Mask Estimator Approaches For Audio Beamforming

A PROJECT REPORT
SUBMITTED IN PARTIAL FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

Master of Technology

IN

SIGNAL PROCESSING

BY

Rohit Kumar

Under the guidance of

Dr. Sriram Ganapathy

Electrical Communication/Electrical Engineering
Indian Institute of Science
Bangalore – 560 012 (India)

August, 2021

Declaration of Originality

I, **Rohit Kumar**, with SR No. **04-02-02-10-42-18-1-16100** hereby declare that the material presented in the thesis titled

Mask Estimator Approaches For Audio Beamforming

represents original work carried out by me in the **Department of Electrical Engineering** at **Indian Institute of Science** during the years **2018-2020**.
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: 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: Dr. Sriram Ganapathy Advisor Signature

DEDICATED TO

My parents.

Acknowledgements

I would like to express my sincere and deepest gratitude to Dr. Sriram Ganapathy for providing me the opportunity to work under his guidance. His expertise, invaluable guidance, constant encouragement, patience, and healthy criticism added considerably to my experience. I gratefully acknowledge his invaluable suggestions towards improvements of my technical writing skill. I am highly thankful to Anurenjan P. R. (PhD student of LEAP lab), Anirudh Sreeram (Project Assistant) and Venkat K(Project Assistant), for helping me out to understand various concepts especially regarding ASR and Networks. Also I would like to express my sincere gratitude and want to say thankyou to my senior R.Manoj Kumar(MTech SP), Sagar Govindrao Ingale and my friends Gururaj K, M V Praveen, Ranga Shaarad Ayyagari and Kanishak Vaidya who have always guided me,motivated me and helped me, throughout my journey in IISC. In the end I extend my thanks to all my lab mates for their kind support. I am also grateful to all my professors for imparting immense knowledge during the course work. I sincerely acknowledge Chandrika Thunga for providing help related to administration.

I must express my gratitude to my parents and friends for their constant support during the entire course work. It would have been impossible to accomplish master degree without their encouragement.

Abstract

Beamforming is a family of algorithms and performs a spatial filtering operation that makes it possible to map the distribution of the sources at a certain distance from the microphones and therefore locate the strongest source. The state-of-art methods for acoustic beamforming in multi-channel ASR are based on a neural mask estimator that predicts the presence of speech and noise, which in turn used to determine spatial filter coefficients value. These models are trained using a paired corpus of clean and noisy recordings (teacher model). In this thesis, we attempt to move away from the requirements of having supervised clean recordings for training the mask estimator. The models based on signal enhancement and beamforming using multi-channel linear prediction serve as the required mask estimate. In this way, the model training can also be carried out on real recordings of noisy speech rather than simulated ones alone done in a typical teacher model. We propose two model in this thesis, both based on Unsupervised Mask estimation, and several experiments performed on noisy and reverberant environments in the CHiME-3 corpus as well as the REVERB challenge corpus highlight the effectiveness of the proposed approaches. Both the method that we discuss are novel method, where the first model only deals with the real data, the second model deals with complex data i,e complex short time Fourier transform features to obtain the mask estimate.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Acoustic Beamforming

Acoustic Beamforming as best described in the paper [1] is a signal processing technique based on far-field microphone array measurements. Beamforming is a family of algorithms and performs a spatial filtering operation that makes it possible to map the distribution of the sources at a certain distance from the microphones and therefore locate the strongest source. The beamforming approaches in general can be divided into two categories one is **Time domain algorithms**, which perform beamforming as a delayed and weighted summation of the multiple spatially separated microphones to provide an enhanced audio signal.

$$L(t, x_0) = \frac{4\pi}{M} \sum_{m=1}^{M} p_m(x_0, t + t_0)|x - x_0| \tag{1.1}$$

where p_m is the signal measured by each microphone,M is the number of microphone,x_0 is the source position and x is the microphone location.

The second is **Frequency domain algorithms**, where the beamforming operation in frequency domain determines the spatial filter coefficients $w(m, k)$ to obtain the enhanced signal,

$$z(k, n) = \sum_{m=0}^{M-1} w(m, k)\ y^m(k, n) \tag{1.2}$$

where $z(k, n)$ is the beamformed signal. In this thesis we explore the later family of algorithm in which there is an approach to beamforming using a generalized eigen value(GEV) formulation [2]. This 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 [2] 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) [3].

1.2 Application of Beamforming

The most common application of beamforming is in automatic speech recognition (ASR). The automatic speech recognition (ASR) in noisy/reverberant multi-channel environments continue to be a challenging task. The improvement of ASR solutions in such environments are key to several applications like smart speakers, home automation and in meeting transcription systems. Another application of beamforming can be to localise the target speaker in the presence of multiple speaker in the **cocktail party** problem.

1.3 Outline of Contributions

In this thesis first we will propose a novel combination of multi-channel Linear Prediction (MCLP) based beamforming method [4] with mask estimation based GEV beamforming for addressing the problem of unsupervised mask estimation and beamforming. The MCLP based algorithm generates a "clean" version of the audio that is bootstrapped to a DNN mask estimation process. Using this simple approach, we show that the model can also be effectively trained on real recordings where there are no parallel clean recordings. With several ASR experiments on CHiME-3 and REVERB challenge dataset, we show that the proposed approach performs on par with the oracle mask estimation methods. In addition, the approach significantly improves over a DNN mask estimator trained on an out-of-domain supervised dataset.Also in this thesis we will propose a complex architecture to estimate the speech and noise presence probability. We also explore Gaussian-weighted Self-Attention, which is a type of relative attention used in speech signal. The idea behind using relative attention is that in speech signals the current frame is highly correlated with the context frames compare to far away frames, and therefore attention should be based on neighboring frames.

1.4 Road Map for Rest of the Thesis

The rest of the thesis is organised as follows. In Chapter 2, we discuss some of the background required to understand this thesis. First we will discuss the signal model that will be using uniformly throughout the thesis. