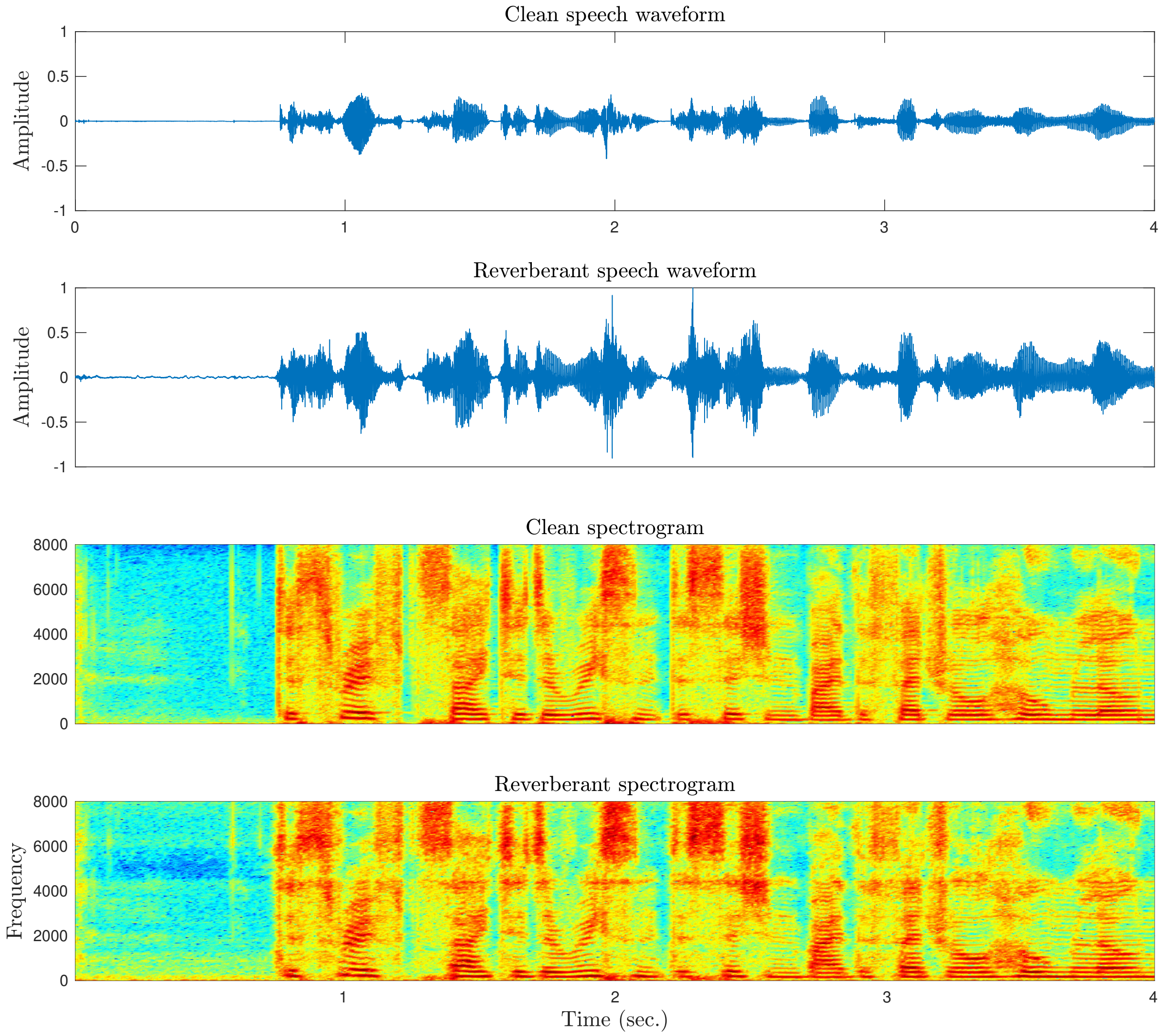


Figure 1.2: *Waveforms and spectrograms corresponding to clean and reverberated signal.*

1.1.2 Impact on applications - Speech recognition, Speaker recognition, Listening

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 [4]. For *e.g.* Peddinti *et al.,* [10] reports a 75% rel. increase in word error rate (WER) when signals from a far-field array microphone are used in place of those from headset microphones in the ASR systems, both during training and testing.

The applications like speaker recognition are also affected by reverberation. Ladislav Mosner *et al.,* in [11] reports a 3-fold increase in equal error rate (EER) in reverberant conditions compared to the clean counter part. The significant reduction in performance of the speaker recognition system, trained with clean speech was reported in [12], when tested with speech from distant microphone. Speaker verification on short utterances in uncontrolled noisy and reverberant environment conditions is one of the most challenging and highly demanding tasks. This was confirmed by the VOiCES from a Distance challenge 2019 (VOiCES 2019 challenge) [13, 14].

We address the problem of reverberation and noise in two directions. In the first case, we enhance the signal quality using some pre-processing. The typical steps involve dereverberation followed by beamforming. Here, the aim is to improve the speech signal to noise ratio or other signal quality measures. Another approach is to jointly model speech enhancement and an ASR system. This allows us to optimize the enhancement module with ASR cost.

1.2 Speech enhancement for far-field ASR

In most of the present day ASR systems, the multi-channel far-field recording is subject to 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) [15, 16] and perform a beamforming step, where all the available channels are combined to form a single channel [17]. In joint speech enhancement approaches, optimization of the speech enhancement model is jointly done with ASR model [18, 19, 20]. Here, we provide a brief introduction of the techniques. Detailed discussion on these techniques are available in Chapter 2.

1.2.1 Weighted Prediction Error (WPE) enhancement

Weighted prediction error (WPE) is a statistical model-based speech dereverberation approach that can cancel the late reverberation of a reverberant speech signal captured by distant microphones without prior knowledge of the room impulse responses. With this approach, the generative model of the captured signal is composed of a source process, which is assumed to be a Gaussian process with a time-varying variance, and an observation process modeled by a delayed linear prediction (DLP). The optimization objective for the dereverberation problem is derived to be the sum of the squared prediction errors normalized by the source variances; hence, this approach is referred to as variance-normalized delayed linear prediction (NDLP) [15, 16].

1.2.2 Multi-channel enhancement - beamforming

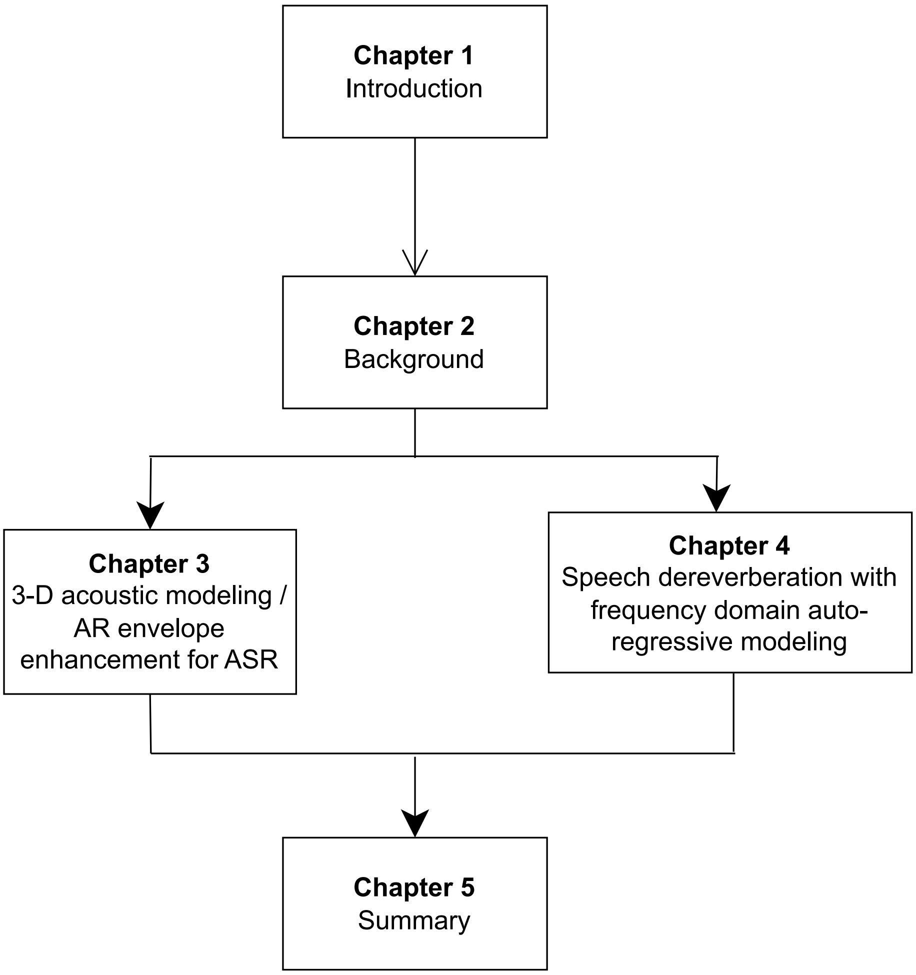
The method of beamforming performs a delayed and weighted summation of the multiple spatially separated microphones to provide an enhanced audio signal. The advancements to the basic beamforming using blind reference-channel selection and two-step time delay of arrival (TDOA) estimation with Viterbi postprocessing has been proposed to improve the beamforming algorithm [17].

An alternate approach to beamforming using a generalized eigen value (GEV) formulation [21] involves a spatial filtering in the complex short-time Fourier transform (STFT) domain. The filter is derived by solving an eigen value problem that maximizes the variance in the “signal” direction while minimizing the variance in the “noise” direction [21] or by keeping the variance in the target direction to be unity while minimizing the variance in the other directions (minimum variance distortionless response (MVDR) beamforming) [22].

1.2.3 Joint speech enhancement

The joint training and optimization of the speech enhancement model with the downstream ASR acoustic model has been a quite popular approach in the recent times. The initial attempt in [18] incorporates a DNN based speech separation model coupled with a DNN based acoustic model. In Bo Wu et. al. [19], unification of separately trained speech enhancement neural model and the acoustic model was proposed, in which the unified or the joint model is further trained to improve the ASR performance. Another work of Bo Wu et. al. [20] also explored an end-to-end deep learning approach, where, the DNN based dereverberation front end leverages the knowledge about the reverberation time. The ASR cost was further improved by jointly training this reverberation time aware-DNN enhancement module and the ASR acoustic module.

1.3 Organization of the thesis


Chapter 1
Introduction
Chapter 2
Background
Chapter 3
3-D acoustic modeling / AR envelope enhancement for ASR
Chapter 4
Speech dereverberation with frequency domain auto-regressive modeling
Chapter 5
Summary

1.4 Contributions of this thesis

In this section, we briefly outline the contributions of the thesis. The details of the works mentioned here can be found in the subsequent chapters.

1.4.1 3-D Acoustic modeling framework for ASR (Chap. 3)

The conventional approach to automatic speech recognition in multi-channel reverberant conditions involves a beamforming based enhancement of the multi-channel speech signal followed by a single channel neural acoustic model. We propose to model the multi-channel signal directly using a convolutional neural network (CNN) based architecture which performs the joint acoustic modeling on the three dimensions of time, frequency and channel [23]. The features that are input to the 3-D CNN are extracted by modeling the signal peaks in the spatio-spectral domain using a multi-variate autoregressive modeling approach. This AR model is efficient in capturing the channel correlations in the frequency domain of the multi-channel signal. The experiments are conducted on the CHiME-3 and REVERB Challenge dataset using multi-channel reverberant speech. In these experiments, the proposed 3-D feature and acoustic modeling approach provides significant improvements over an ASR system trained with beamformed audio (average relative improvements of 16% and 6% in word error rates for CHiME-3 and REVERB Challenge datasets, respectively).

1.4.2 Auto-regressive envelope enhancement for ASR (Chapter 3)

We develop a neural model for speech dereverberation using the long-term sub-band envelopes of speech. The sub-band envelopes are derived using frequency domain linear prediction (FDLP), which performs an autoregressive estimation of the Hilbert envelopes [24]. The neural dereverberation model estimates the envelope gain which when applied to reverberant signals suppresses the late reflection components in the far-field signal. The dereverberated envelopes are used for feature extraction in speech recognition. Further, the sequence of steps involved in envelope dereverberation, feature extraction and acoustic modeling for ASR can be implemented as a single neural processing pipeline which allows the joint learning of the dereverberation network and the acoustic model [25]. The end-to-end (E2E) automatic speech recognition (ASR) systems are commonly used. In a similar analogy to our works on hybrid HMM-DNN ASR systems, we extended the approach to E2E ASR systems as well. The average relative improvements of 10-24% over the baseline system are observed on REVERB challenge dataset and the VOiCES dataset.

1.4.3 Speech dereverberation with Envelope-Carrier Modeling (Chap. 4)

In this line of work, we investigate the effect of reverberation in temporal envelopes and the corresponding carrier components. The effect of reverberation can be approximated as convolution of the long-term sub-band envelopes of clean speech with the envelope of room impulse response function [25, 26]. Here we show that the non linear relationship between the reverberant carrier and clean carrier can be learned by a deep learning model. Dual-path long short time model named, DPLSTM is used to parallelly learn the mapping between both the reverberant envelopes and the carrier. Further, joint learning of the speech enhancement model with the end-to-end ASR model in a single neural model is proposed.

Various ASR experiments are performed on the REVERB challenge dataset [27] as well as the VOiCES dataset [14, 13]. The experiments, show that the proposed method improves over the state-of-the-art speech enhancement system and ASR systems.

Chapter 2

Background

The details required to understand the rest of the thesis is discussed in this chapter. Section 2.2 discusses the frequency domain linear prediction (FDLP) and its details. The details of different beamforming algorithms are discussed in Section 2.4. This is followed by a discussion on different automatic speech recognition approaches in Section 2.5. Performance measures used in speech recognition and enhancement methods are detailed in Section 2.6. This is followed by a survey of speech enhancement approaches in the spectral domain in Section 2.7.

2.1 AR modeling of temporal envelopes

Conventional speech analysis techniques are based on estimating the spectral content of relatively short (about 15-25 ms) segments of the signal. An alternate way to describe a speech signal is a summation of amplitude modulated frequency bands, where each frequency band consists of a smooth envelope (gross structure) modulating a carrier signal (fine structure). The analytic signal (AS) forms a suitable candidate for such an envelope-carrier decomposition with the squared magnitude of the AS, called the Hilbert envelope, representing the smooth structure and the phase component of the AS representing the fine structure. We begin by the definition of discrete version of analytic signal.

2.1.1 Discrete-Time Analytic Signal

The discrete version of analytic signal (AS) [28] of $x[n]$ is found out by the following procedure.

1. Compute the N-point DFT sequence $X[k]$.

2. Find the N-point DFT of the AS as,

$$X_a[k] = \begin{cases} X[0] & \text{for } k = 0 \\ 2X[k] & \text{for } 1 \leq k \leq \frac{N}{2} - 1 \\ X[\frac{N}{2}] & \text{for } k = \frac{N}{2} \\ 0 & \text{for } \frac{N}{2} + 1 \leq k \leq N \end{cases} \tag{2.1}$$

2.1.2 Discrete Cosine Transform (DCT)

Let $x[n]$ denote an N-point discrete sequence. The type-I odd DCT [29] $y[k]$ for $k = 0, 1, \ldots, N-1$ is given by,

$$y[k] = 4 \sum_{n=0}^{N-1} c_{n,k} x[n] \cos\left(\frac{2\pi nk}{M}\right) \tag{2.2}$$

where $c_{n,k} = 1$ for $n, k > 0$ and $c_{n,k} = \frac{1}{2}$ for $n, k = 0$ and $c_{n,k} = \frac{1}{\sqrt{2}}$ for the values of n, k where only one of the index is 0 and $M = 2N - 1$. The DCT defined by equation 2.2 is a scaled version of the orthogonal DCT with a factor of $2\sqrt{M}$.

2.2 Frequency Domain Linear Prediction

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, i.e. square of its Hilbert envelope [30]. Temporal envelope is given by the inverse Fourier transform of the auto-correlation function of DCT,

$$e(t) = F^{-1}\left\{Autocorr(y[k])\right\} \tag{2.3}$$

where $y[k]$ is the DCT of a signal $x[n]$ having N- points. The auto-correlation of the DCT signal is defined as,

$$r_y[\tau] = \frac{1}{N}\sum_{k=|\tau|}^{N-1} y[k]y[k-|\tau|] \tag{2.4}$$

We use the auto-correlation of the DCT coefficients to predict the temporal envelope of the signal. One of the inherent property of linear prediction is that, it tries to approximate the peaks very well. The FDLP model tries to preserve the peaks in temporal domain.

2.2.1 Auto-correlations of DCT and Hilbert envelope

Let the discrete-time sequence $x[n]$ have a zero-mean property in time and frequency domains, i.e., $x[0] = 0$ and $X[0] = 0$. We can define the even-symmetrized version $q[n]$ of the input signal,

$$q[n] = \begin{cases} x[n] & \text{for } n = 0, \ldots, N-1 \\ x[M-n] & \text{for } n = N, \ldots, M-1 \end{cases} \tag{2.5}$$

where $M = 2N - 1$. An important property of $q[n]$ is that, it has a real spectrum given by,

$$Q[k] = 2\sum_{n=0}^{N-1} x[n]\cos\left(\frac{2\pi nk}{M}\right) \tag{2.6}$$

For signals with the zero-mean property in time and frequency domains, it follows from equations 2.2 and 2.7 that,

$$y[k] = 2Q[k] \tag{2.7}$$

for $k = 0, \ldots, N-1$. Let $\hat{y}$ denote the zero-padded DCT with $\hat{y}[k] = y[k]$ for $k = 0, \ldots, N-1$ and $\hat{y}[k] = 0$ for $k = N, \ldots, M-1$. From the definition of Fourier transform of the analytic signal in equation 2.1 and using the definition of the even symmetric signal in equation 2.5, we can see that,

$$Q_a[k] = \hat{y}[k] \tag{2.8}$$

where $Q_a[k]$ is the DFT of analytic signal of $q[n]$, denoted by $q_a[n]$. Hence the AS spectrum of the even-symmetric signal is equal to the zero-padded DCT signal. The inverse DFT of zero-padded DCT signal $\hat{y}[k]$ is the AS of the even symmetric signal. So it can be shown that,

$$r_y[\tau] = \frac{1}{N} \sum_{n=0}^{M-1} |q_a[n]|^2 e^{-j\frac{2\pi n\tau}{M}} \tag{2.9}$$

i.e., the auto-correlation of the DCT signal and the squared magnitude of the AS (Hilbert envelope) of the even-symmetric signal are Fourier transform pairs. Hence, linear prediction of DCT components results in AR model of the Hilbert envelope of the even-symmetrized signal.

Linear prediction of type-I odd DCT (Equation 2.2) components results in AR model of the Hilbert envelope of the even-symmetrized signal. If $\{a_k\}$ are the estimated linear prediction coefficients, the resulting FDLP envelope is given by,

$$\hat{E}_x(n) = \frac{G}{|\sum_{k=0}^{p} a_k e^{-i2\pi kn}|^2} \tag{2.10}$$

2.3 Weighted Prediction Error (WPE) enhancement

Weighted prediction error (WPE) is a dereverberation algorithm [15, 16]. Let the observed speech signal in m^{th} microphone be represented by $y^m(k,n)$ in the short time Fourier transform (STFT) domain, where k denotes the frequency bin index and n denotes the frame index. This observed signal is corrupted with reverberation and additive noise, $y^m(k,n)$.

$$y^m(k,n) = \sum_{l=0}^{L_h-1} g^m(k,l)\ x(k,n-l) + v^m(k,n) \tag{2.11}$$

where $g^m(k,n)$ is the STFT of the room response function, L_h is the length of room impulse response and $x(k,n)$ is the source signal STFT at m^{th} microphone. Usually, the early part of reverberation caused by the convolution of Room Impulse Response and the original signal can be suppressed using cepstral mean normalization techniques. However, the late reverberation component is the significant artifact. Thus, the signal model in the STFT domain is given by,

$$\begin{aligned}
\mathbf{y}(k,n) &= [y^1(k,n), y^2(k,n), ..., y^M(k,n)]^T && (2.12)\\
\mathbf{g}(k,n) &= [g^1(k,n), g^2(k,n), ..., g^M(k,n)]^T && (2.13)\\
\mathbf{v}(k,n) &= [v^1(k,n), v^2(k,n), ..., v^M(k,n)]^T && (2.14)
\end{aligned}$$

The convolutive model in Eq. (2.11) can be simplified and the signal at the reference microphone (e.g., $m = 1$) can be written as,

$$y^1(k,n) = d^m(k,n) + \sum_{m=0}^{M} \sum_{l=0}^{L_g-1} g^m(k,l) x^m(k,n-\tau-l) + v^1(k,n) \tag{2.15}$$

Here, $d^m(k,n)$ is the sum of the clean signal along with the early part of reverberation. The first term in Eq. (2.11) can be broken down into two parts, where the direct component of speech signal and early reflection component will be considered as the first part $d^m(k,n)$, while the second part will be the late reverberation part $r^m(k,n)$,

$$y^m(k,n) = d^m(k,n) + r^m(k,n) + v^m(k,n) \tag{2.16}$$

The technique of weighted prediction error (WPE) [15, 16] can be used to suppress the late reflection component $r^m(k,n)$. Following the vector notations in Eq. (2.12, 2.13, 2.14), we can write the multi-channel desired signal $\mathbf{y}(k,n)$ and the multi-channel late reverberation signal, $\mathbf{r}(k,n)$ is given by,

$$\mathbf{d}(k,n) = \sum_{l=0}^{L_e-1} \mathbf{g}(k,l)x(k,n-l) \tag{2.17}$$

$$\mathbf{r}(k,n) = \sum_{l=m}^{L_h} \mathbf{g}(k,l)x(k,n-l) \tag{2.18}$$

where L_e is the length of room impulse response upto which it is assumed to be part of early reverberation. In weighted Prediction Error (WPE) based dereverberation, the late reverberation component is removed from the speech assuming that $\mathbf{y}(k,n)$ follows the model given by,

$$\mathbf{y}(k,n) = \mathbf{d}(k,n) + \sum_{l=m}^{N} \mathbf{W}(k,l)^H \mathbf{y}(k,n-l) \tag{2.19}$$

where $\mathbf{W}(k,l)$ is an $M \times M$ linear prediction matrix and N is the order of the linear predictor. H is the Hermitian operator. Here the dereverberation is considered as a maximum likelihood estimation problem where the target is to minimize the second term in equation 2.19.

$$\tilde{\mathbf{W}} = \arg\min_{\tilde{\mathbf{W}}} \sum_{n} \frac{||\mathbf{y}(k,n) - \sum_{l=m}^{N} \mathbf{W}(k,l)^H \mathbf{y}(k,n-l)||}{\sigma_n^2} \tag{2.20}$$

where σ_n is the time varying quantity. Once we have $\tilde{W}$, we can solve equation 2.19 and obtain our desired signal i.e.

$$\mathbf{d}(k,n) = \mathbf{y}(k,n) - \sum_{l=m}^{N} \tilde{\mathbf{W}}(k,l)^H \mathbf{y}(k,n-l) \tag{2.21}$$

2.4 Beamforming

The conventional method of processing the multi-channel audio signal involves the spatial filtering performed via beamforming [31, 32].

2.4.1 Acoustic Beamforming

The acoustic beamforming system is based on the weighted- delay & sum microphone array theory, which is a generalization of the well known weighted-delay& sum beamforming technique [17]. The signal output $y[n]$ is expressed as the weighted sum of the different channels as follows,

$$y[n] = \sum_{m=1}^{M} W_m[n] x_m[n - TDOA^{m,ref}[n]] \tag{2.22}$$

where $W_m[n]$ is the relative weight for microphone m (out of M microphones) at instant n, with the sum of all weights equals to 1, $x_m[n]$ is the signal for each channel, and $TDOA^{(m,ref)}[n]$ (Time Delay of Arrival) is the relative delay between each channel and the reference channel, in order to obtain all signals aligned with each other at each instant n. $TDOA^{(m,ref)}[n]$ is estimated via cross correlation techniques once every several acoustic frames, using $GCC - PHAT$ (Generalized Cross Correlation with Phase Transform) [33, 34].

The overall channel weighting factor is used to normalize the input signals to match the file's available dynamic range. It is useful for low amplitude input signals since the beamformed output has greater resolution and therefore can be scaled appropriately to minimize the quantization errors generated by scaling it to the output sampling requirements. In each window the maximum value is found and these max values are averaged over the entire recording. The weighting factor $W_m[n]$ is obtained directly from this average.

The computation of the time delay of arrival ($TDOA$) between each of the channels and the reference channel is computed in segments of 250 ms. Top N values of $TDOA$ given by

$GCC-PHAT$ algorithm are used to find the best $TDOA$. For each analysis window we obtain a vector $TDOA_n^i$ for microphone i with $m = 1, ...M, i \neq ref$ with its corresponding correlation values from $GCC - PHAT$.

A Viterbi postprocessing technique is applied to the computed delays is used to select the appropriate delay to be used among the best $GCC - PHAT$ values computed previously. The aim here is to maximize speaker continuity avoiding constant delay switching in the case of multiple speakers, and to filter out undesired beam steering towards spurious noises present in the room. A two-step Viterbi decoding of the best $TDOA$ is used. The first step consists of a local (single-channel) decoding where the two best delays are chosen from the best delays computed for that channel at every segment. The second decoding step considers all combinations of two-best delays across all channels, and selects the final single TDOA value that is most consistent across all channels.

2.4.2 Generalized Eigen Value (GEV) Beamforming

The beamforming operation in frequency domain determines the spatial filter coefficients $w(m, k)$ to obtain the enhanced signal,

$$z(k, n) = \sum_{m=0}^{M-1} w(m, k)\ y^m(k, n) \tag{2.23}$$

where $z(k, n)$ is the beamformed signal. The main goal of GEV beamforming is to determine the spatial filter coefficients $\mathbf{w}(k) = [w(0, k), .., w(M-1, k)]^T$ such that the SNR at the output of the filter is maximized [21], i.e.,

$$\mathbf{w}_{GEV}(k) = \arg\max_{\mathbf{w}(k)} \frac{\mathbf{w}^H(k)\hat{\mathbf{\Phi}}_{XX}(k)\mathbf{w}(k)}{\mathbf{w}^H(k)\hat{\mathbf{\Phi}}_{VV}(k)\ \mathbf{w}(k)} \tag{2.24}$$

where $\hat{\mathbf{\Phi}}_{XX}$ and $\hat{\mathbf{\Phi}}_{VV}$ are power spectral density (PSD) estimates of the clean signal and noise respectively. H is the Hermitian operator.

The most successful approach to the estimation of clean and noise PSD is through the use of a neural mask estimator (described next). Once the PSD matrices are estimated, the solution to the optimization given in Eq. (2.24) is the eigen vector corresponding to maximum eigen value of the matrix $\hat{\mathbf{\Phi}}_{VV}^{-1}\hat{\mathbf{\Phi}}_{XX}$.

2.4.3 MVDR Beamforming

The most commonly used beamforming method is MVDR based beamforming. This formulation tries to minimize the residual noise keeping the constraint that the signal from preferred source direction being distortionless,

$$\mathbf{w}_{MVDR}(k) = \frac{\hat{\mathbf{\Phi}}_{VV}^{-1}(k)\mathbf{d}}{\mathbf{d}^H\hat{\mathbf{\Phi}}_{VV}^{-1}(k)\ \mathbf{d}} \tag{2.25}$$

where $\mathbf{d}$ specifies the preferred direction of arrival.

2.4.4 Neural Mask Estimator

As proposed in [35, 36], the neural mask estimators are deep feed-forward/recurrent networks that are trained to predict the speech presence probability in each time-frequency bin. In simulated settings (where $y^m(k,n)$ and $x(k,n)$ are available), the deep model is trained with magnitude STFT $|y^m(k,n)|$ coefficients for patch of frames n and all frequency bins to predict the ideal binary mask (IBM). The IBM is obtained by thresholding the ratio of magnitude STFT $\frac{|y^m(k,n)|}{|x(k,n)|}$ with a different threshold applied for voiced and unvoiced regions of the audio [35]. The output of the mask estimator performs a sigmoid non-linearity and these outputs are interpreted as speech presence probability estimators, $s(k,n)$, and noise presence probability estimators $u(k,n)$. Once the mask estimator is trained, the PSD matrices needed in Eq. (2.24)

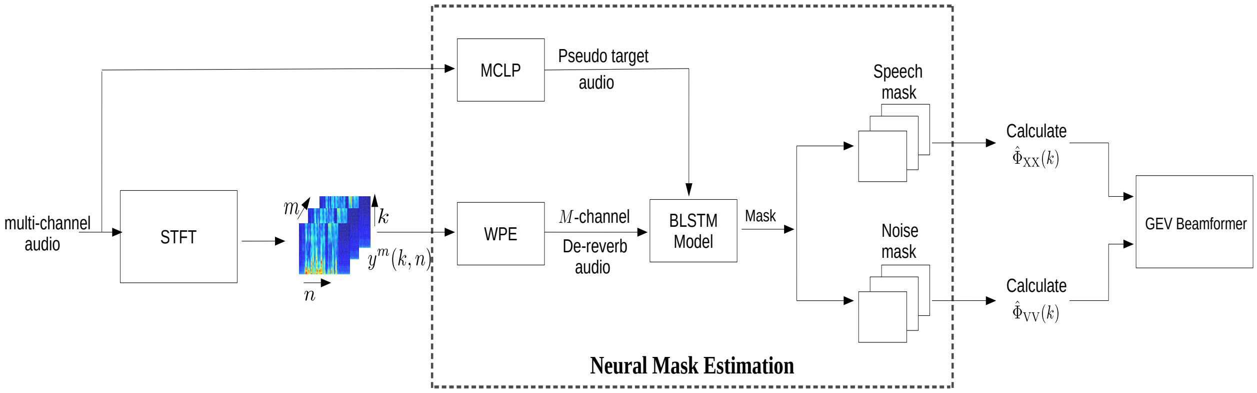


Figure 2.1: *Block schematic of the unsupervised neural network based mask estimation.*

for $\mathbf{y}(k,n) = [y^0(k,n),..,y^{M-1}(k,n)]^T$ are given as,

$$\hat{\mathbf{\Phi}}_{XX}(k) = \frac{\sum_n s(k,n)\mathbf{y}(k,n)(\mathbf{y}(k,n))^H}{\sum_n s(k,n)} \tag{2.26}$$

$$\hat{\mathbf{\Phi}}_{NN}(k) = \frac{\sum_n u(k,n)\mathbf{y}(k,n)(\mathbf{y}(k,n))^H}{\sum_n u(k,n)} \tag{2.27}$$

2.4.4.1 Unsupervised Mask Estimation

One of the limitations of the neural mask estimation described above is the need for simulated data with parallel clean and noisy multi-channel recordings to train the deep model. Hence, the real multi-channel recordings cannot be used in the neural mask training. In [37], R. Kumar *et al.* proposed a method to remove the requirement of having simulated settings by generating unsupervised pseudo targets for the real (and simulated) multi-channel recordings. The block schematic of the unsupervised mask estimation algorithm is given in Fig. 2.1.

2.5 Automatic speech recognition - Hybrid, E2E ASR, Joint modeling

The Automatic Speech Recognition (ASR) is the process of generating the transcription (word sequence) of an utterance, given the speech waveform. Essentially, an ASR system identifies an input sequence, $X = \{x_1, x_2 \ldots, x_T\}$ of length T as a sequence of words (generally letters, phonemes, sub-words or words etc.), $W = \{w_1, w_2, \ldots, w_N\}$. Here, $x_t \in \mathbb{R}^D$, is the speech input vector at time t, of dimension D. Let $\mathcal{W}$ denotes the possible vocabulary in the language and $\mathcal{W}^*$ is the collection of all sequences possible using the elements in $\mathcal{W}$. The task of ASR is to find the most likely label sequence W^* given X.

2.5.1 Hybrid ASR

History of automatic speech recognition dated back to 1950's when Dreyfus Graf [38, 39] tried to represent speech signal as the output of band pass filters and used for transcription. The SPHINX system [40] developed by Kai-Fu Lee using Hidden Markov Model (HMM) to model transitions from one phoneme to another, which is represented by a Gaussian Mixture Model (GMM), was a mile stone in automatic speech recognition (HMM-GMM systems).

Once the deep learning models were introduced, there were attempts to integrate HMM-GMM system with deep learning framework. Hybrid HMM-DNN dates back to 1990's Boulard's connectionist HMM work [41]. Dahl *et al.* [42] came up with Context-Dependent Pre-Trained Deep Neural Networks for ASR, popularly known as Hybrid ASR. It achieved significant performance gains compared to traditional HMM-GMM system in ASR tasks.

The three building blocks (see Figure 2.2) of conventional HMM-DNN hybrid ASR are,

- Acoustic model $P_\theta(X/S)$: describes the probability of observing X from the hidden state S
- Pronunciation model $P(S/W)$: It captures the connection between the hidden sequence

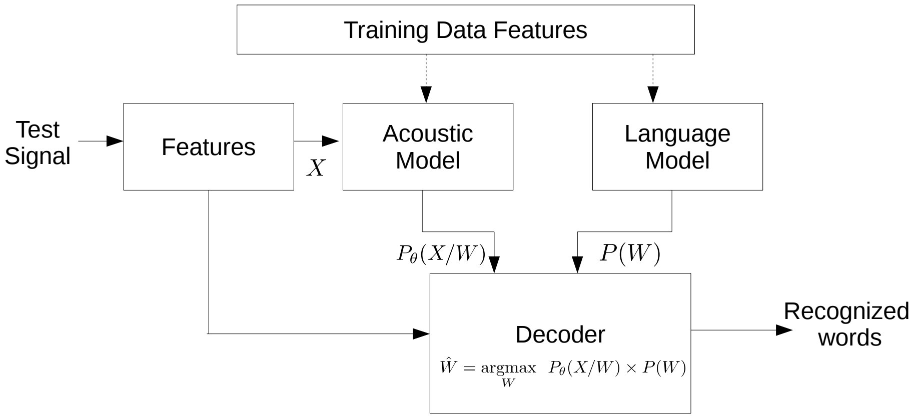


Figure 2.2: *Hybrid ASR block schematic.*

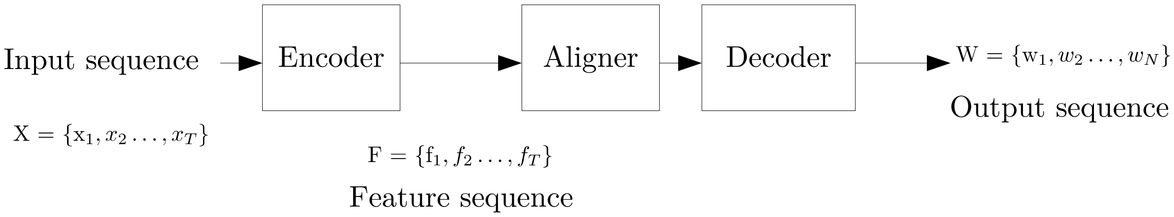
S and the language sequence W. Usually, the hidden sequence S represents the sequence of context dependent phonemes and W is the sequence of words in the language. The pronunciation model, also called dictionary, contains the mapping between all the words in the language and the corresponding context dependent phonemes.

- Language model $P(W)$: Represents the probability of a particular word sequence W. It could either be an n-gram language model or a recurrent neural network based language model [43].

The three modules in HMM based ASR are trained independently and serve different purposes. The training process is complex and difficult to optimize globally. There are conditional independent assumptions that may be violated in reality.

2.5.2 End-to-end ASR

Due to the inherent limitations of the hybrid ASR and the thrust for a deep neural network based solution, more focus was given to an end-to-end solution for ASR. The end-to-end system directly maps the speech signal to a word sequence. Most end-to-end ASR systems have three parts: encoder, which maps speech input sequence to feature sequence; aligner, which captures



Input sequence
Encoder
Aligner
Decoder
W = {w$_1, w_2 \ldots, w_N$}
Output sequence
X = {x$_1, x_2 \ldots, x_T$}
F = {f$_1, f_2 \ldots, f_T$}
Feature sequence

Figure 2.3: *Functional structure of end-to-end system.*

the alignment between feature sequence and language; decoder, which decodes the final word sequence (see Figure 2.3). This division may not always exist, since an end-to-end system is a complete model, which generates the transcription of the given utterance directly. Thus, the modularity in a conventional hybrid ASR is not present here. Multiple modules are merged into a single model for joint training. This enables an evaluation criteria which is highly relevant in the final global optimum, at the training time [44]. It directly maps input acoustic sequence to the text sequence, and does not require further processing to improve recognition performance [45].

The end-to-end model can be divided into three different categories depending on their implementations of soft alignment,

- Connectionist Temporal Classification (CTC): CTC first enlists all possible hard alignments, then it does soft alignment by clubbing these hard alignments. CTC assumes that output labels are independent of each other in the hard alignments [44].

- RNN-transducer: it also enumerates all possible hard alignments and then aggregates them for soft alignment. But the difference here is, RNN-transducer does not make independent assumptions about labels when enlisting hard alignments, so it is different from CTC in terms of path definition and probability calculation [46, 47].

- Attention-based: this method no longer enumerates all possible hard alignments, but uses attention mechanism to directly calculate the soft alignment information between input data and output label [48].

The continuity, monotonicity, etc. of the attention mechanism naturally do not exist in CTC because CTC itself requires continuity and monotonicity of soft alignment between input and output. In [49], S. Watanabe *et al.* attempted to combine CTC and attention in the training and inference process to solve these issues.

2.6 Performance metric - SRMR, PESQ, DNSMOS, Subjective test, WER

To evaluate the performance of dereverberation algorithms, subjective and/or objective quality and intelligibility measurement methods are needed. The most widely used subjective method is ITU-T 800[50], where a panel of listeners are asked to rate the quality/intelligibility of the audio. Commonly, subjective quality tests have listeners rate the quality of the speech signal on a pre-specified scale. Typically the mean opinion score (MOS) is used. The MOS score is a rating from 1 to 5 of the perceived quality of a speech recording, 1 being the lowest score and 5 the highest for excellent quality.

Objective methods can be broadly classified into intrusive and non-intrusive based on whether a clean reference signal is required or not. Intrusive methods need the clean reference signal and quantifies the distance metric between the clean one and noisy counter part. Non-intrusive measures, on the other hand, do not depend on a reference signal.

2.6.1 Speech to reverberation modulation energy ratio (SRMR)

Speech to reverberation modulation energy ratio is a non intrusive measure. Here a representation is obtained using an auditory-inspired filter bank analysis of critical band temporal envelopes of the signal. Modulation spectral information is used to get an adaptive measure termed speech to reverberation modulation energy ratio [51, 52]. The SRMR ranges from 1 to ∞, higher the better.

2.6.2 Perceptual Evaluation of Speech Quality (PESQ)

Perceptual Evaluation of Speech Quality is an intrusive measure, which requires a clean reference signal. The technique was developed to model subjective tests commonly used in telecommunications [53, 54]. PESQ is the result of integration of perceptual analysis measurement system (PAMS) [55, 56] and PSQM99 [57]. PESQ returns a score from 1.0 to 4.5, with higher scores indicating better quality.

2.6.3 Non-Intrusive Speech Quality Assessment (NISQA)

With the recent advancements in deep neural networks, researchers have came up with models predicting the MOS directly, without any reference signal. Non-intrusive Speech Quality Assessment (NISQA) framework is such a model where the neural model directly predicts the MOS score [58] (range 1 to 5). NISQA is trained and evaluated on a large set of 59 training datasets ($72,903$ files), 18 validation sets ($9,567$ files), and 4 test sets (952 files).

2.6.4 ASR performance metrics

- WER (Word Error Rate): Word error rate is the most common measure used to evaluate the performance of automatic speech recognition system. It is given by the equation,

  $$WER = \frac{SUB + DEL + INS}{N} \tag{2.28}$$

  where SUB is the number of substitutions, DEL is the number of deletions, INS is the number of insertions, and N is the number of words in the reference.

- CER (Character Error Rate): In a similar line we can define character error rate as well, where the numbers in the above equation are replaced by corresponding numbers for characters.

2.7 Relevant prior works - Spectrum enhancement

Speech enhancement based on spectral subtraction is the most simple and one of the earliest methods. It is based on the idea that the noise is additive in nature, can be estimated/predicted and this when subtracted from the noisy speech, gives the estimate of clean speech. Spectral subtraction based algorithms were initially proposed by Weiss *et al.* [59] in the correlation domain and later by Boll [60] in the Fourier transform domain.

2.7.1 Basic subtractive algorithm

Assume that $y[n]$, is the noise-corrupted input signal. $y[n]$ is made from the clean speech signal $x[n]$ and the additive noise signal, $d[n]$, that is,

$$y[n] = x[n] + d[n] \tag{2.29}$$

In the Fourier domain the relation becomes,

$$Y[f] = X[f] + D[f] \tag{2.30}$$

Expressing $Y[f]$ in polar form,

$$Y[f] = |Y[f]| \times e^{\phi_Y(f)} \tag{2.31}$$

where $|Y[f]|$ is the magnitude spectrum and $e^{\phi_Y(f)}$ is the phase spectrum of the noisy signal. In a similar analogy, noisy spectrum $D[f]$ can be represented as $|D[f]| \times e^{\phi_D(f)}$. Suppose we estimate the noise magnitude spectrum ($|\hat{D}[f]|$) by averaging over the silent parts of the speech. Let the the noise phase part be replaced by noisy signal phase $e^{\phi_Y(f)}$. Assuming independence between speech and noise, we can arrive at an estimate of the clean signal spectrum as,

$$X[f] = [|Y[f]| - |\hat{D}[f]|] \times e^{\phi_Y(f)} \tag{2.32}$$

The estimated clean signal is obtained by performing inverse Fourier transform on $X[f]$.

Equation 2.32 summarizes the underlying principle of spectral subtraction. Compute the magnitude spectrum of the noisy speech and keep an estimate of the noise spectrum when speech is not present. Subtract the noise magnitude spectrum from the noisy speech magnitude spectrum and, finally, take the inverse Fourier transform of the difference spectra (using the noisy phase) to produce the enhanced speech signal.

2.7.2 Wiener Filtering

Wiener filtering approach derives the enhanced signal by optimizing a mathematically tractable error criterion, the mean-square error. We pose the enhancement problem as a filtering problem as below.

$$\hat{x}[n] = \sum_{k=-\infty}^{\infty} h[k]y[n-k] \tag{2.33}$$

The noisy signal $y[n]$ goes through a linear time invariant filter, with impulse response $h[n]$ and produces an estimate of the clean signal $\hat{x}[n]$ [61]. We design the system $h[n]$ in such a way that the output, $\hat{x}[n]$ is as close to the desired signal $x[n]$. This can be done by computing the estimation error, $e[n]$. The optimal filter that minimizes the estimation error is called the *Wiener filter.*

We compute the filter coefficients $h[k]$ so that the estimation error, $x[n] - \hat{x}[n]$ is minimized in a statistical sense. The mean square of the estimation error is commonly used as a criterion for minimization, and the optimal filter coefficients are derived in either time or frequency domain.

In the frequency domain Equation (2.33) becomes,

$$\hat{X}[f] = H[f]Y[f] \tag{2.34}$$

where $\hat{X}[f]$, $H[f]$ and $Y[f]$ are the discrete-time Fourier transforms of $\hat{x}[n]$, $h[n]$ and $y[n]$, respectively. The estimation error at a particular frequency f_k is given by,

$$\begin{aligned} E[f_k] &= X[f_k] - \hat{X}[f_k] && (2.35) \\ &= X[f_k] - H[f_k]Y[f_k] && (2.36) \end{aligned}$$

The mean square error is given by,

$$\begin{aligned} \mathcal{E}[|E[f_k]|^2] &= \mathcal{E}\Big(\{X[f_k] - H[f_k]Y[f_k]\} \times \{X[f_k] - H[f_k]Y[f_k]\}\Big) && (2.37) \\ &= \mathcal{E}[|X[f_k]|^2] - H[f_k]P_{yx}[f_k] - H^*[f_k]P_{xy}[f_k] + |H[f_k]|^2 P_{yy}[f_k] && (2.38) \end{aligned}$$

where $P_{yy}[f_k] = \mathcal{E}\big(|Y[f_k]|^2\big)$, is the power spectrum of $y[n]$, $P_{xy}[f_k] = \mathcal{E}\big(X[f_k]Y[f_k]\big)$ is the cross power spectrum of $x[n]$ and $y[n]$.

For finding the optimal $H[f_k]$, we find the complex derivative of $\mathcal{E}[|E[f_k]|^2]$ with respect to $H[f_k]$ and equate it to zero. Solving for $H[f_k]$, the general form of wiener filter in the frequency domain as,

$$H[f_k] = \frac{P_{xy}[f_k]}{P_{yy}[f_k]} \tag{2.39}$$

Equation (2.39) comprise the general equations of the Wiener filters in the frequency domain.

2.7.3 Neural enhancement and dereverberation attempts

With the advent of deep neural networks, there has been a surge in speech enhancement based on neural model.

2.7.3.1 Enhancement and dereverberation - neural network based

In [62], Hu *et al.* used a deep complex convolution recurrent neural network based model for phase aware speech enhancement. For speech enhancement, Xu et. al. [63] devised a mapping from noisy speech to clean speech using a supervised neural network. In a similar manner, ideal ratio mask based neural mappings [64] have been explored for speech separation tasks. On the dereverberation front, Zhao et. al. proposed an LSTM model for late reflection prediction in the spectrogram domain [65]. Han et. al [66] developed a spectral mapping approach using the log-magnitude inputs and Williamson et. al [67] proposed a mask-based approach for dereverberation on the complex short-term Fourier transform. In a different line of work, speech enhancement in the time domain was pursued by Pandey et. al [68].

The application of speech dereverberation as a pre-processing step for downstream applications like ASR have been explored in several works (for example, [69, 70, 71]). The recent years have seen the use of recurrent neural network architectures for dereverberation. For example, Maas et. al [72], utilized a recurrent neural network (RNN) to establish mapping between noise-corrupted input features and their corresponding clean targets. Also, the use of a context-aware recurrent neural network-based convolutional encoder-decoder architecture was investigated by Santos et. al. [73].

2.7.3.2 Robust multi-channel ASR

In the design of front-end for robust ASR, Generalized sidelobe canceller (GSC) [74, 75] is a common approach. It was introduced by Li et. al in [76], where the authors proposed a neural network-based generalized side-lobe canceller. To combine spectral and spatial information from multiple channels using attention layers, an end-to-end multi-channel transformer was investigated in [77]. In another attention modelling approach, the streaming ASR model based on monotonic chunk-wise attention was proposed by Kim et. al in [78]. Ganapathy et. al. [6] proposed a 3-D CNN model for far-field ASR, where the data from all the microphones are chosen as input to the ASR system without a beamforming step.

2.7.3.3 Joint modeling of enhancement and ASR

The initial attempt proposed by Wang et. al. [18] incorporates a DNN based speech separation model coupled with a DNN based acoustic model. The key idea is to concatenate a deep neural network (DNN) based speech separation frontend and a DNN- based acoustic model to build a larger neural network, and jointly adjust the weights in each module. This way, the separation frontend is able to provide enhanced speech desired by the acoustic model and the acoustic model can guide the separation frontend to produce more discriminative enhancement.Sequence training is applied to the jointly trained DNN so that the linguistic information contained in the acoustic and language models can be back-propagated to influence the separation frontend at the training stage. To further improve the robustness, more noise- and reverberation-robust features were added.

In Bo Wu et. al. [19], unification of separately trained speech enhancement neural model and the acoustic model was proposed, in which the unified or the joint model is further trained to improve the ASR performance. a unified deep neural network (DNN) approach to achieve both high-quality enhanced speech and high-accuracy automatic speech recognition (ASR) simultaneously on the recent REverberant Voice Enhancement and Recognition Benchmark (RE-VERB) Challenge. These two goals are accomplished by two proposed techniques, namely DNN-based regression to enhance reverberant and noisy speech, followed by DNN-based multi-condition training that takes clean-condition, multi-condition and enhanced speech all into consideration. They first report on superior objective measures in enhanced speech to those listed in the 2014 REVERB Challenge Workshop. They then show that in clean-condition training, we obtain the best word error rate (WER) of 13.28% on the 1-channel REVERB simulated evaluation data with the proposed DNN-based pre-processing scheme. A competitive single-system WER of 8.75% with the proposed multi-condition training strategy and the same less-discriminative log power spectrum features used in the enhancement stage is obtained. Finally by leveraging upon joint training with more discriminative ASR features and improved neural network based language models a state-of-the-art WER of 4.46% is attained with a single ASR system, and single-channel information. Another state-of-the-art WER of 4.10% is achieved through system combination.

We propose an integrated end-to-end automatic speech recognition (ASR) paradigm by joint learning of the front-end speech signal processing and back-end acoustic modeling. We believe that "only good signal processing can lead to top ASR performance" in challenging acoustic environments. This notion leads to a unified deep neural network (DNN) framework for distant speech processing that can achieve both high-quality enhanced speech and high-accuracy ASR simultaneously. Our goal is accomplished by two techniques, namely: (i) a reverberation-time-aware DNN based speech dereverberation architecture that can handle a wide range of reverberation times to enhance speech quality of reverberant and noisy speech, followed by (ii) DNN-based multicondition training that takes both clean-condition and multicondition speech into consideration, leveraging upon an exploitation of the data acquired and processed with multichannel microphone arrays, to improve ASR performance. The final end-to-end system is established by a joint optimization of the speech enhancement and recognition DNNs. The recent REverberant Voice Enhancement and Recognition Benchmark (REVERB) Challenge task is used as a test bed for evaluating our proposed framework. We first report on superior objective measures in enhanced speech to those listed in the 2014 REVERB Challenge Workshop on the simulated data test set. Moreover, we obtain the best single-system word error rate (WER) of 13.28% on the 1-channel REVERB simulated data with the proposed DNN-based pre-processing algorithm and clean-condition training. Leveraging upon joint training with more discriminative ASR features and improved neural network based language models, a low single-system WER of 4.46% is attained. Next, a new multi-channel-condition joint learning and testing scheme delivers a state-of-the-art WER of 3.76% on the 8-channel simulated data with a single ASR system. Finally, we also report on a preliminary yet promising experimentation with the REVERB real test data.

In Bo Wu et. al. [19], unification of separately trained speech enhancement neural model and the acoustic model was proposed, in which the unified or the joint model is further trained to improve the ASR performance. Another work of Bo Wu et. al. [20] also explored an end-to-end deep learning approach, where, the DNN based dereverberation front end leverages the knowledge about the reverberation time. The ASR cost was further improved by jointly training this reverberation time aware-DNN enhancement module and the ASR acoustic module.

Recently, Tesch et. al. [79] proposed a non- linear spatial filter realized by a DNN as well as its interdepen- dency with temporal and spectral processing by carefully con- trolling the information sources (spatial, spectral, and temporal) available to the network. In a traditional setting, linear spatial filtering (beamforming) and single-channel post-filtering are commonly performed separately. Their analyses revealed that in particular spectral information should be processed jointly with spatial information as this increases the spatial selectivity of the filter.

In [80], a full-band and sub-band fusion model, named as FullSubNet, for single-channel real-time speech enhancement was proposed. Full-band and sub-band refer to the models that input full-band and sub-band noisy spectral feature, output full-band and sub-band speech target, respectively. The sub-band model processes each frequency independently. Its input consists of one frequency and several context frequencies. The output is the prediction of the clean speech target for the corresponding frequency. These two types of models have distinct characteristics. The full-band model can capture the global spectral context and the long-distance cross- band dependencies. However, it lacks the ability to modeling signal stationarity and attending the local spectral pattern. The sub-band model is just the opposite. In our proposed FullSubNet, we connect a pure full-band model and a pure sub-band model sequentially and use practical joint training to integrate these two types of models' advantages.

In the work by Zhou et. al. [81], a new learning target based on reverberation time shortening (RTS) for speech dereverberation was proposed. The learning tar- get for dereverberation is usually set as the direct-path speech or optionally with some early reflections. This type of target suddenly truncates the reverberation, and thus it may not be suitable for net- work training. The proposed RTS target suppresses reverberation and meanwhile maintains the exponential decaying property of re- verberation, which will ease the network training, and thus reduce signal distortion caused by the prediction error. Moreover, this work experimentally study to adapt our previously proposed FullSubNet speech denoising network to speech dereverberation. Experiments show that RTS is a more suitable learning target than direct-path speech and early reflections, in terms of better suppressing reverber- ation and signal distortion. FullSubNet [80] is able to achieve outstanding dereverberation performance.

2.8 Chapter summary

This chapter discussed about the details of frequency domain linear prediction. It also brief about different beamforming methods. A discussion on automatic speech recognition can also be found. This is followed by the discussion on performance measures for speech recognition and enhancement methods. This is followed by a survey of speech enhancement approaches in the spectral domain.

Chapter 3

Dereverberation of Sub-band Temporal Envelopes for Far-Field ASR

In this chapter we propose two methods of dereverberation for far-field ASR. In the first part, we have proposed a new framework of multi-channel feature extraction using spatio-spectral autoregressive (SSAR) modeling. In this method, we propose a 3-D CLSTM model for neural beamforming that allows the modeling of multi-channel audio features directly. The experiments are conducted on the CHiME-3 and REVERB Challenge dataset using multi-channel reverberant speech. In these experiments, the proposed 3-D feature and acoustic modeling approach provides significant improvements over an ASR system trained with beamformed audio (average relative improvements of 16% and 6% in word error rates for CHiME-3 and REVERB Challenge datasets respectively).

In the second part, we propose a new neural model for dereverberation of temporal envelopes and joint learning of the acoustic model to improve the ASR cost. The joint learning framework combines the envelope dereverberation framework, feature pre-processing and acoustic modeling into a single neural pipeline. This framework is elegant and the model can be learned using a joint loss function. Several experiments are performed on the REVERB challenge

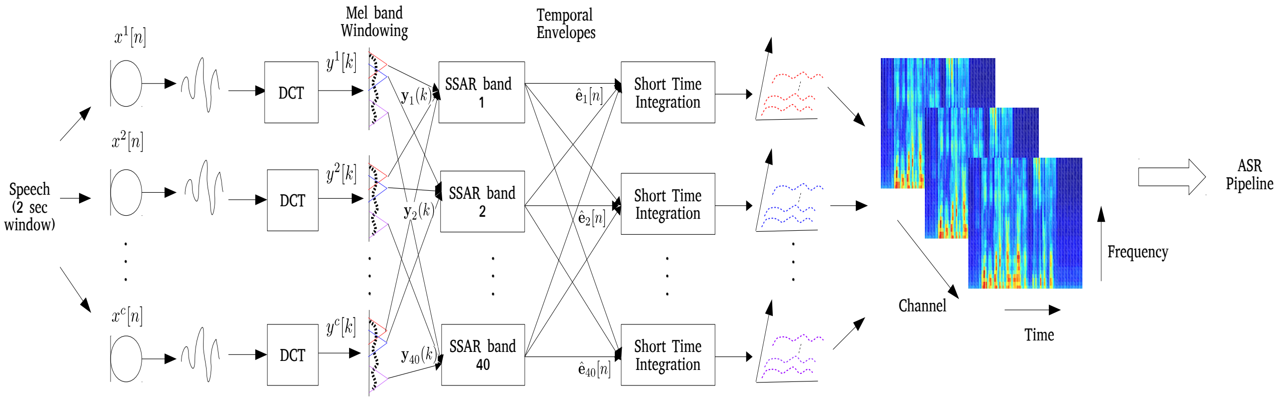


Figure 3.1: *Block schematic of the 3-D feature extraction method using SSAR modeling.*

dataset, CHiME-3 dataset and VOiCES dataset. In these experiments, the joint learning of envelope dereverberation and acoustic model yields significant performance improvements over the baseline ASR system based on log-mel spectrogram as well as other past approaches for dereverberation (average relative improvements of 10-24% over the baseline system). A detailed analysis on the choice of hyper-parameters and the cost function involved in envelope dereverberation is also provided.

Rest of the chapter is organized as follows. Section 3.1 discusses the details of the proposed 3-D acoustic modeling framework. Details about the proposed dereverberation method on temporal envelopes can be found in section 3.2.

3.1 3-D Acoustic modeling for far-field multi-channel speech recognition

Previously, many works have focused on far-field speech recognition using multiple microphones [82, 83]. The technique of beamforming attempts to find the time delay between channels and boosts the signal by weighted and delayed summation of the individual channels [84, 85]. This approach is still the most widely used technique for ASR in multi-channel reverberant environments [86]. We propose an approach to avoid the beamforming step by exploiting multi-channel features within the ASR framework. A feature extraction step is proposed, that is based on autoregressive (AR) modeling exploiting the joint correlation among the frequency dimensions and channel dimensions present in the signal. The block schematic is shown in Figure 3.1 The AR approach efficiently models the peaks in the spatio-spectral (SS) domain of the multi-channel signal. We refer to these features as spatio-spectral autoregressive (SSAR) features. Then, a novel neural network architecture is proposed for multi-channel ASR which contains network-in-network (NIN) in a 3-D convolutional neural network (CNN) architecture.

3.1.1 Related Prior Work

While the original goal of beamforming [17] is directed towards signal enhancement, the beamforming cost can be modified for maximizing the likelihood [87]. Recently, Swietojanski *et al.* [88] proposed the use of features from each channel of the multi-channel speech directly as input to a convolutional neural network based acoustic model. Here, the neural network is seen as a replacement for conventional beamformer. Joint training of a more explicit beamformer with the neural network acoustic model has been proposed by Xiao *et al.*, [89]. Training of neural networks, which operate on the raw signals that are optimized for the discriminative cost function of the acoustic model, has also been recently explored. These approaches are termed as *Neural Beamforming* approaches as the neural network acoustic model subsumes the functionality of the beamformer [90, 91]. Previously, Ganapathy *et al.* had explored the use of 3-D CNN models in [92], where the network was fed with the spectrogram features of all channels. The beamforming using a deep neural network based mask estimation method and a generalized eigenvalue approach has also shown improved results for ASR [93]. However, the DNN mask based methods require parallel noisy and clean data to train the mask estimators.

3.1.2 Spatio-Spectral Autoregressive (SSAR) Model For Feature Extraction

The FDLP analysis, presented in Section 2.1, is extended to model multi-channel speech. The

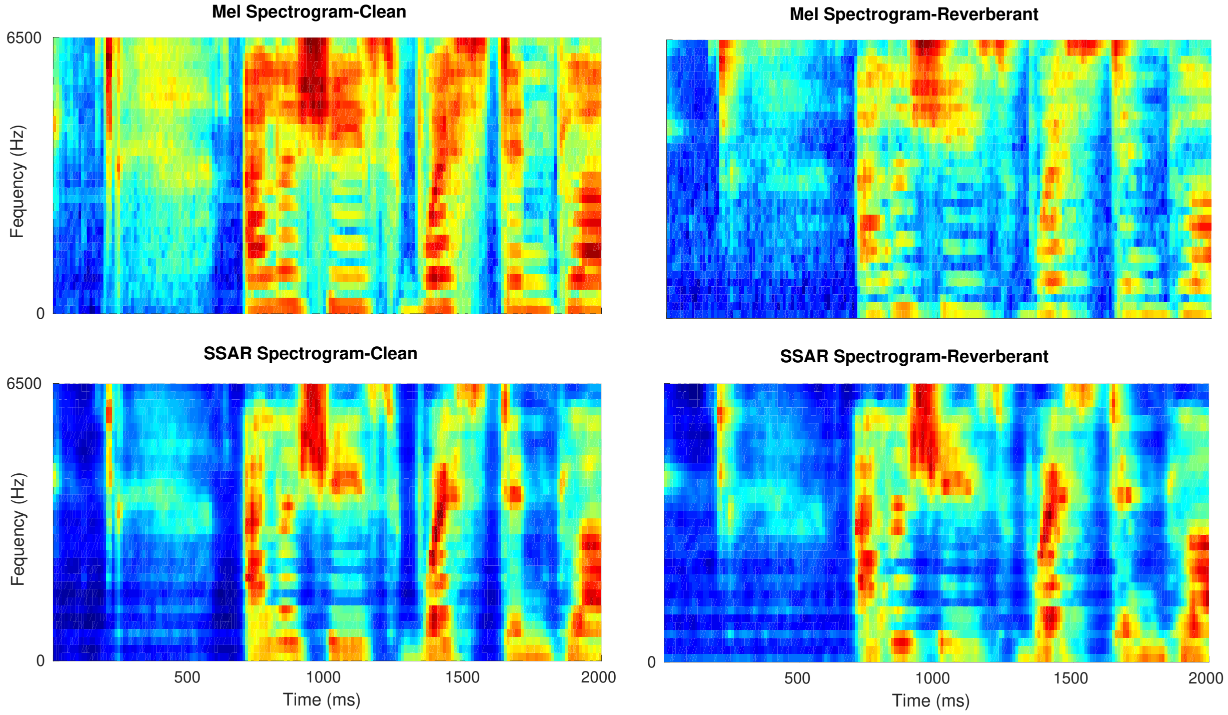


Figure 3.2: *Comparison of spectrogram estimation using SSAR modeling with conventional mel spectrogram for clean (near-room) and reverberant speech (far-room) recordings from the REVERB Challenge dataset.*

proposed feature extraction is shown in Fig. 3.1. Let the multi-channel signal be denoted as $x^c[n]$ for $c = 1, \ldots, C$ channels. Each $x^c[n]$ is processed using a long-window (2000 ms) discrete cosine transform (DCT) to generate $y^c[k]$. The DCT signal is decomposed into sub-bands using mel spaced Gaussian shaped windows. Since the channels capture the same underlying acoustic environment, their sub-band DCT components are highly correlated. These correlations can be captured in a spatio-spectral autoregressive model.

Let C denote the number of channels and let $y_i^c(k)$ denote the DCT transformed signal for the i^{th} sub-band, c^{th} channel and k denotes the DCT component index. Let $\mathbf{y}_i(k) = [y_i^1(k), \ldots, y_i^C(k)]^T$ denote the C dimensional vector containing the k^{th} DCT component for the

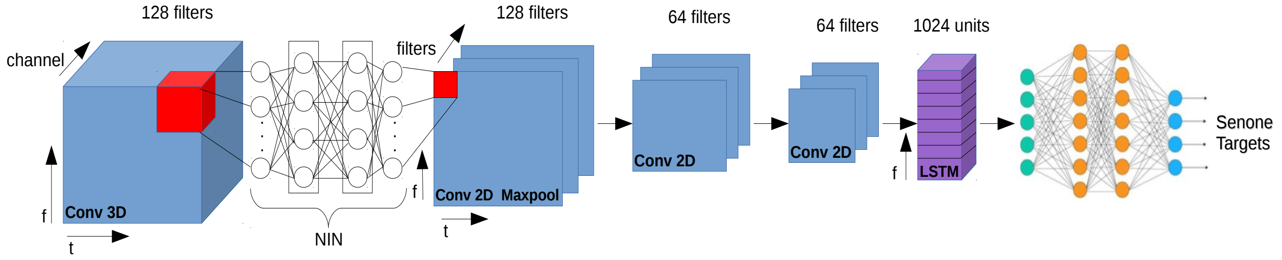


Figure 3.3: *3-D Conv-LSTM (3-D CLSTM) architecture used in multi-channel ASR which has NIN 1^{st} layer performing 3-D CNN, 2-D CNN, and LSTM layers.*

i^{th} band for all the spatial channel locations. The spatio-spectral AR model is given by,

$$\mathbf{y}_i(k) = \sum_{l=1}^{p} \mathbf{A}_i^l \mathbf{y}_i(k-l) + \boldsymbol{\epsilon}_i(k) \tag{3.1}$$

where p is the AR model order, $\mathbf{A}_i^l$ denotes the multi-variate AR model coefficient matrices [94] for the i^{th} sub-band and $\boldsymbol{\epsilon}_i$ denotes the residual noise with a co-variance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. The model parameters $\mathbf{A}_i^l$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ can be solved using an auto-correlation analysis based method [94]. As in conventional linear prediction, the model parameters of Eq. (3.1) represent the AR model of the envelope of the sub-band i. The forward prediction model in the frequency domain is characterized by the multi-dimensional filter $\mathbf{H}_i[z]$, where $\mathbf{H}_i[z] = \mathbf{I}_C + \mathbf{A}_i^1 z^{-1} + \mathbf{A}_i^2 z^{-2} + \ldots + \mathbf{A}_i^p z^{-p}$

If $\mathbf{e}_i[n]$ denotes the envelope for sub-band i for all the C channels, i.e., ($\mathbf{e}_i[n] = \left[e_i^1[n] \; e_i^2[n] \ldots e_i^C[n] \right]^T$), then the multi-variate AR estimate is,

$$\hat{\mathbf{e}}_i[n] = diag(\mathbf{H}_i[n]^{-1} \hat{\boldsymbol{\Sigma}}_e \mathbf{H}_i^*[n]^{-1}) \tag{3.2}$$

where $\mathbf{H}_i[n] = \mathbf{H}_i[z]|_{z=\exp^{-j2\pi n}}$, the multi-variate AR estimate, $\hat{\mathbf{e}}_i[n] = \left[\hat{e}_i^1[n] \; \hat{e}_i^2[n] \ldots \hat{e}_i^C[n] \right]^T$ and $\mathbf{H}^*[n]$ is the conjugate-transpose of $\mathbf{H}[n]$. By estimating $\hat{\mathbf{e}}_i[n]$ for each sub band, we reconstruct the temporal envelopes of all the channels and all sub bands. The sub band envelopes $\hat{\mathbf{e}}_i[n]$ are integrated with a Hamming window over a 25 ms window with a 10 ms shift. The feature vector for a given channel is obtained by splicing the integrated envelopes of all sub-bands.

The spectrographic representation from the proposed SSAR approach is compared with the mel spectrogram in Fig. 3.2. In this figure, a region of the "clean" speech spectrogram (recorded from multiple channels in a near room environment) is compared with spectrogram for the "reverberant" condition (recorded from multiple microphones in far room environment). These files are part of the REVERB Challenge dataset. As seen in this figure, the spectrogram representation from the mel filter-bank analysis is modified by the effect of reverberation. In contrast, the proposed SSAR model spectrogram, which focuses on the peaks in the spatio-spectral regions of the audio signal, is more robust to reverberant artifacts. In the ASR experiments described later, we show that this peak modeling property improves the WER performance in far field conditions.

3.1.3 3-D Multi-channel Acoustic Model

The proposed 3-D CLSTM architecture is shown in Fig. (3.3). The input data to the model consists of 21 frames and 40 bands from all the C channels. The input data to 3-D CLSTM model is a 3-D tensor of size $C \times 21 \times 40$ in the first layer. The first layer implements a network-in-network (NIN) based 3-D convolution operations [95]. The next layer is a 2-D CNN layer with 128 kernels of size 3×3 in the second layer. This is followed by max-pooling and two 2-D CNN layers with 64 filters of kernel size 3×3. The output of the convolution layers is fed to an LSTM layer which performs frequency recurrence operations. This is followed by two fully connected layers of 1024 neurons and the output layer of softmax activations predicts the senone classes. We use dropout (DP) of 10% for the two hidden layers in NIN, while 20% dropout is added to every layer following NIN. The batch normalization technique is applied after activation for regularization. The cross entropy loss is used as the training criterion with Adam optimizer. The model training is performed using PyTorch software. The first layer performs the equivalent of neural beamforming while successive layers have only 2-D $t \times f$

Table 3.1: *Word Error Rate (%) in CHiME-3 dataset for beamformed audio for different baseline model architectures with FBANK features.*

Experiments	Dev		Eval	
	Real	Simu	Real	Simu
BF-FBANK-DNN	8.9	10.9	17.0	17.3
BF-FBANK-2D CNN	7.3	9.9	14.3	15.6
+Dropout	6.5	9.0	13.6	13.8
+Batchnorm, Adam	6.4	8.9	**13.5**	13.7
+LSTM (2-D CLSTM)	**6.4**	**8.6**	13.8	**12.9**

representation. For the baseline experiments, the 2-D CLSTM architecture is used along with beamformed audio. The 2-D CLSTM architecture used in the case of beamformed audio, is a special case of the proposed 3-D CLSTM architecture, where the input is a 2-D spectrogram of size 21 × 40 and normal 2-D convolution operations are performed in the initial layer. The increase in the number of parameters for 3-D CLSTM model compared to 2-D CLSTM model is less than 1%. For this work, we also do not use mask based neural beamforming methods [93], as these models require a mask estimation with paired clean and noisy speech data.

3.1.4 Experiments and Results

The experiments are performed on CHiME-3 and REVERB Challenge datasets. For the baseline model, multiple architectures are experimented using beamformed signal processed with filterbank energy features (denoted as BF-FBANK). The FBANK features are 40 band log-mel spectrogram extracted every 25 ms windows with a shift of 10 ms. We use the Kaldi toolkit for deriving the senone alignments used in the PyTorch deep learning framework. A hidden Markov model - Gaussian mixture model (HMM-GMM) system is trained with MFCC (Mel Frequency Cepstral Coefficients) features to generate the alignments for training the CLSTM model. We use a recurrent neural network (RNN) based language model (LM) in the ASR decoding [96].

Table 3.2: *Word Error Rate (%) in CHiME-3 dataset for different features and model configurations.*

Experiments	Dev			Eval		
	Real	Simu	Avg	Real	Simu	Avg
BF-FBANK (2-D CLSTM)	6.1	8.4	7.3	13.0	12.7	12.9
MC-FBANK (3-D CLSTM)	7.2	7.2	7.2	15.4	9.1	12.2
SSAR (3-D CLSTM)	7.0	7.0	**7.0**	12.8	9.1	**10.9**

3.1.4.1 CHiME-3 ASR

The CHiME-3 dataset for the ASR has multiple microphone tablet device recording in four different environments, namely, public transport (BUS), cafe (CAF), street junction (STR) and pedestrian area (PED). For each of the above environments, real and simulated data are present. The real data consists of 6 channel recordings from WSJ0 corpus sampled at 16 kHz spoken in the four varied environments. The simulated data was made by mixing clean utterances with the environment noise. The training dataset consists of 1600 (real) noisy recordings and 7138 (simulated) noisy recordings from 83 speakers. The development (Dev) and evaluation (Eval) dataset consists of 1640 (410 × 4) recordings from 4 speakers and 1320 (330 × 4) recordings from 4 other speakers respectively.

3.1.4.2 CHiME-3 ASR Results

The ASR results for various feature and model architectures on the beamformed audio (beamformed from 5 channel recordings) are reported in Table 3.1. The 2-D CNN architecture gives a significant improvement over the DNN. Adding dropouts helped improve the performance further. Batch normalization and Adam optimizer also showed marginal improvement over the 2-D CNN model with dropout. Finally, we propose a new CLSTM architecture with the LSTM recurring over frequency. This served as the baseline for our experiments on the multi-channel data.

Table 3.3: *Word Error Rate (%) for different noise conditions in CHiME-3 dataset on SSAR with 3-D CLSTM architecture and the BF-FBANK 2-D CLSTM architecture (average results on Dev. and Eval. set.).*

Noise	BF-FBANK-2D-CLSTM		SSAR-3D-CLSTM	
	Dev	Eval	Dev	Eval
BUS	**7.0**	12.7	7.8	**8.2**
CAF	8.6	17.5	**8.0**	**11.7**
PED	5.8	13.5	**5.2**	**9.4**
STR	7.9	13.1	**7.0**	**10.1**

3.1.4.3 Discussion

The summary of the results for various feature and model configurations for the CHiME-3 ASR task, with one iteration of fine tuning on real development data are reported in Table 3.2. The beamformed audio with filter-bank features (BF-FBANK) provides the baseline results to compare other feature and model configurations. The MC-FBANK improves the baseline results by performing the neural beamforming using the 3-D CLSTM model without changing the feature representations. Finally, the combination of SSAR features and 3-D CLSTM model provides the best ASR performance. This approach yields average relative improvements of 4% on the development dataset and about 16% on the evaluation dataset over the baseline BF-FBANK configuration.

The results for various noise types in the CHiME-3 ASR evaluation for the baseline (BF-FBANK-2D-CLSTM) and the proposed 3-D feature and modeling (SSAR-3D-CLSTM) are reported in Table 3.3. The proposed approach yields consistent improvements on all the noise types in CHiME-3 dataset except in the development data for the BUS noise type. These results highlight the benefits of modeling the multi-channel audio in the joint time-frequency-channel domain during the feature extraction as well as in the acoustic modeling stages.

Table 3.4: *Word Error Rate (%) in REVERB dataset for different features and model configurations with RNN-LM.*

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

3.1.4.4 REVERB Challenge ASR

The REVERB Challenge dataset for ASR consists of 8 channel recordings with real and simulated reverberation conditions. The simulated data is comprised of reverberant utterances generated (from the WSJCAM0 corpus) obtained by artificially convolving clean WSJCAM0 recordings with the measured room impulse responses (RIRs) and adding noise at an SNR of 20 dB. The simulated data has six different reverberation conditions. The real data, which is comprised of utterances from the multi-channel Wall Street Journal audio-visual (MC-WSJ-AV) corpus, consists of utterances spoken by human speakers in a noisy reverberant room. The training set consists of 7861 utterances (92 speakers) from the clean WSJCAM0 training data by convolving the clean utterances with 24 measured RIRs. The development (Dev.) and evaluation (Eval.) datasets consists of 1663 (1484 simulated and 179 real) recordings and 2548 (2176 simulated and 372 real) recordings respectively. The Dev. and Eval. datasets have 20 and 28 speakers respectively. The ASR configuration for 2-D and 3-D models as well as the feature configurations used in CHiME-3 ASR setup are used in the same manner for these experiments. The Kaldi ASR build is used to generate the senone alignments that are used in the acoustic modeling performed in the PyTorch software.

The results for the various ASR experiments on REVERB Challenge are shown in Table 3.4. These results are consistent with those observed in the CHiME-3 dataset. Both the autoregressive model based feature extraction and the 3-D CLSTM modeling provide noticeable improvements over the baseline BF-FBANK-2D-CLSTM model. Further, the combined use of 3-D features and 3-D acoustic models yields significant improvements over the baseline. The proposed approach yields average relative improvements of 3% on the development dataset and about 6% on the evaluation dataset over the baseline BF-FBANK configuration.

3.1.5 Section Summary

In this section, we have proposed a new framework of multi-channel features using spatio-spectral autoregressive (SSAR) modeling. Using a 3-D CLSTM model for neural beamforming that allows the modeling of multi-channel audio features directly, we perform speech recognition experiments on the CHiME-3 dataset as well as on the REVERB Challenge dataset. These experiments indicate that the SSAR modeling in the spatio-spectral domain generates robust representations of speech. The use of multi-channel features in a 3-D convolutional long-short term memory (CLSTM) model architecture further improves the ASR performance. The analysis of results also highlight the incremental benefits achieved for various features and model configurations.

3.2 Dereverberation of Autoregressive Envelopes for Far-field Speech Recognition

In this direction of work, we analyze the effect of reverberation on sub-band Hilbert envelopes. We show that the effect of reverberation can be approximated as convolution of the long-term sub-band envelopes of clean speech with the envelope of room impulse response function. In order to compensate for the late reverberation component in the envelope, we explore a Wiener filtering approach where the Wiener filter gain is computed using a deep neural network (DNN). The gain estimation network is implemented using a convolutional-long short term memory (CLSTM) model. The gain is multiplied with the sub-band envelopes to suppress reverberation artifacts. The sub-band envelopes are converted to spectrographic features through integration and used for deep neural network based ASR. The sub-band envelopes are derived using the autoregressive modeling framework of frequency domain linear prediction (FDLP) [97, 98].

The steps involved in envelope dereverberation, feature extraction and acoustic modeling for ASR can all be implemented as neural network layers. Therefore, we also propose an approach for joint learning of the speech dereverberation model with the ASR acoustic modeling network as a single neural model. Various ASR experiments are performed on the REVERB challenge dataset [27] as well as the CHiME-3 dataset [99]. In these experiments, we show that the proposed approach improves over the state-of-the-art ASR systems based on log-mel features as well as other past approaches proposed for speech dereverberation and denoising based on deep learning. In addition, we also extend the approach to large vocabulary speech recognition on the VOiCES dataset [14, 13].

The related prior work is discussed in Section 3.2.1. This section also discusses the key contributions from the proposed work. Section 3.2.2 provides details regarding the reverberation artifacts and autoregressive envelope estimation using frequency domain linear prediction. In Section 3.2.3, we discuss the envelope dereverberation model, feature extraction as well as the joint approach to dereverberation with acoustic modeling for ASR. The ASR experiments and results are discussed in Section 3.2.4. Various model parameter choices and additional analyses are reported in Section 3.2.5. This is followed by a summary of the work in Section 3.2.6.

3.2.1 Related prior work

Xu et. al. in [63] attempted to find a mapping function from noisy and clean signals using supervised neural network, which is used for enhancement in the testing stage. In a similar manner, speech separation problem is also explored with ideal ratio mask based neural mapping [64]. Zhao et. al. proposed a LSTM model for late reflection prediction in the spectrogram domain for reverberant speech [65]. A spectral mapping approach using the log-magnitude inputs was attempted by Han et. al [66]. A mask based approach to dereverberation on the complex short-term Fourier transform domain was explored by Williamson et. al [67].

Speech enhancement for speech recognition based on neural networks has been explored in [69, 70, 71]. In Maas et. al [72], a recurrent neural network is used to map noise-corrupted input features to their corresponding clean versions. A context aware recurrent neural network based convolutional encoder-decoder architecture was used in [73] to map the power spectral features of noisy and clean speech. In a recent work by Pandey et. al [68], the speech enhancement is learned in the time domain itself, but using a matrix multiplication to convert the time domain signal into frequency domain and the frequency domain loss is used for training. This approach uses mean absolute error between the STFT frames of the clean and noisy speech for training.

The joint learning of the speech enhancement neural model and the acoustic model was attempted in [18]. Here, a DNN based speech separation model is coupled with a DNN based acoustic model and the weights are adjusted jointly. Bo Wu et. al. [19] proposed to unify the speech enhancement neural model and the acoustic model trained separately, and then the joint model is further trained to improve the ASR performance. The power spectrum in the log domain was used as features in the enhancement stage. Bo Wu et. al. [20] also explored an end-to-end deep learning approach in, where the knowledge about reverberation time is incorporated in DNN based dereverberation front end. This reverberation time aware-DNN enhancement module and ASR acoustic module are further trained jointly to improve the ASR cost.

The key contributions from this thread of work can be summarized as follows,

- Deriving a signal model for reverberation effects on sub-band speech envelopes and posing the dereverberation problem as a gain estimation problem.

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

- Joint learning of the dereverberation model parameters and the acoustic model for ASR in a single neural pipeline.

- Illustrating the performance benefits of the proposed approach for multiple ASR tasks.

We use FDLP features [100] for far-field speech. Further, several ASR experiments with the joint modeling approach are also conducted in this work.

3.2.2 Sub-band Envelopes - Effect of Reverberation and Autoregressive Estimation

We present the signal model for reverberation and the autoregressive model for estimating the sub-band envelopes [101, 102].

3.2.2.1 Signal model

When speech is recorded in far-field reverberant environment, the data collected in the microphone is modeled as

$$r(t) = x(t) * h(t), \tag{3.3}$$

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 $h(t) = h_e(t) + h_l(t)$, where $h_e(t)$ and $h_l(t)$ represent the early and late reflection components.

Let $x_q(n)$, $h_q(n)$ and $r_q(n)$ denote the decimated sub-band clean speech, room-response and the reverberant speech signal respectively. Here $q = 1, .., Q$ denotes the sub-band index and n denotes the decimated time-index (frame). Assuming an ideal band-pass filtering we can write (using Eq. 3.3),

$$r_q(n) = x_q(n) * h_q(n) \tag{3.4}$$

In the proposed model, we explore the modeling of the sub-band temporal envelopes. In order to extract the envelopes, the analytic signal based demodulation is proposed. The analytic representation of a real-valued signal is the complex signal consisting of the original function (real part) and the Hilbert transform (imaginary part). The negative frequency components of the analytic signal are zero-valued. By representing the real-valued functions in analytic domain, the extraction of the modulation components (like envelopes and carrier signals) is facilitated. Now, the analytic signal of the sub-band signal $r_q(n)$ is denoted as $r_{aq}(n)$, where $r_{aq}(n) = r_q(n) + j\mathcal{H}[r_q(n)]$. Here, $\mathcal{H}[.]$ is the Hilbert operator. It can be shown that [97],

$$r_{aq}(n) = \frac{1}{2}[x_{aq}(n) * h_{aq}(n)], \tag{3.5}$$

If two signals have a modulating envelope on the same modulating sinusoidal carrier signal (single AM-FM signal), the convolution operation of the two signals will have an envelope which is the convolution of the two envelopes, i.e., the envelope of the convolution of the two signals is the convolution of the envelope of the signals. For sub-band speech signals, this envelope convolution model will form a good approximation if the sub-band signals are narrow-band.

Then, for band-pass filters with narrow band-width, we get the following approximation between the sub-band envelope (defined as the magnitude of the analytic signal) components of the reverberant signal and those of the clean speech signal.

$$m_{rq}(n) \simeq \frac{1}{2} m_{xq}(n) * m_{hq}(n), \tag{3.6}$$

where $m_{rq}(n)$, $m_{xq}(n)$, $m_{hq}(n)$ denote the sub-band envelopes of reverberant speech, clean speech and room response respectively. We can further split the envelope into early and late reflection coefficients.

$$m_{rq}(n) = m_{rqe}(n) + m_{rql}(n), \tag{3.7}$$

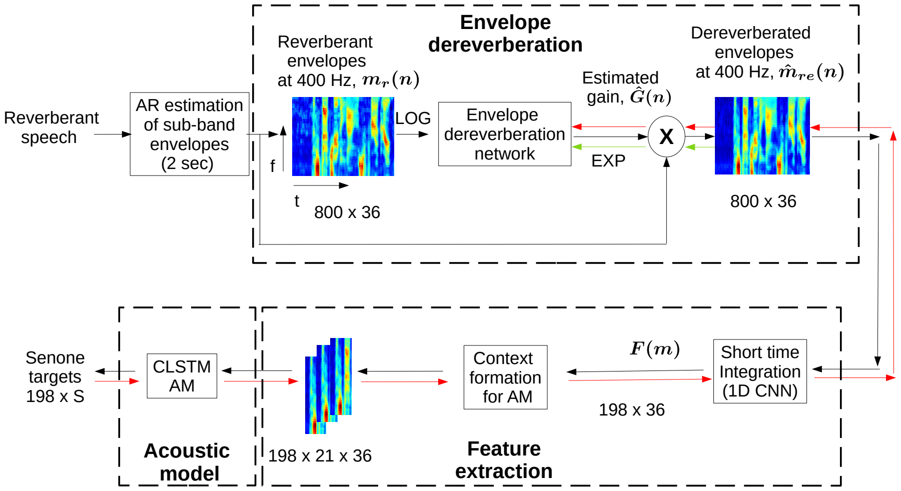

Envelope dereverberation
Reverberant speech
AR estimation of sub-band envelopes (2 sec)
Reverberant envelopes at 400 Hz, $m_r(n)$
f
t
800 x 36
LOG
Envelope dereverberation network
Estimated gain, $\hat{G}(n)$
EXP
X
Dereverberated envelopes at 400 Hz, $\hat{m}_{re}(n)$
800 x 36
Senone targets 198 x S
CLSTM AM
Acoustic model
198 x 21 x 36
Context formation for AM
Feature extraction
$F(m)$
198 x 36
Short time Integration (1D CNN)

Figure 3.4: *Block schematic of envelope dereverberation model, the feature extraction module and acoustic model (Here $m_r(n)$, $\hat{G}(n)$, $\hat{m}_{re}(n)$ and $F(m)$ are given by Eq.s (5, 8-11), the subscript q is dropped to indicate the fact that, signals from all the bands are considered). 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 (E_{CE}), and green arrows denote the backward propagation with mean square error loss (E_{MSE}). Here, S is the total number of senone targets.*

3.2.2.2 Autoregressive modeling of sub-band envelopes

Refer to Section 2.1 for details on AR modeling of sub-band envelopes. The LP envelope estimated using the prediction on the DCT components provides an all-pole model of the sub-band envelopes $m_{rq}(n)$.

3.2.3 Envelope Dereverberation and Joint Modeling

The proposed framework (Figure 3.4), consists of three modules, (i) envelope dereverberation, (ii) feature extraction and (iii) ASR acoustic model.

3.2.3.1 Neural dereverberation network

As seen in Eq. (3.7), the FDLP envelope of reverberant speech can be expressed as sum of the direct component (early reflection) and those with the late reflection. In the envelope dereverberation model, our aim is to input the envelope of the reverberant sub-band temporal envelope $m_{rq}(n)$ to predict the late reflection components $m_{rql}(n)$. Once this prediction is achieved, the late reflection component can be subtracted from the sub-band envelope to suppress the artifacts of reverberation. A similar analogy to this envelope subtraction approach is the spectral subtraction model where the noise and clean power spectral density (PSD) gets added in noisy speech PSD. If Gaussian assumptions are made for PSD components [103], the Wiener filtering approach to noisy speech enhancement provides the minimum mean squared error, where the noisy PSD is multiplied by the gain of the filter. In a similar manner, we pose the dereverberation problem as an envelope gain estimation problem.

The envelope gain (G_q) is defined as,

$$G_q(n) = \frac{\hat{m}_{rqe}(n)}{\hat{m}_{rqe}(n) + \hat{m}_{rql}(n)} \tag{3.8}$$

The gain $G_q(n)$ is estimated using the input sub-band envelope $m_{rq}(n)$. With the gain estimate, the dereverberated envelope can be computed as,

$$\hat{m}_{rqe}(n) = G_q(n) m_{rq}(n) \tag{3.9}$$

The product model of enhancement is inspired by Wiener filtering principles. This sub-band envelope gain estimation is achieved using a deep neural network model in the proposed work. Following the model training, the dereverberation is achieved by multiplying the estimated sub-band envelope gain with the sub-band envelope of reverberant speech.

The block schematic of the envelope dereverberation model is shown in Figure 3.4. The input to the dereverberation model is the FDLP sub-band envelope of the reverberant speech. The model is trained to learn the sub-band envelope gain, which is the ratio of the clean envelopes (direct component) with the reverberant envelopes. During the model training, the model inputs are either far-field microphone recordings or the simulated reverberant recordings.

The model targets are the envelope gain (Eq 8) computed using either the close talking/near-room microphone corresponding to the far-field microphone data, or the clean close-talking microphone data for the simulated reverberant training data. Thus, model is trained with paired data to estimate the gain.

As the envelopes and the gain parameters are positive in nature, the model implementation in the neural architecture uses a logarithmic transform at the input and the estimated gain is transformed by an exponential operation. Specifically, the input to the dereverberation model is the set of sub-band envelopes $\{log(m_{rq})(n)\}_{q=1}^{Q}$, where Q is the number of sub-bands. The model is trained to predict the log-gain $\{log(G_q)\}_{q=1}^{Q}$. The sub-band dereverberated envelope is,

$$\hat{m}_{rqe} = exp\big[(log(\hat{G}_q(n)) + log(m_{rq}(n))\big] \tag{3.10}$$

where $\hat{G}_q(n)$ is the estimate of the gain from the model.

The entire model developed is applicable only on long analysis windows (which are typically greater than the T60 of the room response function). Hence, the proposed approach operates on long temporal envelopes of the order of 2 sec. duration. From the reverberant speech and the corresponding clean speech, the FDLP sub-band envelopes corresponding to 2 sec. non-overlapping segments are extracted. If the input sampling rate is 16 kHz, a 2 sec. segment will correspond to 32k samples ($t = \{1..32000\}$). The FDLP envelopes are extracted at a sampling rate of 400 Hz. Thus, 2 sec. segment of audio corresponds to 800 envelope samples ($n = \{1..800\}$) for each sub-band.

The input 2-D data of sub-band envelopes (800 samples from 36 mel sub-bands) are fed to a set of convolutional layers where the first two layers have 32 filters each with kernels of size of 41 × 5. The next two CNN layers have 64 filters with 21 × 3 kernel size. All the CNN layer outputs with ReLU activations are zero padded to preserve the input size and no pooling operation is performed. The output of the CNN layers are reshaped to perform time domain recurrence using 3 layers of LSTM cells. The first two LSTM layers have 1024 cells while the last layer has 36 cells corresponding to the size of the target signal (envelope gain). The training criteria is based on the mean square error between the target and predicted output. The model is trained with stochastic gradient descent using Adam optimizer [104].

3.2.3.2 Feature Extraction and Acoustic Modeling

For feature extraction, the sub-band envelopes are integrated in short Hamming shaped windows of size 25 ms with a shift of 10 ms [105]. A 25 ms slice at 400 Hz sampling (FDLP envelopes are sampled at 400 Hz) corresponds to 10 samples and the hop size of 10 ms corresponds to 4 samples.

The windowed FDLP envelopes are multiplied with a Hamming shaped window (size of 10) and accumulated. This window is shifted by 4 samples. A log compression is applied to limit the dynamic range of values. Following this integration, a 2 sec. chunk of 800 × 36 sub-band FDLP envelopes becomes 198 × 36.

In particular, let $\hat{m}_{rqe}(n)$ denote the dereverberated sub-band envelope obtained using Eq. (3.10). Further, let $w(n)$ denote a Hamming window of size 10 (corresponding to 25 ms at 400Hz sampling). Then, the features for ASR are extracted as,

$$F_q(m) = log(\hat{m}_{rqe}(m) * w(m)) \tag{3.11}$$

where $*$ is the convolution operation, and F_q denotes the scalar feature of qth sub-band. Here, m denotes the feature frame index at 10ms sampling (100 Hz). The features for all the Q sub-bands are spliced to form the final feature vector for ASR model training.

The set of operations described above for short-term integration can be implemented as a 1-D CNN layer with a fixed Hamming shaped kernel size of 10 and a stride 4. A log non-linearity is applied on the convolution output.

The integrated envelopes are used as time-frequency representations for ASR training. A context of 21 frames, with 10 frames on the left and 10 frames on the right is used in the acoustic model training.

3.2.3.3 Acoustic Model

The architecture of the acoustic model is based on convolutional long short term memory (CLSTM) networks (Figure 3.4). The acoustic model corresponds to 2-D CLSTM network described in [100], consisting of 4 layers of CNN, a layer of LSTM with 1024 units performing recurrence over frequency and 3 fully connected layers with batch normalization.

3.2.3.4 Joint learning

As shown in Figure 3.4, the three modules of (i) envelope dereverberation, (ii) feature extraction and context formation and (iii) the ASR acoustic modeling can be combined into a single neural end-to-end framework[1]. The intermediate envelope integration step is implemented as a 1-layer of 1-D convolutions with Hamming shaped kernel and log non-linearity. The context creation for acoustic features in the given segment is also performed as a fixed 1-D convolution layer. In this manner, the entire processing pipeline can be performed using an elegant joint learning approach.

For generating mini-batches in the model training, a 2 sec. speech segment is read along with the corresponding frame level targets (198 frames of senone targets for the 2 sec. segment). The entire joint neural network is trained using a combination of ASR cross entropy training criterion and mean square loss between the clean and reverberant envelopes. The gradients from the ASR loss at the input of the acoustic model (computed for each senone target) is accumulated over all the frames in the given 2 sec. segment. This accumulated gradient is of size 198×36 which corresponds to the size of the integrated envelopes. This ASR loss function when further back-propagated through fixed 1-D CNN layer provides a gradient matrix of size 800×36. The gradient w.r.t. mean square error (MSE) between the target envelopes and the

[1] The implementation of the work can be found in https://github.com/iiscleap/FDLP_Envelope_Dereverberation

dereverberation model outputs is combined with the ASR based gradient for training the joint model. The two gradients are indicated by two different backward arrows in Figure 3.4.

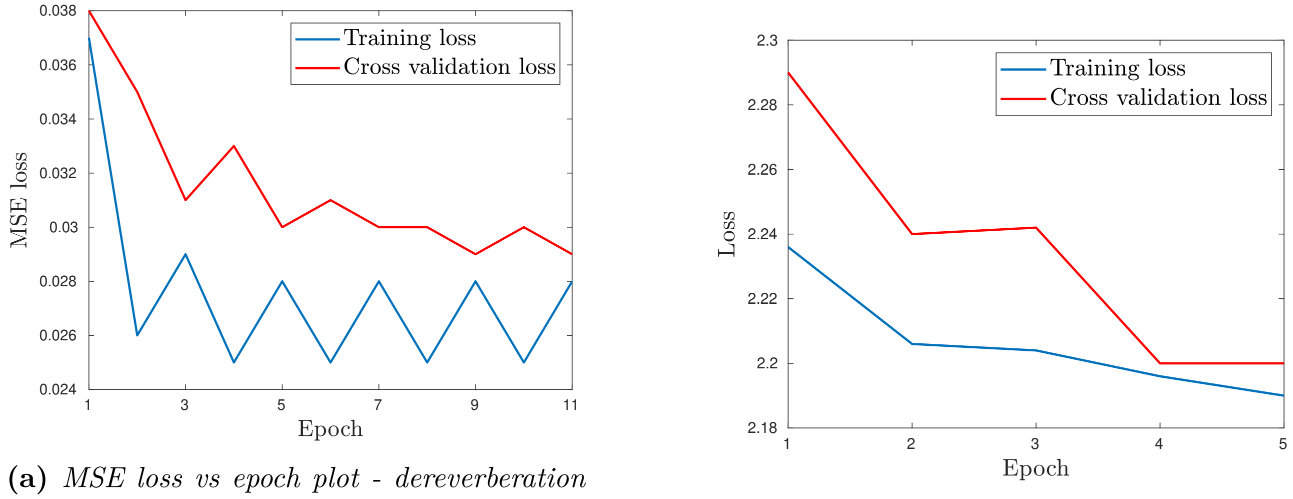
Joint loss function

The separate deverberation model is trained to minimize the mean square error loss, E_{MSE}, which is the squared error between the reverberant envelope and the clean counter part. For joint training, we have two loss functions, one is the mean square error loss, E_{MSE} for a mini-batch and the cross-entropy loss, E_{CE} between the senone targets and the corresponding posteriors for the same mini-batch. We use a combination of these two losses. Thus the final joint loss, E_{Total} is given by,

$$E_{Total} = E_{CE} + \mu \times E_{MSE}, \tag{3.12}$$

where μ is a regularization parameter, which decides the share of E_{MSE} in the joint loss, E_{Total}. In all our ASR experiments, we have used regularization parameter $\mu = 0.4$. The absolute value of the two loss functions (different dynamic range in Figure 3.5) does not have an impact as the model is trained with the gradient of the losses. The regularization constant μ controls the trade-off between the two loss functions. The variation of the MSE loss in the envelope dereveberation network is shown in Figure 3.5. The joint loss function on the training and validation data is also shown in this Figure. While the MSE loss trained with a higher learning rate exhibits oscillatory behavior, the joint loss function is relatively smooth. The final joint model is used in our ASR experiments.

A visualization of the dereverberation, achieved for the sub-band envelope of one single sub-band (10 th mel-band), is shown in Figure 3.6. The sub-band envelopes of reverberant signal deviate from their clean signal counterparts (as explained in Sec. 3.2.2). Using the dereverberation model proposed in this paper, we find that the FDLP envelopes are more closely matched with the clean signal envelopes. In Section 4.4, we compare the performance



(a) *MSE loss vs epoch plot - dereverberation model training*

(b) *Loss vs epoch plot - joint model training*

Figure 3.5: *Variation of training loss and cross validation loss with training epochs. The envelope dereverberation model and the ASR model are pre-trained before the joint learning step.*

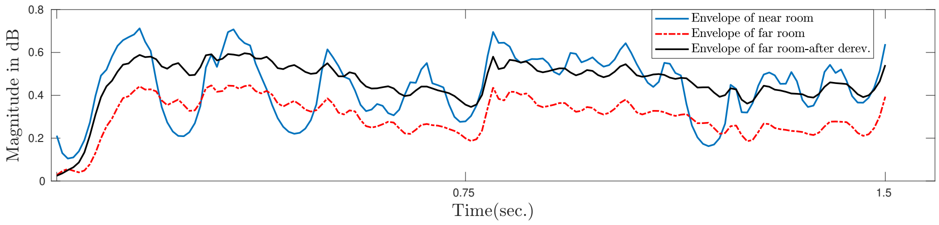


Figure 3.6: *Comparison of temporal envelopes, FDLP envelopes for clean, reverberant speech and reverberant speech after proposed joint learning based dereverberation, recordings from the REVERB Challenge dataset.*

of the CLSTM acoustic model architecture with other model architectures for dereverberation.

3.2.4 Experiments and results

The experiments are performed on REVERB challenge [27] and CHiME-3 [99] datasets. For the baseline model, we use WPE enhancement [16] along with unsupervised GEV beamforming [37]. This signal is processed with filter-bank energy features (denoted as BF-FBANK). The FBANK features are 36 band log-mel spectrogram with frequency range of $200-6500$ Hz. This is the same frequency decomposition used in the FDLP and FDLP-dereverberation experiments. The acoustic model is the 2-D CLSTM network described in [100].

Table 3.5: *Word Error Rate (%) in REVERB dataset for different features and proposed dereverberation method. Here prop denotes proposed work in this paper.*

Model Features	Dev			Eval		
	Real	**Simu**	**Avg**	**Real**	**Simu**	**Avg**
BF-FBANK	19.1	6.1	12.6	14.7	**6.5**	10.6
BF-FDLP [100]	17.8	6.8	12.3	14.0	7.0	10.5
BF-FBANK + CLSTM derevb. (prop)	17.3	5.5	11.4	13.1	6.9	10.0
BF-FBANK + spectral mapping derevb.[106]	15.8	**5.2**	10.5	12.8	6.7	9.7
BF-FBANK + context aware derevb.[73]	19.6	6.9	13.2	17.5	9.0	13.2
BF-FBANK + end to end derevb.[20]	-	-	-	24.8	7.9	16.4
BF-FDLP + CLSTM derevb. (prop)	16.3	5.6	10.9	13.4	7.1	10.2
BF-FDLP + CLSTM derevb. + joint (prop)	**15.2**	5.6	**10.4**	**12.1**	7.1	**9.6**

3.2.4.1 ASR framework

We use the Kaldi toolkit [107] for deriving the senone alignments used in the PyTorch deep learning framework for acoustic modeling. A hidden Markov model - Gaussian mixture model (HMM-GMM) system is trained with MFCC (Mel Frequency Cepstral Coefficients) features [108] to generate the alignments for training the CLSTM acoustic model. A tri-gram language model [109] is used in the ASR decoding and the best language model weight obtained from development set is used for the evaluation set.

3.2.4.2 REVERB Challenge ASR

The details of REVERB Challenge dataset is given in Section 3.1.4.4.

Discussion

Table 3.5 shows the WER results for experiments on REVERB challenge dataset. The WPE along with unsupervised GEV beamformed signal is used for all the ASR experiments (denoted as BF). The BF-FDLP baseline by itself is better than the BF-FBANK baseline (average relative improvements of 2% on the development set and about 1% on the evaluation set). For a fair comparision of the proposed approach, we have applied a similar dereverbaration method on BF-

FBANK baseline. Here, we have trained the neural model with log-mel features corresponding to 2 sec. duration with all the 36 mel-bands jointly. This approach is denoted as BF-FBANK + CLSTM derevb. (prop). Average relative improvements of 10% on the development set and about 6% on the evaluation set is achieved compared to the BF-FBANK baseline.

(BF-FBANK + spectral mapping derevb. [106]) corresponds to the work by Kun Han et. al. Here, a 3-layer deep neural network of 2570 units is used as the dereverberation neural model. The network is fed with 257-dimensional log-magnitude STFT features from a frame of 25 m.sec. A context window of 10-frames (5-left and 5-right) is selected and the network tries to predict the central frame. The work by Santos et. al. is implemented as (BF-FBANK + context aware dereverberation [73]). A CNN-GRU based encoder-decoder model is input with the entire utterance at the 257-STFT magnitude level features and is trained to predict the clean utterance. The results for end-to-end dereverberation network (joint learning) proposed in [20] is also compared with the proposed work in Table 3.5.

Finally, applying the proposed neural model based dereverberation on BF-FDLP baseline (denoted as BF-FDLP + CLSTM derevb. (prop)) yields average relative improvements of 13% on the development set and about 4% on the evaluation set, compared to the BF-FBANK baseline. After joint training this further improves to 17% and 9% respectively. The improvement in real condition is much more than that of simulated data. Average relative improvements of 20% on the real development set and about 18% on the real evaluation set, compared to the BF-FBANK baseline, is achieved by the proposed method. This suggests that, even though the jointly learned neural model is trained only with simulated reverberation, it generalizes well on unseen real data.

3.2.4.3 CHiME-3 ASR

Refer Section 3.1.4.1 for the details of CHiME-3 dataset.

Table 3.6: *Word Error Rate (%) in CHiME-3 dataset for different features and proposed dereverberation method. Here prop denotes proposed work in this paper.*

Model Features	Dev			Eval		
	Real	**Simu**	**Avg**	**Real**	**Simu**	**Avg**
BF-FBANK	7.8	8.0	7.9	14.0	9.7	11.8
BF-FDLP	7.0	8.1	7.5	12.0	10.0	11.0
BF-FBANK + CLSTM derevb. (prop)	7.2	8.3	7.7	12.9	9.8	11.4
BF-FBANK + spectral mapping derevb.[106]	8.0	10.0	9.0	14.3	12.3	13.3
BF-FBANK + context aware derevb.[73]	7.7	9.9	8.8	13.4	13.3	13.3
BF-FDLP + CLSTM derevb. (prop)	7.2	7.9	7.5	13	9.6	11.3
+ spec. reg. (prop)	**6.9**	8.0	7.4	11.8	9.8	10.8
+ spec. reg. + joint (prop)	7.0	**7.7**	**7.3**	**11.7**	**9.3**	**10.5**

Discussion

The WER results for experiments on CHiME-3 dataset are shown in Table 3.6. The FDLP baseline, denoted as BF-FDLP is better than the FBANK baseline (BF-FBANK). We observe average relative improvements of 8% on the development set and about 12% on the evaluation set when comparing BF-FDLP and BF-FBANK baseline systems. It can also be seen from Table 3.6 that the proposed dereverberation method improves the FBANK-baseline system. The results based on the implementation of works done by Han et. al. [106] and Santos et. al. [73] degrade the word error rates compared to the BF-FBANK baseline.

In the CHiME-3 dataset, we observed that the significant cause of degradation in the signal quality came from the additive noise sources. On further investigation, we found that the dereverberation model also resulted in smoothing of the spectral variations in the FDLP spectrogram. In order to circumvent this issue, we regularized the MSE loss with a term that encouraged the spectral channels to be uncorrelated. The regularization parameter was kept at 0.05. Using this regularized MSE loss, we further improved the BF-FDLP-Dereverberation system results over the dereverberation approach with MSE loss alone. These experiments suggest that even when the audio data does not contain significant late reflection components (like CHiME-3 dataset), the proposed approach improves significantly over the baseline method

Table 3.7: *WER in VOiCES dataset for different features and proposed dereverberation method. Here prop denotes proposed work in this paper.*

Model Architecture	Dev	Eval
BF-FBANK	55.5	66.6
BF-FDLP	51.5	62.6
BF-FDLP + CLSTM derevb. (prop)	52.8	62.4
+ joint. (prop)	**49.9**	**59.8**

(average relative improvements of 10.3 % over the baseline BF-FBANK system in the real development condition and 23.5 % on real evaluation condition).

3.2.4.4 VOiCES corpus ASR

Since the REVERB challenge dataset and CHiME-3 dataset are relatively smaller datasets, we wanted to establish the efficacy of the proposed dereverbaration method on a larger dataset. Thus, we experimented with VOiCES challenge dataset. VOiCES corpus [14] is released as part of "The voices from a distance challenge 2019" [13] of Interspeech 2019. For the ASR fixed conditons track, the training set consists of 80-hours subset of LibriSpeech corpus [110]. The training set has close talking microphone recordings from 427 different speakers from quiet environment. The development and evaluation sets consists of 19 hours and 20 hours of distant microphone recordings of varying room, environment and noise conditions. The significant difference between the training set and development/evaluation set makes the challenge even more difficult. We have used the same acoustic model configurations and hence these results reflect the impact of acoustic mismatch condition in ASR.

Discussion

The WER results for VOiCES corpus is given in Table 3.7. As seen, the baseline FDLP, denoted by BF-FDLP, provides at a better WER compared to the baseline FBANK. denoted as BF-FBANK. This is further improved with joint learning based dereverberation. The final WER shows improvement in both development set and evaluation set. A relative WER improvement

Table 3.8: *Word Error Rate (%) in REVERB dataset using different model architectures for dereverberation(without joint training)*

Architecture	**Dev**			**Eval**		
	Real	**Simu**	**Avg**	**Real**	**Simu**	**Avg**
BF-FDLP [100]	17.8	6.8	12.3	14.0	7.0	10.5
Neural Dereverberation						
3 layer DNN	17.3	5.4	11.3	14.2	6.9	10.5
5 layer CNN	**16**	5.9	11	13.5	7.2	10.3
4 layer CNN + 3 layer DNN	17.9	5.6	11.7	14.4	**6.7**	10.5
7 layer LSTM (1024 units each)	17	**5.3**	11.1	14.2	7.4	10.8
7 layer Resnet	17.4	7.9	12.6	14.8	10.3	12.5
CNN + DNN + LSTM [2,2,3]	19.1	6.8	13.0	15.5	7.9	11.7
CLSTM (4-CNN + 3-LSTM)	16.3	5.6	**10.9**	**13.4**	7.1	**10.2**

of 10% in both development set and evaluation set over the baseline FBANK system is observed in these experiments.

3.2.5 Analysis

In this section, the effect of different neural network architectures and various parameters like regularization parameter, λ, FDLP model order, p on WER are reported in Tables 4-6 and Figure 3.7.

3.2.5.1 Architecture of Dereverberation Model

Table 3.8 shows the WER for different neural network architectures. We initially explore a DNN of three feed forward layers. A slight improvement in development set is seen over the FDLP baseline, BF-FDLP. The relative improvement in WER becomes appreciable as we move to 5 layer CNN. The architecture with LSTM alone is promising. We also explore a Resnet [111] style architecture which was successful in image recognition. A combination of CNN, DNN and LSTM did not perform well compared to the baseline. Finally the CNN + LSTM combination

Table 3.9: *WER for various regularization to alleviate spectral smearing CHiME3 dataset. The regularization term is the cross correlation of the spectral bands. In absence of reverberation, alleviating spectral smearing improves the WER.*

Regularizer weight, λ	**Dev**			**Eval**		
	Real	**Simu**	**Avg**	**Real**	**Simu**	**Avg**
$\lambda = 0.3$	7	8.1	7.5	12.3	9.9	11.1
$\lambda = 0.1$	7	8.2	7.6	12.5	10	11.2
$\lambda = 0.05$	**6.9**	**8**	**7.4**	**11.8**	9.8	**10.8**
$\lambda = 0.02$	7.2	8.4	7.8	12.5	10	11.2
$\lambda = 0$	7.2	7.9	7.5	13	**9.6**	11.3
BF-FBANK	7.8	8.0	7.9	14.0	9.7	11.8

provides the best WER.

Table 3.10: *WER for various regularization to alleviate spectral smearing REVERB dataset. The regularization term is the cross correlation of the spectral bands. In the presence of significant reverberation, the extra regularization did not improve the performance.*

Regularizer weight, λ	**Dev**			**Eval**		
	Real	**Simu**	**Avg**	**Real**	**Simu**	**Avg**
$\lambda = 0.0$	**16.3**	5.6	**10.9**	**13.4**	7.1	**10.2**
$\lambda = 0.05$	16.6	**5.5**	11.0	14	6.9	10.4
$\lambda = 0.1$	16.9	5.6	11.2	13.7	6.9	10.3
$\lambda = 0.2$	17.2	5.5	11.3	14.2	6.8	10.5
BF-FBANK	19.1	6.1	12.6	14.7	**6.5**	10.6

3.2.5.2 Spectral Correlation Loss

As reported in Table 3.6 on CHiME-3 dataset, an extra term in the loss function which encourages the spectral bands to be uncorrelated improves the ASR performance on noisy data when the data is corrupted by additive noise with minimal reverberation artifacts. Table 3.9, 3.10 shows the effect of the regularization weight, λ on WER in CHiME-3 and REVERB datasets respectively for the spectral correlation loss used in the model learning. The introduction of the spectral correlation loss improves the WER in CHiME-3 dataset. The best results are obtained for a choice of $\lambda = 0.05$.

The introduction of spectral correlation loss does not benefit the REVERB challenge dataset. We hypothesize that this may due to the more dominant effect of temporal smearing seen in the REVERB challenge dataset. For the experiments on the VOiCES corpus, the spectral correlation loss was not used.

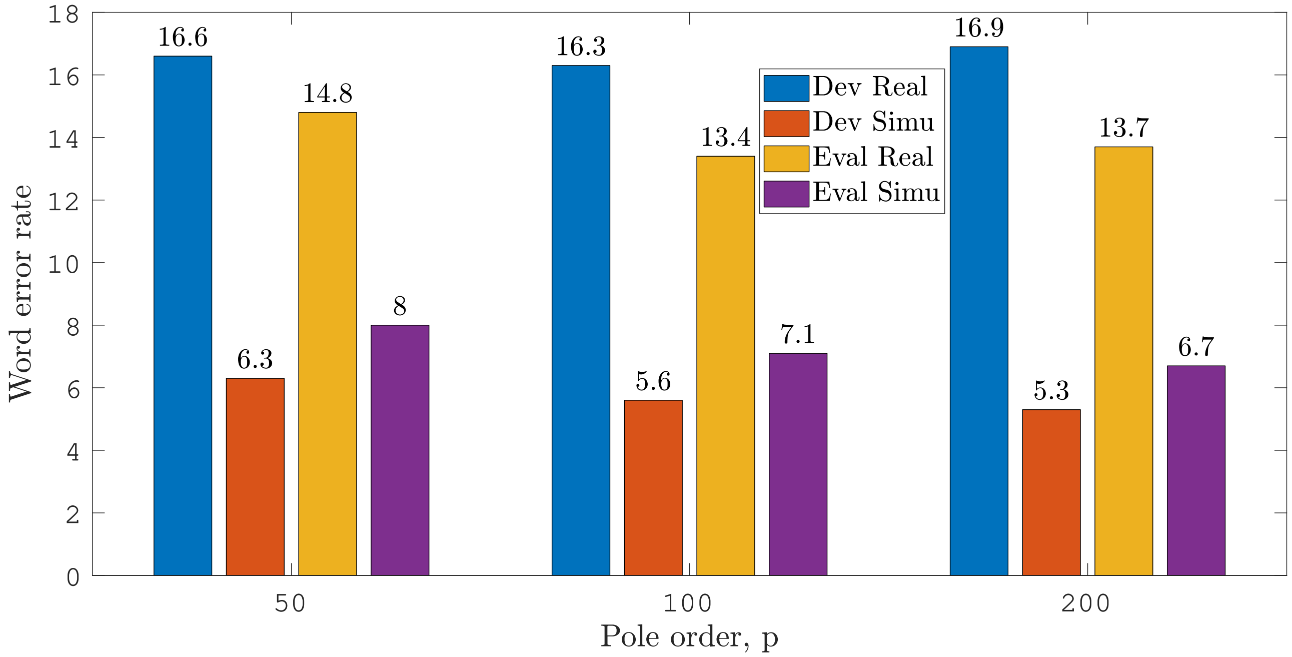


Figure 3.7: *WER (%) for various model-order, p in FDLP model for REVERB dataset*

3.2.5.3 Choice of FDLP model order

Figure 3.7 shows the effect of model order, p used in the FDLP envelope estimation on the WER for the REVERB challenge dataset. The model order p is the number of "past" samples used in the auto-regressive modeling of the sub-band DCT signal for a 2 sec window. While the WER results on the simulated conditions improve with higher model order of the FDLP, the performance on the real conditions is observed to be the best for about 100 poles per 2 sec. of audio in each sub-band. All the other experiments reported in the paper use the 100 poles per 2 sec. window of the audio signal.

3.2.5.4 Discussion on Performance Gains

All the results reported in Table 3.5, Table 3.6 and Table 3.7 use a strong baseline system with GEV based beamforming and weighted prediction error (WPE) based enhancement. Hence, we note that all systems use the same pre-processing pipeline and the gains observed over the baseline system are in addition to these enhancement steps. Furthermore, we also ensure that the baseline FBANK based system, the neural enhancement methods explored in the past and the proposed approach have the same sub-band decomposition, feature normalization, acoustic model and language model settings. In this way, the results highlight the effectiveness of the proposed work in suppressing reverberation distortions.

The methods proposed previously based on neural enhancement and dereverberation improve the performance of the baseline system on the REVERB challenge dataset. However, as seen in Table 3.6, in the presence of additive noise conditions on the CHiME-3 dataset, most of these prior works degrade the performance compared to the BF-FBANK baseline system. In this regard, the method proposed in this paper provides significant performance improvements on all three datasets. Further, the results consistently highlight the performance gains of using the joint neural learning framework.

3.2.6 Chapter Summary

In this section, we propose a new neural model for dereverberation of temporal envelopes and joint learning of the acoustic model to improve the ASR cost. The joint learning framework combines the envelope dereverberation framework, feature pre-processing and acoustic modeling into a single neural pipeline. This framework is hence elegant and the model can be learned using a joint loss function. Using the proposed neural dereverberation approach and joint learning, we perform speech recognition experiments on the REVERB challenge dataset as well as on the CHiME-3 dataset. These experiments indicate that the proposed neural dereverberation approach generalizes well on unseen conditions. The analysis of results also highlight the incremental benefits achieved for different choice of hyper-parameters and model architecture settings. The application of the proposed approach for large vocabulary speech recognition experiments on VOiCES dataset further emphasizes the performance benefits.

Chapter 4

Speech dereverberation with Envelope-Carrier Modeling

4.1 Introduction

One of the approaches to deal with the adverse far-field conditions is to develop a front-end which performs signal enhancement. Several techniques for dereverberation like signal processing based (for example, weighted prediction error (WPE) [16]), mask estimation based (for example, time-frequency mask estimation [67]) and multi-channel beamforming based (for example, time-delay estimation [17], generalized eigen-value [21, 37]) have been explored to improve the signal quality. On the other hand, another effective approach for system development in reverberant conditions is that of multi-condition training [112]. Here, either simulated or real far-field data is used to the train the models. However, even with these pre-processing and multi-condition training methods, the beamformed signal contains significant amount of temporal smoothing which adversely impacts the ASR performance [10]. In this chapter, we investigate the effect of reverberation on the sub-band signals of speech using an envelope-carrier decomposition. The extraction of the sub-band envelope is achieved using the autoregressive (AR) modeling approach in the spectral domain, termed as frequency domain linear prediction

(FDLP(Chapter 2)). Chapter 3 showed that a feature level enhancement with the FDLP envelope improves speech recognition performance [25, 26]. However, the prior works did not allow the reconstruction of the audio signal for quality improvement. Further, the enhancement of the carrier signal was not addressed in the previous work primarily due to the challenges in the dealing with the impulsive nature of the carrier signal.

In this chapter, we propose a novel approach to the joint dereverberation of the envelope and carrier signals using a neural modeling framework. We develop a dual path long short term memory (DPLSTM) architecture for the dereverberation of the temporal envelope and carrier signals. Following this step, the sub-band modulation and synthesis step generates the reconstructed audio signal. The neural enhancement and sub-band synthesis can also be implemented as a part of the larger neural pipeline for downstream tasks like ASR, thereby enabling the joint learning of the ASR and dereverberation model parameters. We refer to the proposed approach as Dual path dereverberation using Frequency domain Auto-Regressive modeling (DFAR) and the joint end-to-end model as E2E-DFAR.

Various ASR experiments are performed on the REVERB challenge dataset [27] as well as the VOiCES dataset [14, 13]. In these experiments, we show that the E2E-DFAR method improves over the state-of-the-art speech enhancement systems and ASR systems for the respective tasks.

4.1.1 Key contributions

The key contributions from this chapter, can be summarized as follows,

- Proposing an analysis for dereveberation with a sub-band decomposition (achieved using a perfect reconstruction quadrature mirror filter bank (QMF)) and envelope-carrier demodulation (achieved using frequency domain linear prediction (FDLP)).

- Proposing a dual-path long short time memory model named, DPLSTM for the dereverberation of sub-band envelope and carrier signals. This approach is termed as DFAR.

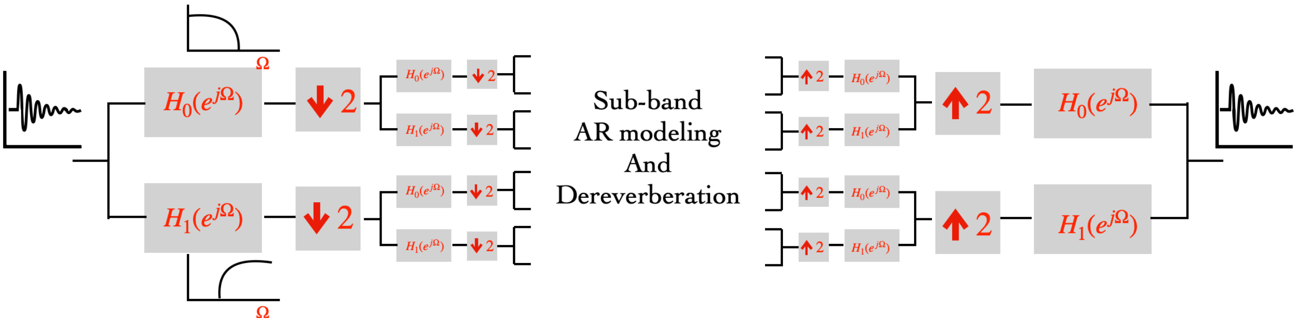


Figure 4.1: *Illustration of a 4-channel uniform QMF decomposition using a 2-stage binary QMF tree. In our work, we use* 64*-channel decomposition, using a 6-way binary tree.*

- Developing a joint learning scheme, where the ASR model and the DFAR model are optimized in a single end-to-end framework. This model is referred to as the E2E-DFAR.
- Evaluating the proposed approaches on speech quality improvement tasks as well as on ASR tasks on two benchmark datasets - REVERB challenge dataset and the VOiCES dataset.

The rest of the chapter is organized as follows. Section 4.2 provides details regarding the proposed approach. The ASR experiments and results are discussed in Section 4.4. Various model parameter choices and additional analyses are reported in Section 4.4.1.1. This is followed by a conclusion of the work in Section 4.5.

4.2 Proposed DFAR approach

In this section, we provide the details of the sub-band decomposition, frequency domain linear prediction and dual-path LSTM model, all of which form parts of the proposed DFAR model.

4.2.1 Quadrature Mirror Filter (QMF)

A quadrature mirror filter (QMF) is a filter whose magnitude response is a mirror reflection at quadrature frequency ($\frac{\pi}{2}$) of another filter. In signal processing, the QMF filter-pairs are used for the design of perfect reconstruction filter banks. Let $H_0(e^{j\Omega})$ and $H_1(e^{j\Omega})$ denote low-pass and high-pass filter's frequency domain function, where Ω is the digital frequency. In addition to the quadrature property ($H_1(e^{j\Omega}) = H_0(e^{j(\Omega-\pi)})$), the filters used in QMF filter-banks also satisfy the complimentary property,

$$|H_0(e^{j\Omega})|^2 + |H_1(e^{j\Omega})|^2 = 1. \tag{4.1}$$

The design of sub-band decomposition scheme with QMF involves a series of filtering and down-sampling operations for the analysis [113]. The synthesis is achieved by up-sampling and filtering operations. A tree-like structure can be formed using a recursive decomposition operation, thereby enabling a finer sub-band analysis. The down-sampling process enables a critical rate of processing, where the sum of the number of samples in each sub-band equals the number of the samples in the full-band signal.

In this work, we use an uniform 64-band Quadrature Mirror Filter bank (QMF) for decomposing the input signal into 64 uniformly spaced frequency bands. Inspired by the audio decomposition scheme outlined in Motlicek et. al. [114], we use a 6-level binary tree structure. The schematic of the sub-band decomposition is shown in Fig. 4.1. For the implementation in a neural pipeline, the down-sampling operation is equivalent to a stride, while the up-sampling operation is that of un-pooling.

4.2.2 Autoregressive modeling of temporal envelopes

The application of linear prediction model in the frequency domain, an approach called frequency domain linear prediction (FDLP), enables the modeling of the temporal envelopes of a signal with an autoregressive (AR) model [115, 105]. The sub-band signal is transformed to the spectral domain using a discrete cosine transform (DCT), where a linear prediction model is applied. This technique was discussed in detail in Chapter 2.

Let $x_q[n]$ and $e_q[n]$ denote the sub-band signal and envelope respectively, the corresponding carrier (remaining residual signal), $c_q[n]$ is found by sample wise division of the signal $x_q[n]$ by the estimated envelope $e_q[n]$.

$$c_q[n] = \frac{x_q[n]}{\sqrt{e_q[n]}} \tag{4.2}$$

The division operation in the expression above is well defined as the envelope given in Eq. (**??**) is always positive. Further, the modeling of the temporal envelopes using the AR model ensures that the peaks of the sub-band signal in the time-domain are well represented, which tend to belong to the higher signal-to-noise (SNR) regions of the speech signal [102, 116].

4.2.3 Effect of reverberation on envelope and carrier signals

The effect of reverberation on the time-domain speech signal can be expressed in the form of a convolution operation,

$$y[n] = x[n] * r[n], \tag{4.3}$$

where $x[n]$ denotes the clean speech signal, $r[n]$ is the impulse response of the room and $y[n]$, is the reverberant speech signal. The room response function can be further split into two parts, $r[n] = r_e[n] + r_l[n]$, where $r_e[n]$ and $r_l[n]$ are the early and late reflection components, respectively.

The details are already discussed in Chapter 3.

Envelope enhancement: A neural model can be used to learn late reflection component $e_{xql}[n]$ from the sub-band temporal envelope $e_{xq}[n]$. The predicted late reflection component can be subtracted from the sub-band envelope to suppress the artifacts of reverberation.

We pose the problem in the log domain to reduce the dynamic range of the envelope magnitude. The neural model is trained with reverberant sub-band envelopes $(log\ (e_{xq}[n]))$ as input. The model outputs the gain (in the log domain, i.e., $log\ \frac{e_{sq}[n]}{e_{xq}[n]}$). This gain is added in the log-domain to generate dereverberated signal envelope $(log\ (\hat{e}_{sq}[n])$.

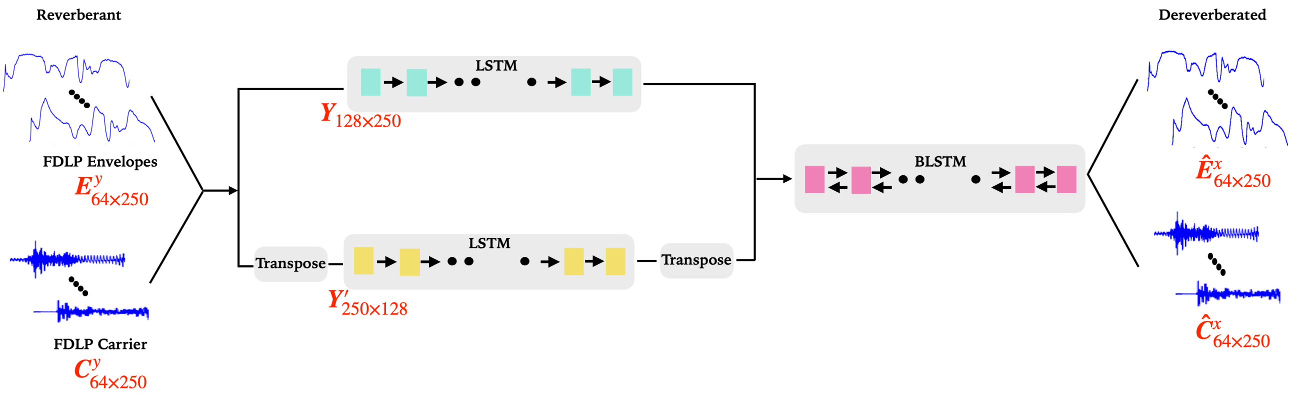


Figure 4.2: *The dual path LSTM model architecture for envelope-carrier dereverberation. The top LSTM path models the recurrence along the time dimension while the one on the bottom models the recurrence along the frequency dimension.*

Envelope-carrier dereverberation model: In a similar manner, the non-linear mapping between the reverberant carrier, $c_{xq}[n]$ and clean carrier, $c_{xq}[n]$, can be learned using a neural network. A neural model is trained with reverberant sub-band carrier ($c_{xq}[n]$) as input and model outputs the residual (an estimate of the late reflection component, $c_{xql}[n]$), which when added with the reverberant carrier generates the estimate of source signal carrier ($\hat{c}_{sq}[n]$). Instead of independent operations of dereverberation of the envelope and the carrier, we propose to learn the mapping between clean and reverberant versions of both the envelope and the carrier in a joint model. The input to the neural model is the sub-band reverberant envelope spliced with the corresponding carrier signal. The network is trained to output the late reflection components of both the envelope and carrier. With this approach, the model also learns the non-linear relationships between the envelope and carrier signals for the dereverberation task. From the model output, the estimate of the clean sub-band signal $\hat{s}_q[n]$ is generated. In our implementation, the audio signal is divided into non overlapping segments of 1 sec. length and passed through the envelope-carrier dereverberation model. The model is outlined in Fig. 4.3.

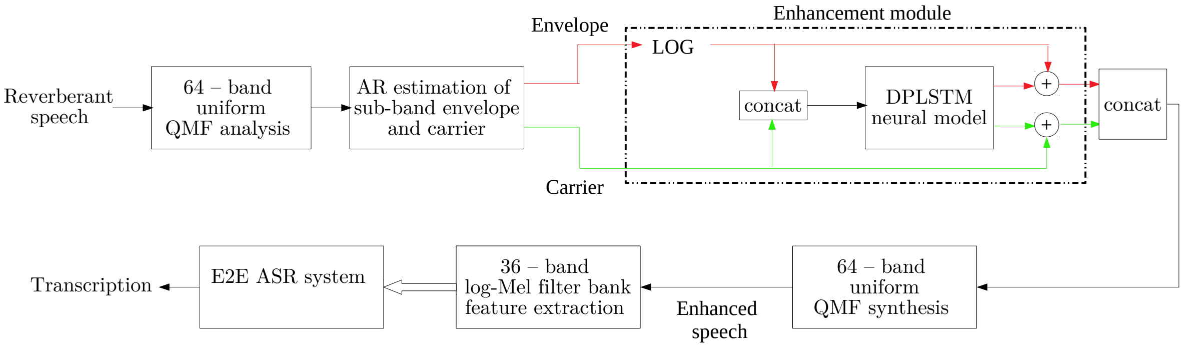

Enhancement module
Envelope
Reverberant speech
64 – band uniform QMF analysis
AR estimation of sub-band envelope and carrier
LOG
concat
DPLSTM neural model
+
+
concat
Carrier
Transcription
E2E ASR system
36 – band log-Mel filter bank feature extraction
Enhanced speech
64 – band uniform QMF synthesis

Figure 4.3: *Block schematic of speech dereverberation model, the feature extraction module and the E2E ASR model. The red arrows denote the envelopes, $e[n]$, and the green arrows represent the carrier, $c[n]$. The entire model can be constructed as an end-to-end neural framework.*

4.2.4 DFAR model architecture using DPLSTM

We propose the dual path long short term model (DPLSTM) for the dereverberation of the envelope-carrier components of the sub-band signal. Our proposed model is inspired by dual path RNN proposed by Luo et. al [117]. The block schematic of the DPLSTM model architecture is shown in Fig. 4.2. For 1 sec. of audio sampled at 16 kHz, the envelope ($\boldsymbol{E}^y$) and carrier ($\boldsymbol{C}^y$) components of the critically sampled sub-band signals (64 channel QMF decomposition) are of length 250. The envelope/carrier signals of all the sub-bands, for the reverberant signal ($\boldsymbol{Y}$), is of size 64 × 250. The combined envelope-carrier input is therefore of size 128 × 250, which forms the input to the DPLSTM model. The DPLSTM model outputs are also of the same size of the input, and the model is trained using the mean squared error (MSE) loss.

The proposed DPLSTM has two paths, one LSTM path models the recurrence along the time dimension, while the other LSTM path models the recurrence along the frequency dimension. We use two separate 3-layer LSTM architectures for these paths. The output dimensions are kept the same as the input dimension for each of these paths. The frequency recurrence LSTM output is transposed and these are concatenated in the frequency dimension. This combined output is fed to a multi layer bi-directional LSTM, which performs recurrence over time. The final output is split into sub-band specific envelope and carrier components. The modulation of the envelope with the respective carrier components generates the sub-band signals, which are passed through the QMF synthesis to generate the full-band dereverberated signal.

4.2.5 Joint learning of dereverberation model for ASR

The joint learning of the envelope-carrier dereverberation module with the E2E ASR architecture is achieved by combining the two separate models to train a single joint neural model. This is shown in Fig. 4.3. Given the deep architecture consisting of convolutions, LSTMs and transformer based layers, we initialize the modules with weights obtained from the independent training of each component. Specifically, the envelope-carrier dereverberation model is trained using MSE loss, which is followed by a sub-band synthesis (right side half of Fig. 4.1). The QMF synthesis is implemented using a 1-D CNN layer to generate the dereverberated speech signal. Further, the E2E ASR architecture is separately trained on the log-mel filter bank features, obtained from the dereverberated speech. The mel-filter bank feature generation can also implemented using a neural framework. Thus, the final model, composed of neural components from the envelope-carrier dereverberation, sub-band synthesis, feature extraction and ASR, can now be jointly optimized using the E2E ASR loss function. This model is refered to as E2E-DFAR model[1]. The trainable components are the DPLSTM model and the ASR model parameters, while the sub-band synthesis and feature extraction parameters are not learnable.

4.2.6 Visualization

Fig. 4.4 shows the clean, reverberant and dereverberated mel spectrograms of an utterance from the REVERB Challenge dataset (far-room). The reverberation effects are visible in the plot depicted in the second panel, where the temporal smearing blurs the details in the spectrogram. The dereverberated spectrogram, shown in the bottom panel, is able to retrieve some of the finer spectral details especially in low frequency regions. Experimental results shows that, these details are useful for in improving the ASR performance and for restoring the speech quality.

[1] The implementation of the work can be found in https://github.com/anurenjan/DFAR

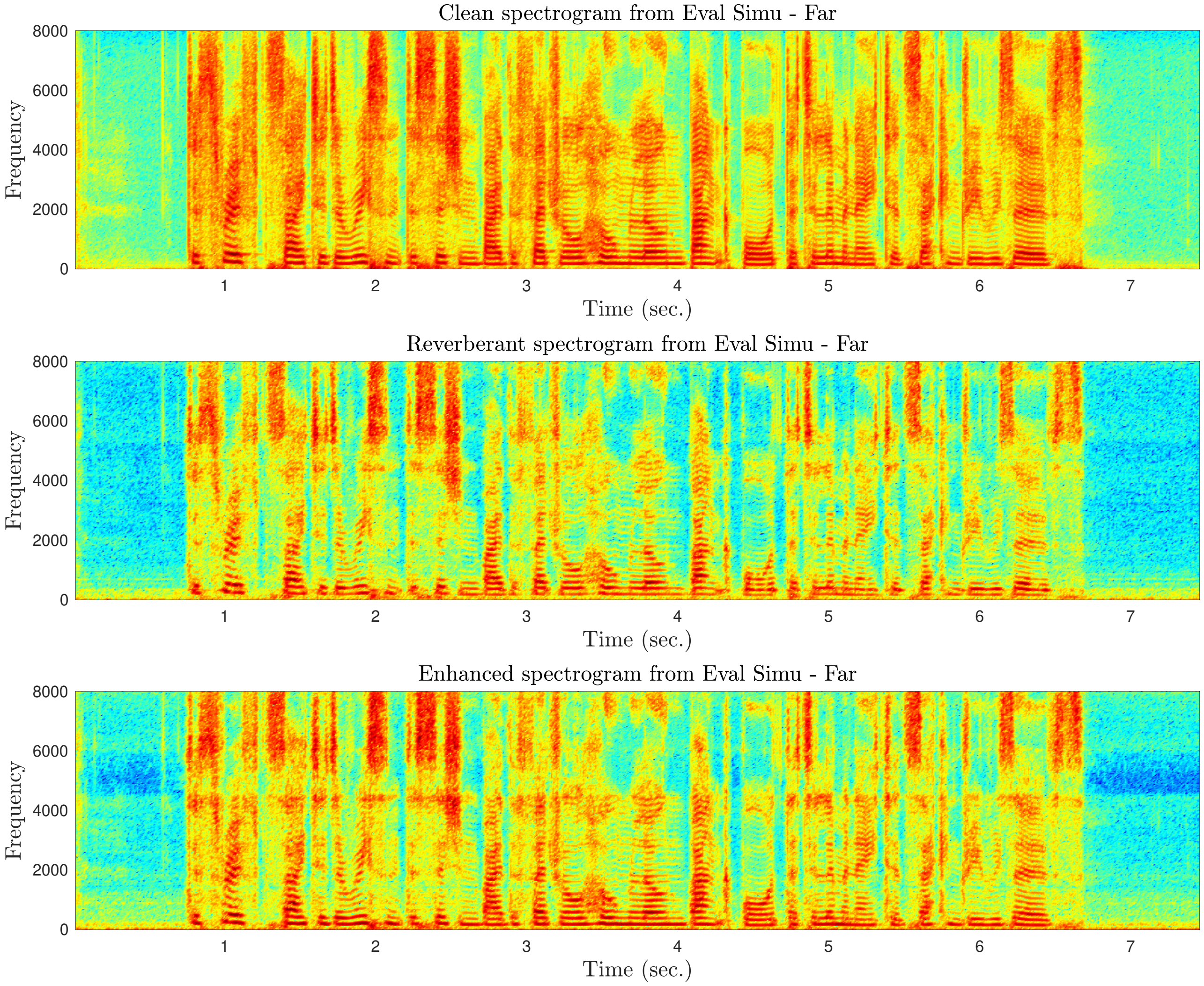


Figure 4.4: *Comparison of mel-spectrograms before and after dereverberation (last two plots) for reverberant speech (far-room) recordings from the REVERB challenge dataset. The clean spectrogram is also shown at the top for reference.*

4.3 Experimental setup

4.3.1 Datasets

4.3.1.1 REVERB Challenge ASR

Refer Section 3.1.4.4 of Chapter 3 for the details of REVERB challenge dataset.

4.3.1.2 VOiCES Dataset

The details of VOiCES dataset is given in Section 3.2.4.4 of Chapter 3.

4.3.2 E2E ASR baseline system

For all the ASR experiments, we use the weighted prediction error based pre-processing [16] and unsupervised generalized eigenvalue (GEV) beamforming [37]. The baseline features are 36-dimensional log-mel filter bank features with frequency range from 200 Hz to 6500 Hz. The ESPnet toolkit [118] is used to perform all the end-to-end ASR experiments, with a Pytorch backend [119]. The model architecture uses 12 conformer encoder layers with 2048 units in the projection layer. The 6-layer transformer architecture with 2048 units in the projection layer serves as the decoder. Both connectionist temporal cost (CTC) loss and attention based cross entropy (CE) loss are used in the training, with CTC-weight set at 0.3 [120]. A single layer of 1000 LSTM cell recurrent neural network is used for language modeling (RNN-LM). For training the model, we use stochastic gradient descent (SGD) optimizer with a batch size of 32. For language model training, the data is augmented from Wall Street Journal (WSJ) corpus.

4.3.3 Performance metrics

4.3.3.1 ASR performance metrics

- **WER/CER** (Word/Character Error Rate): The word/character error rate is given by the ratio of number of word/character insertions, deletions and substitutions in the system output to the total number of words/characters in the reference.

4.3.3.2 Speech quality metrics

- **SRMR**: Speech to reverberation modulation ratio (SRMR) is a non intrusive measure. Here, a representation is obtained using an auditory-inspired filter bank analysis of critical band temporal envelopes of the signal. The modulation spectral information is used to get

Table 4.1: *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**
BF-FBANK (baseline)	12.8	8.7	10.8	11.9	7.9	9.9
BF-FBANK- + env. derevb.	12.7	8.5	10.6	10.1	7.8	9
BF-FBANK- + crr. derevb.	11.2	8.3	9.8	10.8	7.6	9.2
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**

an adaptive measure termed as speech to reverberation modulation energy ratio [51, 52]. A higher value indicates an improved quality of the given speech signal.

- **MOS** (Mean Opinion Score): To evaluate the performance of dereverberation algorithms, subjective quality and intelligibility measurement methods are needed. The most widely used subjective method is the ITU-T standard [50], where a panel of listeners are asked to rate the quality/intelligibility of the audio.

4.4 Experiments and results

The baseline features for ASR are the beamformed log-mel filter-bank energy features (denoted as BF-FBANK).

4.4.1 REVERB Challenge ASR

The word error rates (WER) for the dereverberation experiments are shown in Table 4.1. Note that, all the experiments use the same input features (log-mel filter bank features) along with the same E2E ASR architecture (conformer encoder and transformer decoder). The only difference between the various rows, reported in Table 4.1, is the dereverberation pre-processing applied on the raw audio waveform. All the dereverberation experiments use the DPLSTM architecture described in Sec. 4.2.

4.4.1.1 Various dereverberation configurations

In Table 4.1, the first row is the baseline result with the beamformed audio (unsupervised GEV beamforming [37]) and weighted prediction error (WPE) processing [16]. The second row corresponds to the WER results with envelope based dereverberation alone. The relative improvements of $2 - 9\%$ are seen here compared to the baseline BF-FBANK. Separately, with dereverberation based on the carrier signal alone, a similar improvement is achieved. Further, the dereverberation of the temporal envelope and carrier components in a combined fashion using the DPLSTM model improves the ASR results over the separate dereverberation of envelope/carrier components. Here, average relative improvements of 16% and 19% are seen in the development set and evaluation set respectively, over the BF-FBANK baseline system for the DFAR approach.

The results using the joint learning of the dereverberation network and the E2E ASR model are reported as the last row of Table 4.1. The E2E-DFAR is initialized using the dereverberation model and the E2E model trained separately. The proposed E2E-DFAR model yields average relative improvements of 27% and 34% on the development set and evaluation set respectively over the baseline system. The joint training is also shown to improve over the set up of having separate networks for dereverberation and E2E ASR. While the DFAR model is trained only on simulated reverberation conditions, the WER improvement in real condition is seen to be more pronounced than those observed in the simulated data. This indicates that the model can generalize well to unseen reverberation conditions in the real-world.

4.4.1.2 Comparison with prior works

The comparison of the results from prior works reported on the REVERB challenge dataset is given in Table 4.2. The Table includes results from end-to-end ASR systems [121, 124, 123] as well as the joint enhancement and ASR modeling work reported in [122]. We also compare with our prior work reported in [25]. To the best of our knowledge, the results from the proposed E2E-

Table 4.2: *Comparison of the results with other works reported on the REVERB challenge dataset.*

System	Eval-sim.	Eval-real	Avg.
Subramanian et. al. [121]	6.6	10.6	8.6
Heymann et. al. [122]	-	10.8	-
Fujita et. al. [123]	**4.9**	9.8	7.4
Purushothaman et. al. [25]	7.1	12.1	9.6
Zhang et. al. [124]	-	10.0	-
This work	5.7	**7.3**	**6.5**

Table 4.3: *WER (%) in REVERB dataset for different architectures for the dereverberation model.*

Model Config.	**Dev**			**Eval**		
	Real	**Sim**	**Avg**	**Real**	**Sim**	**Avg**
Baseline	12.8	8.7	10.8	11.9	7.9	9.9
CLSTM	14.5	9.7	12.1	12.4	9.1	10.8
4-layer LSTM	12.5	8.0	10.3	10.1	7.1	8.9
DPLSTM	**10.6**	**7.6**	**9.1**	**9.1**	**6.9**	**8.0**

DFAR consitute the best published ASR performance on the REVERB challenge evaluation dataset (relative improvements of 12% over the recent work by Fujita et. al. [123]).

4.4.1.3 Dereverberation model architecture

The ASR experiments on the REVERB challenge dataset, pertaining to the choice of different model architectures used in the dereverberation model, are listed in Table 4.3. We have experimented with convolutional LSTM (CLSTM) and time-domain LSTM (4-layer LSTM) architecture in addition to the DPLSTM approach. As seen here, the Dual-path recurrence based DPLSTM gives the best word error rate in comparison with the other LSTM neural architectures considered. This may be attributed to the joint time-frequency recurrence compared to the other approaches which perform only time domain recurrence.

Table 4.4: *WER (%) in REVERB dataset for Hyper parameter λ, in $MSE\ \ loss = \lambda \times env.\ loss + (1-\lambda) \times carr.\ loss$.*

Parameter λ	Dev			Eval		
	Real	**Simu**	**Avg**	**Real**	**Simu**	**Avg**
0	12	8.2	10.1	10.4	7.5	9.0
0.2	11.9	8.6	10.3	10.7	7.7	9.2
0.4	11.6	8.2	9.9	10.1	7.2	8.7
0.5	11.3	**7.2**	9.3	9.7	**6.5**	8.1
0.6	**10.6**	7.6	**9.1**	**9.1**	6.9	**8.0**
0.8	13.1	8.7	10.9	10.9	7.9	9.4
1	13.5	8.0	10.8	10.4	6.9	8.7

4.4.1.4 Dereverberation loss function

The MSE loss function used in the DPLSTM model training consists of a combination of loss values from the envelope and the carrier components. During the training, it is possible to control the ratio of the two losses in the overall loss. We experimented with the hyper parameter, λ, which controls the proportion of envelope based loss and carrier based loss in the total loss ($Total\ loss = \lambda \times env.\ loss + (1-\lambda) \times carr.\ loss$). The ASR results for the various choices of the hyper parameter λ are shown in Table 4.4. Empirically, the value of $\lambda = 0.6$ gives the best WER on the REVERB challenge dataset. Further, the choice of $\lambda = 1/0$, corresponding to envelope/carrier only dereverberation, are inferior to other choices of λ, indicating that the joint dereverberation of both the envelope and carrier components is beneficial for far-field ASR.

4.4.2 VOiCES ASR

The ASR setup used in the VOiCES dataset followed the ESPnet recipe with the conformer encoder and a transformer decoder. The rest of the model parameters and hyper-parameters are kept similar to the ones in the REVERB challenge dataset. The WER results on the VOiCES dataset are given in Table 4.5. The dereverberation of the envelope alone provides an absolute improvement of 1.9% and 2.2% on the development and evaluation data respectively, compared

Table 4.5: *Performance (WER %) on the VOiCES dataset.*

Model Config.	**Dev**	**Eval**
FBANK (baseline)	40.3	50.8
+ Env. derevb.	38.4	48.6
+ Env.-carr. derevb. (DFAR)	37.1	45.4
+ E2E-DFAR	**36.4**	**44.7**

to the FBANK baseline system. The dereverberation based on envelope-carrier modeling further improves the results. An absolute improvement of 3.3% / 5.4% on the development / evaluation data is achieved, compared to the FBANK baseline. Further, the joint training on envelope-carrier dereverberation network with the ASR model improves the WER results. We observe relative improvements of 10% and 12% on the development and evaluation data respectively compared to the FBANK baseline.

4.4.3 Speech quality evaluation

A comparison of the SRMR values for different dereverberation approaches is reported in Table 4.6. Here, we compare the baseline unsupervised GEV beamforming [37] and weighted prediction error (WPE) [16] with various strategies for beamforming. The deep complex convolutional recurrent network (DCCRN) based speech enhancement [62] is also implemented on the REVERB dataset, and these results are reported in Table 4.6. While the envelope based dereverberation did not improve the SRMR values, the carrier based dereverberation is shown to improve the SRMR results. Further, the DFAR model also achieves similar improvements in SRMR for all the conditions over the baseline approach (GEV+WPE) and the DCCRN approach.

We have conducted a subjective evaluation to further assess the performance of the dereverberation method. The subjects were asked to rate the quality of the audio on a scale of 1 to 5, 1 being poor and 5 being excellent. The subjects listened to the audio in a relatively quiet room with a high quality Sennheiser headset. We perform the A-B listening test, where the two

Table 4.6: *SRMR values on the REVERB dataset for various signal enhancement strategies.*

Signal	SRMR				
	Dev. Real	Dev. Simu	Eval. Real	Eval. Simu	REVERB Tr_cut
Unsupervised GEV beamforming [37]	5.18	4.1	4.58	4.67	4.23
+ WPE [16]	5.35	4.2	4.61	4.75	4.48
+ DCCRN[62]	5.43	4.37	4.63	4.94	4.67
+ env. derevb. (this work)	4.62	3.83	4.12	4.25	4.11
+ crr. derevb. (this work)	5.52	4.46	4.69	5.27	4.77
+ env. & crr. derevb. [DFAR] (this work)	**5.52**	**4.47**	**4.69**	**5.27**	**4.77**

Table 4.7: *MOS values in REVERB dataset for envelope and carrier based enhancements. The experiments used* 20 *audio samples and recruited* 20 *human listeners.*

	ET Real - near	**ET Real - far**	**ET Simu - near**	**ET Simu - far**
Baseline - GEV [37] + WPE [16]	3.78	3.65	3.74	4.12
+ env.-carr. derevb. [DFAR] (this work)	**3.98**	**3.67**	**4.01**	**4.40**

versions of the same audio file were played, the first one with GEV + WPE dereverberation and the second one with the proposed dereverberation approach. We chose 20 audio samples, from four different conditions (real and simulated data and from near and far rooms) for this evaluation and recruited 20 subjects.

The subjective results are shown in Table 4.7. As seen, the proposed speech dereverberation scheme shows improvement in subjective MOS scores for all the conditions considered. The subjective results validate the signal quality improvements observed in the SRMR values (Table 4.6).

4.5 Chapter summary

In this chapter, we propose a speech dereverberation model using frequency domain linear prediction based sub-band envelope-carrier decomposition. The sub-band envelope and carrier components are processed through a dereverberation network. A novel neural architecture, based on dual path recurrence, is proposed for dereverberation. Using the joint learning of the neural speech dereverberation module and the E2E ASR model, we perform several speech recognition experiments on the REVERB challenge dataset as well as on the VOiCES dataset. These results show that the proposed approach improves over the state of art E2E ASR systems based on mel filterbank features.

The dereverberation approach proposed in this chapter also reconstructs the audio signal, which makes it useful for audio quality improvement applications as well as other speech processing systems in addition to the ASR system. We have further evaluated the reconstruction quality subjectively and objectively on the REVERB challenge dataset. The quality measurements show that the proposed speech dereverberation method improves speech quality over the baseline framework of weighted prediction error. The ablation studies on various architecture choices provides justification for the choice of the DPLSTM network architecture. Given that the proposed model allows the reconstruction of the audio signal, it can be used in conjunction with self-supervised neural approaches for representation learning of speech as well. This will form part of our future investigation, where a joint modeling framework will be considered which will involve dereverberation, representation learning and ASR modeling.

Chapter 5

Summary

5.1 Chapter conclusions

In this thesis, we have proposed several research directions for dereverberation and enhancement of far-field speech using an autoregressive model of sub-band envelopes.

In Chapter 1, we have described the problem of reverberation, its mathematical model and the impact on applications. Also, speech enhancement techniques specifically designed for far-field automatic speech recognition, are detailed. This is followed by the list of contributions from this thesis, where the novel proposals specific to the thesis are discussed. Finally, an organization of the thesis is given.

The background materials required to understand the thesis are given in Chapter 2. Theory of frequency domain linear prediction is analyzed in detail in the first section. This is followed by a discussion of far-field speech enhancement based on beamforming. The description of two broad automatic speech recognition paradigms is given in the next section. The details about the major performance metrics used in speech enhancement and ASR are discussed then. Finally, a survey on prior works in spectral enhancement is highlighted.

Chapter 3 describes feature level dereverberation with FDLP. A joint acoustic model named 3-D SSAR model, which considers three dimensions of time, frequency and channel is proposed in the first part. The proposed 3-D feature and acoustic modeling approach provides significant improvements over an ASR system trained with beamformed audio (average relative improvements of 16% and 6% in word error rates for CHiME-3 and REVERB Challenge datasets respectively).

In the second part of Chapter 3, we propose a new neural model for dereverberation of temporal envelopes and joint learning of the acoustic model to improve the ASR cost. The joint learning framework combines the envelope dereverberation framework, feature pre-processing and acoustic modeling into a single neural pipeline. Experimental results show improvements in the REVERB Challenge, CHiME-3 and VOiCES datasets.

In Chapter 4, we develop a model for dereverberation of audio signal using an envelope-carrier decomposition. The main focus of this chapter is to resynthesize the audio signal back without the reverberation artifacts. We investigate the effect of reverberation in temporal envelopes and corresponding carrier signals. Here, we show that the reverberant carrier can be approximated as the sum of clean carrier and late reflection carrier. The dual-path long short time model, named DPLSTM, is used to parallelly learn the mapping between the reverberant envelopes/carrier and their clean counterparts. Further, joint learning of the speech dereverberation model with the end-to-end ASR model is proposed.

The average relative improvements of 20% on the real development set and about 18% on the real evaluation set, compared to the BF-FBANK baseline is achieved by the joint neural dereverberation method proposed in second part of Chapter 3, on REVERB challenge dataset. The proposed E2E-DFAR model (Chapter 4) yields average relative improvements of 27% and 34% on the development set and evaluation set respectively over the baseline system, in REVERB challenge dataset.

In general, in many of the experiments, the dereverberation applied to FDLP envelopes is more substantial than when it was applied to the Mel-spectrogram features. We hypothesize that this is due to the smoothed representations given by FDLP, which preserves high energy regions quite well. The smooth representations avoids fine details which are more speaker specific.

5.2 Limitations

The key limitations from the thesis can be summarized as below,

5.2.1 Modeling envelopes using long-term windows

All the research work reported in this thesis perform dereverberation using long analysis windows applied on speech sub-bands. Further, the FDLP based processing is computationally intensive, where the linear prediction is applied on the long-sequence of sub-band DCT components. There are additional hyper-parameters like the window-duration, FDLP model order and type of spectral transform used in FDLP (DCT versus DFT). Thus, while the performance benefits have been highlighted through various ASR and speech enhancement experiments, more work is required to optimize the processing pipe-line for online and real-time applications.

5.2.2 Dereverberation for non-ASR applications

The current work had explored dereverberation for speech enhancement and ASR applications only. However, several other speech applications like speaker recognition, language recognition, emotion recognition, key-word spotting, and speech activity detection may also benefit from the dereveberation strategies developed in this thesis. Many of these applications have shifted to a fully neural processing pipeline, thus also allowing the dereverberation to learnt jointly for these applications as well. Since these applications have not been experimented in this thesis, it is unclear if the current approach will also benefit these tasks.

5.2.3 Dereverberation and Self-supervised Learning

The field of self-supervised representations learning has made wide-spread impact in the design of speech systems. The representation learning frameworks like wav2vec have shown promising results for speech recognition and other speech applications. This thesis has not explored the application of the dereverberation models along with these representations. But, if we are reconstructing the waveform, we may be able to perform self-supervision based representations on the dereverberated audio.

5.3 Future work

Most of the dereverberation approaches were trained with relatively small amounts of audio ($<$ 100h). The future efforts could also explore larger speech dereverberation data settings using the conformer style architectures.

We have reported the improvements in WER for ASR due to envelope based dereverberation on different datasets, in Chapter 3. The neural model learned on simulated data generalized well on real data. The applicability of the proposed method on non-ASR tasks like speaker recognition, language identification and emotion recognition should be explored in future.

The dereverberation approach proposed in Chapter 4 also reconstructs the audio signal, which makes it useful for audio quality improvement applications as well as other speech processing systems in addition to the ASR system. Given that the proposed model allows the reconstruction of the audio signal, it can be used in conjunction with self-supervised neural approaches for representation learning of speech as well. This will form part of our future investigation, where a joint modeling framework will be considered which will involve dereverberation, representation learning and ASR modeling.

Appendix A: Frequency domain linear prediction (FDLP)

5.A Conventional Signal Analysis

Conventionally, signal analysis for speech/audio signals is done by windowing the signal into short-term frames (typically of the order of 20-30ms) followed by an estimation of spectrum within each frame. A sequence of these short-term frames contain the signal information which are processed by subsequent stages. For speech signals, most of the information captured by such an analysis relates to formant structure of speech.

5.A Linear prediction based spectrogram - LP spectrogram

Linear prediction can be used to reveal the spectral content of a signal [125]. A smooth spectral estimate, where the peaks are preserved and finer details are left out, is obtained 5.1. By fitting an auto-regressive model on the 20-30ms frame and finding the model parameter using linear prediction in time (termed TDLP, for time domain liner prediction), we can find TDLP spectrogram as shown in figure 5.2.

However, speech/audio signals have information spread across longer temporal context of the order of 200ms or more. For example, even a basic speech unit like a phoneme lasts

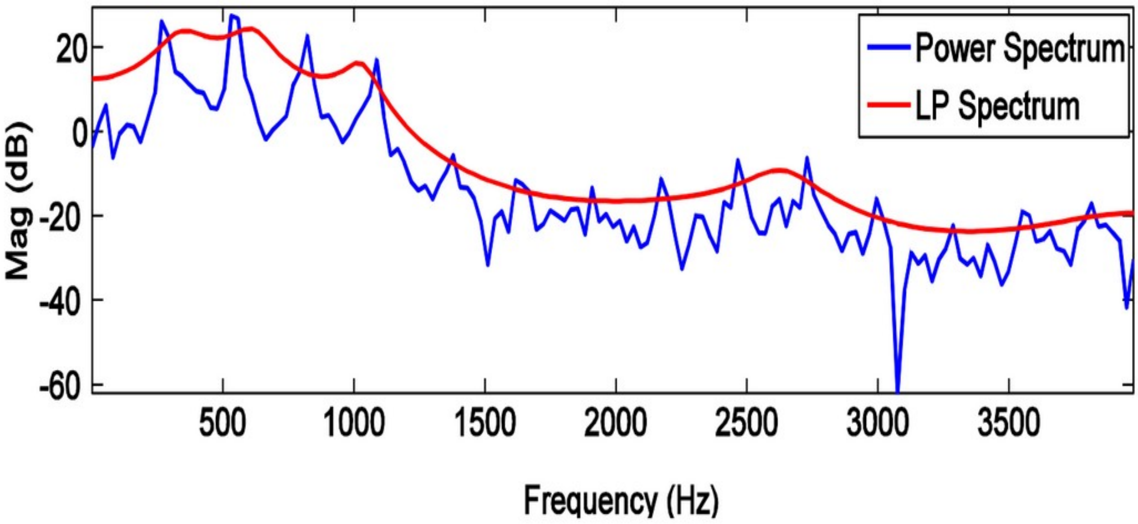


Figure 5.1: *By doing Linear prediction in time, we get a smooth spectral representation of the signal , Figure Credit:[1]*

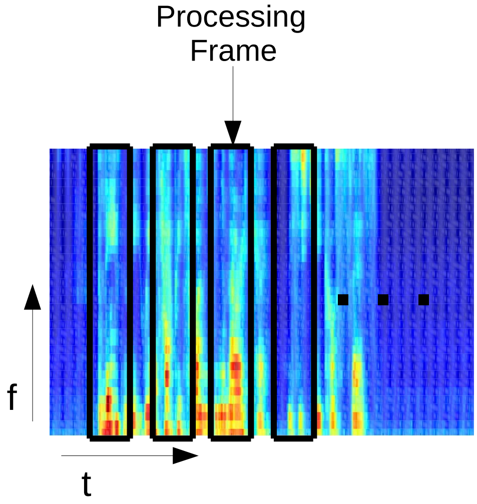


Figure 5.2: *TDLP spectrogram - A smooth spectral representation, at the same time , preserving the peaks.*

for 70-80ms. The choice of 20-30ms in a short-term spectrum approach may be ad-hoc as it does not address the time-frequency compromise inherent in any signal analysis. Specifically, conventional approaches sample the spectrum at a preset rate before the application of any further processing.

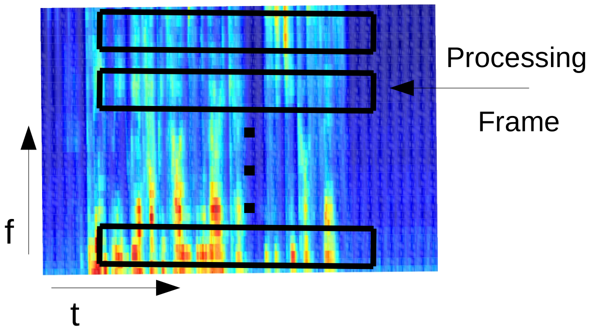


Figure 5.3: *Just as TDLP estimates the spectral envelope of a signal, FDLP estimates the temporal envelope of the signal.*

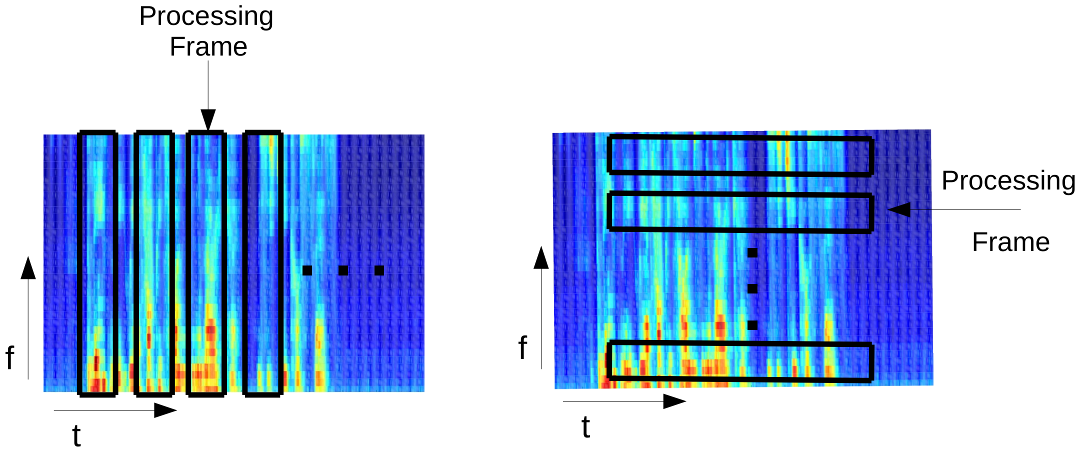


Figure 5.4: *Overview of time-frequency energy representation for TDLP and FDLP.*

An alternate way to construct the same 2-D representation is to process a single frequency band over a long duration and stack these bands in a row-wise manner. This is shown in Fig. 5.3. This is termed frequency domain linear prediction or FDLP in short. Linear prediction in the spectral domain was first proposed by Kumaresan et. al [126].

Fig. 5.4 shows a comparison of TDLP and FDLP based spectral estimation. 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.

Bibliography

[1] Sriram Ganapathy. *Signal analysis using autoregressive models of amplitude modulation*. PhD thesis, Johns Hopkins University, 2012. xviii, 86

[2] Matthias Wölfel and John McDonough. *Distant speech recognition*. John Wiley & Sons, 2009. 1

[3] James A Moorer. About this reverberation business. *Computer music journal*, pages 13–28, 1979. 2

[4] Thomas Hain, Lukáš Burget, John Dines, Philip N Garner, František Grézl, Asmaa El Hannani, Marijn Huijbregts, Martin Karafiat, Mike Lincoln, and Vincent Wan. Transcribing meetings with the amida systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):486–498, 2012. 2, 4

[5] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur. Low latency acoustic modeling using temporal convolution and LSTMs. *IEEE Signal Processing Letters*, 25(3):373–377, 2018. 2

[6] Sriram Ganapathy and Vijayaditya Peddinti. 3-D cnn models for far-field multi-channel speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5499–5503. IEEE, 2018. 2, 28

[7] Aleksei Gusev, Vladimir Volokhov, Tseren Andzhukaev, Sergey Novoselov, Galina Lavren-

tyeva, Marina Volkova, Alice Gazizullina, Andrey Shulipa, Artem Gorlanov, Anastasia Avdeeva, et al. Deep speaker embeddings for far-field speaker recognition on short utterances. *arXiv preprint arXiv:2002.06033*, 2020. 2

[8] Qin Jin, Tanja Schultz, and Alex Waibel. Far-field speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2023–2032, 2007. 2

[9] Takuya Yoshioka et al. Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition. *IEEE Signal Processing Magazine*, 29(6):114–126, 2012. 2

[10] Vijayaditya Peddinti, Yiming Wang, Daniel Povey, and Sanjeev Khudanpur. Low latency acoustic modeling using temporal convolution and LSTMs. *IEEE Signal Processing Letters*, 2017. 4, 64

[11] Ladislav Mošner, Pavel Matějka, Ondřej Novotnỳ, and Jan Honza Černockỳ. Dereverberation and beamforming in far-field speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5254–5258. IEEE, 2018. 4

[12] Qin Jin, Tanja Schultz, and Alex Waibel. Far-field speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2023–2032, 2007. 4

[13] M. Nandwana, J. Van Hout, M. McLaren, C. Richey, A. Lawson, and M. Barrios. The voices from a distance challenge 2019 evaluation plan. *arXiv preprint arXiv:1902.10828*, 2019. 4, 9, 44, 58, 65

[14] C. Richey, M. Barrios, et al. VOiCES obscured in complex environmental settings (voices) corpus. *arXiv preprint arXiv:1804.05053*, 2018. 4, 9, 44, 58, 65

[15] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang. Blind speech dereverberation with multi-channel linear prediction based

on short time fourier transform representation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 85–88. IEEE, 2008. 4, 5, 13, 14

[16] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang. Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE TASLP*, 18(7):1717–1731, 2010. 4, 5, 13, 14, 54, 64, 73, 75, 78, 79

[17] Xavier Anguera, Chuck Wooters, and Javier Hernando. Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2011–2022, 2007. 5, 16, 35, 64

[18] Zhong-Qiu Wang and DeLiang Wang. A joint training framework for robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):796–806, 2016. 5, 6, 29, 45

[19] Bo Wu, Kehuang Li, Zhen Huang, Sabato Marco Siniscalchi, Minglei Yang, and Chin-Hui Lee. A unified deep modeling approach to simultaneous speech dereverberation and recognition for the reverb challenge. In *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, pages 36–40. IEEE, 2017. 5, 6, 29, 31, 45

[20] B. Wu, K. Li, F. Ge, Z. Huang, M. Yang, S. M. Siniscalchi, and C. Lee. An end-to-end deep learning approach to simultaneous speech dereverberation and acoustic modeling for robust speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1289–1300, 2017. 5, 6, 31, 45, 55, 56

[21] Ernst Warsitz and Reinhold Haeb-Umbach. Blind acoustic beamforming based on generalized eigenvalue decomposition. *IEEE Transactions on audio, speech, and language processing*, 15(5):1529–1539, 2007. 5, 17, 64

[22] Takuya Higuchi, Nobutaka Ito, Takuya Yoshioka, and Tomohiro Nakatani. Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise. In *2016*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5210–5214. IEEE, 2016. 6

[23] 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)*, pages 6964–6968, 2020. 8

[24] Anurenjan Purushothaman, Anirudh Sreeram, Rohit Kumar, and Sriram Ganapathy. Deep Learning Based Dereverberation of Temporal Envelopes for Robust Speech Recognition. In *Proc. Interspeech 2020*, pages 1688–1692, 2020. 8

[25] Anurenjan Purushothaman, Anirudh Sreeram, Rohit Kumar, and Sriram Ganapathy. Dereverberation of autoregressive envelopes for far-field speech recognition. *Computer Speech & Language*, 72:101277, 2022. 8, 9, 65, 75, 76

[26] 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)*, pages 6964–6968. IEEE, 2020. 9, 65

[27] Keisuke Kinoshita et al. The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In *IEEE WASPAA*, pages 1–4, 2013. 9, 44, 54, 65

[28] Lawrence Marple. Computing the discrete-time" analytic" signal via fft. *IEEE Transactions on signal processing*, 47(9):2600–2603, 1999. 11

[29] Stephen A Martucci. Symmetric convolution and the discrete sine and cosine transforms. *IEEE Transactions on Signal Processing*, 42(5):1038–1051, 1994. 11

[30] Edward Bedrosian. The analytic signal representation of modulated waveforms. *Proceedings of the IRE*, 50(10):2071–2076, 1962. 11

[31] Barry D Van Veen and Kevin M Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP magazine*, 5(2):4–24, 1988. 16

[32] Hamid Krim and Mats Viberg. Two decades of array signal processing research. *IEEE Signal Processing magazine*, 1996. 16

[33] Charles Knapp and Glifford Carter. The generalized correlation method for estimation of time delay. *IEEE transactions on acoustics, speech, and signal processing*, 24(4):320–327, 1976. 16

[34] Michael S Brandstein and Harvey F Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 375–378. IEEE, 1997. 16

[35] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach. Neural network based spectral mask estimation for acoustic beamforming. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 196–200, 2016. 18

[36] Tomohiro Nakatani, Nobutaka Ito, Takuya Higuchi, Shoko Araki, and Keisuke Kinoshita. Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 286–290. IEEE, 2017. 18

[37] R. Kumar, A. Sreeram, A. Purushothaman, and S. Ganapathy. Unsupervised neural mask estimator for generalized eigen-value beamforming based ASR. In *IEEE ICASSP*, pages 7494–7498, 2020. 19, 54, 64, 73, 75, 78, 79

[38] Jean Dreyfus-Graf. Sonograph and sound mechanics. *The Journal of the Acoustical Society of America*, 22(6):731–739, 1950. 20

[39] W Chengyou, L Diannong, K Tiesheng, C Huihuang, and T Chaojing. Automatic speech recognition technology review. acoust. *Electr. Eng*, 49:15–21, 1996. 20

[40] Kai-Fu Lee. On large-vocabulary speaker-independent continuous speech recognition. *Speech communication*, 7(4):375–379, 1988. 20

[41] Herve Bourlard and Nelson Morgan. Continuous speech recognition by connectionist statistical methods. *IEEE Transactions on Neural Networks*, 4(6):893–909, 1993. 20

[42] George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42, 2011. 20

[43] James H Martin. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall, 2009. 21

[44] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772, 2014. 22

[45] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4835–4839. IEEE, 2017. 22

[46] Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012. 22

[47] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013. 22

[48] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 22

[49] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253, 2017. 23

[50] PREPUBLISHED RECOMMENDATION. Itu-tp. 808. 2018. 23, 74

[51] Tiago H Falk, Chenxi Zheng, and Wai-Yip Chan. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1766–1774, 2010. 23, 74

[52] João F. Santos, Mohammed Senoussaoui, and Tiago H. Falk. An improved non-intrusive intelligibility metric for noisy and reverberant speech. In *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 55–59, 2014. 23, 74

[53] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE, 2001. 24

[54] P Recommendation. 862: Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Feb*, 14:14–0, 2001. 24

[55] Antony Rix, Richard Reynolds, and Mike Hollier. Perceptual measurement of end-to-end speech quality over audio and packet-based networks. In *Audio Engineering Society Convention 106*. Audio Engineering Society, 1999. 24

[56] Antony W Rix and Michael P Hollier. The perceptual analysis measurement system for robust end-to-end speech quality assessment. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pages 1515–1518. IEEE, 2000. 24

[57] John G Beerends and Jan A Stemerdink. A perceptual speech-quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, 42(3):115–123, 1994. 24

[58] Gabriel Mittag and Sebastian Möller. Non-intrusive speech quality assessment for super-wideband speech communication networks. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7125–7129, 2019. 24

[59] Mark R Weiss, Ernest Aschkenasy, and Thomas W Parsons. Study and development of the intel technique for improving speech intelligibility. Technical report, NICOLET SCIENTIFIC CORP NORTHVALE NJ, 1975. 25

[60] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120, 1979. 25

[61] Norbert Wiener, Norbert Wiener, Cyberneticist Mathematician, Norbert Wiener, Norbert Wiener, and Cybernéticien Mathématicien. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*, volume 113. MIT press Cambridge, MA, 1949. 26

[62] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv preprint arXiv:2008.00264*, 2020. 28, 78, 79

[63] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):7–19, 2014. 28, 44

[64] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018. 28, 44

[65] Yan Zhao, DeLiang Wang, Buye Xu, and Tao Zhang. Late reverberation suppression using recurrent neural networks with long short-term memory. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5434–5438, 2018. 28, 44

[66] Kun Han, Yuxuan Wang, and DeLiang Wang. Learning spectral mapping for speech dereverberation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4628–4632, 2014. 28, 44

[67] Donald S Williamson and DeLiang Wang. Time-frequency masking in the complex domain for speech dereverberation and denoising. *IEEE/ACM transactions on Audio, Speech, and Language processing*, 25(7):1492–1501, 2017. 28, 45, 64

[68] A. Pandey and D. Wang. A new framework for CNN-based speech enhancement in the time domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(7):1179–1188, 2019. 28, 45

[69] Martin Wöllmer, Zixing Zhang, Felix Weninger, Björn Schuller, and Gerhard Rigoll. Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6822–6826, 2013. 28, 45

[70] Zhuo Chen, Shinji Watanabe, Hakan Erdogan, and John R Hershey. Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015. 28, 45

[71] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 91–99. Springer, 2015. 28, 45

[72] Andrew L Maas, Tyler M O'Neil, Awni Y Hannun, and Andrew Y Ng. Recurrent neural network feature enhancement: The 2nd chime challenge. In *Proceedings The 2nd CHiME Workshop on Machine Listening in Multisource Environments held in conjunction with ICASSP*, pages 79–80, 2013. 28, 45

[73] Joao Felipe Santos and Tiago H Falk. Speech dereverberation with context-aware recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(7):1236–1246, 2018. 28, 45, 55, 56, 57

[74] Lloyd Griffiths and CW Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on antennas and propagation*, 30(1):27–34, 1982. 28

[75] Sharon Gannot, David Burshtein, and Ehud Weinstein. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, 49(8):1614–1626, 2001. 28

[76] Guanjun Li, Shan Liang, Shuai Nie, Wenju Liu, and Zhanlei Yang. Deep neural network-based generalized sidelobe canceller for dual-channel far-field speech recognition. *Neural Networks*, 141:225–237, 2021. 28

[77] Feng-Ju Chang, Martin Radfar, Athanasios Mouchtaris, Brian King, and Siegfried Kunzmann. End-to-end multi-channel transformer for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888. IEEE, 2021. 28

[78] Chanwoo Kim, Abhinav Garg, Dhananjaya Gowda, Seongkyu Mun, and Changwoo Han. Streaming end-to-end speech recognition with jointly trained neural feature enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6773–6777. IEEE, 2021. 28

[79] Kristina Tesch and Timo Gerkmann. Insights into deep non-linear filters for improved multi-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:563–575, 2022. 31

[80] Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li. Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6633–6637. IEEE, 2021. 31, 32

[81] Rui Zhou, Wenye Zhu, and Xiaofei Li. Speech dereverberation with a reverberation time shortening target. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 31

[82] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals. Hybrid acoustic models for distant and multichannel large vocabulary speech recognition. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 285–290. IEEE, 2013. 34

[83] Ritwik Giri, Michael L Seltzer, Jasha Droppo, and Dong Yu. Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning. In *2015*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5014–5018. IEEE, 2015. 34

[84] Matthias Wölfel and John McDonough. *Distant speech recognition*. John Wiley & Sons, 2009. 34

[85] Marc Delcroix et al. Strategies for distant speech recognition in reverberant environments. *EURASIP Journal on Advances in Signal Processing*, 2015(1):60, 2015. 34

[86] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. The fifth'chime'speech separation and recognition challenge: Dataset, task and baselines. *arXiv preprint arXiv:1803.10609*, 2018. 34

[87] Michael L Seltzer, Bhiksha Raj, and Richard M Stern. Likelihood-maximizing beamforming for robust hands-free speech recognition. *IEEE Transactions on speech and audio processing*, 12(5):489–498, 2004. 35

[88] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals. Convolutional neural networks for distant speech recognition. *IEEE Signal Processing Letters*, 21(9):1120–1124, 2014. 35

[89] Xiong Xiao et al. Deep beamforming networks for multi-channel speech recognition. In *ICASSP*, pages 5745–5749. IEEE, 2016. 35

[90] Tara N Sainath et al. Multichannel signal processing with deep neural networks for automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(5):965–979, 2017. 35

[91] Tsubasa Ochiai, Shinji Watanabe, Takaaki Hori, John R Hershey, and Xiong Xiao. A unified architecture for multichannel end-to-end speech recognition with neural beamforming. *IEEE Journal of Selected Topics in Signal Processing*, 2017. 35

[92] Sriram Ganapathy and Vijayaditya Peddinti. 3-d cnn models for far-field multi-channel speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5499–5503. IEEE, 2018. 35

[93] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach. A generic neural acoustic beamforming architecture for robust multi-channel speech processing. *Computer Speech & Language*, 46:374–385, 2017. 35, 39

[94] PP Vaidyanathan. The theory of linear prediction. *Synthesis lectures on signal processing*, 2(1):1–184, 2007. 37

[95] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 38

[96] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010. 39

[97] Samuel Thomas, Sriram Ganapathy, and Hynek Hermansky. Recognition of reverberant speech using frequency domain linear prediction. *IEEE Signal Processing Letters*, 15:681–684, 2008. 44, 47

[98] Sriram Ganapathy and Madhumita Harish. Far-field speech recognition using multivariate autoregressive models. In *Interspeech*, pages 3023–3027, 2018. 44

[99] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. The third 'chime'speech separation and recognition challenge: Dataset, task and baselines. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 504–511. IEEE, 2015. 44, 54

[100] Anurenjan Purushothaman, Anirudh Sreeram, and Sriram Ganapathy. 3-D acoustic modeling for far-field multi-channel speech recognition. In *IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), pages 6964–6968, 2020. 46, 52, 54, 55, 59

[101] Sriram Ganapathy. Multivariate autoregressive spectrogram modeling for noisy speech recognition. *IEEE signal processing letters*, 24(9):1373–1377, 2017. 46

[102] Sriram Ganapathy, Petr Motlicek, and Hynek Hermansky. Autoregressive models of amplitude modulations in audio compression. *IEEE transactions on audio, speech, and language processing*, 18(6):1624–1631, 2009. 46, 68

[103] Rainer Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE transactions on speech and audio processing*, 13(5):845–856, 2005. 49

[104] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 51

[105] Sriram Ganapathy. *Signal analysis using autoregressive models of amplitude modulation.* PhD thesis, Johns Hopkins University, 2012. 51, 67

[106] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang. Learning spectral mapping for speech dereverberation and denoising. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(6):982–992, 2015. 55, 56, 57

[107] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011. 55

[108] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, volume 270, pages 1–11, 2000. 55

[109] Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992. 55

[110] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, pages 5206–5210. IEEE, 2015. 58

[111] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 59

[112] M. L. Seltzer, D. Yu, and Y. Wang. An investigation of deep neural networks for noise robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7398–7402, 2013. 64

[113] Parishwad P Vaidyanathan. *Multirate systems and filter banks*. Pearson Education India, 2006. 67

[114] Petr Motlicek, Sriram Ganapathy, Hynek Hermansky, and Harinath Garudadri. Scalable wide-band audio codec based on frequency domain linear prediction. Technical report, IDIAP, 2007. 67

[115] Marios Athineos and Daniel PW Ellis. Frequency-domain linear prediction for temporal features. 2003. 67

[116] Sriram Ganapathy, Sri Harish Mallidi, and Hynek Hermansky. Robust feature extraction using modulation filtering of autoregressive models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(8):1285–1295, 2014. 68

[117] Yi Luo, Zhuo Chen, and Takuya Yoshioka. Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 46–50, 2020. 70

[118] S. Watanabe, T. Hori, et al. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*, 2018. 73

[119] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 73

[120] S. Karita, N. Chen, et al. A comparative study on transformer vs RNN in speech applications. In *2019 IEEE ASRU*, pages 449–456. IEEE, 2019. 73

[121] A. Subramanian, X. Wang, S. Watanabe, T. Taniguchi, D. Tran, and Y. Fujita. An investigation of end-to-end multichannel speech recognition for reverberant and mismatch conditions. *arXiv preprint arXiv:1904.09049*, 2019. 75, 76

[122] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani. Joint optimization of neural network-based wpe dereverberation and acoustic model for robust online ASR. In *ICASSP*, pages 6655–6659. IEEE, 2019. 75, 76

[123] Y. Fujita, A. Subramanian, M. Omachi, and S. Watanabe. Attention-based asr with lightweight and dynamic convolutions. In *ICASSP*, pages 7034–7038. IEEE, 2020. 75, 76

[124] W. Zhang, A. Subramanian, X. Chang, S. Watanabe, and Y. Qian. End-to-end far-field speech recognition with unified dereverberation and beamforming. *arXiv preprint arXiv:2005.10479*, 2020. 75, 76

[125] John G Proakis. *Digital signal processing: principles, algorithms, and applications, 4/E.* Pearson Education India, 2007. 85

[126] Ramdas Kumaresan and Ashwin Rao. Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications. *The Journal of the Acoustical Society of America*, 105(3):1912–1924, 1999. 87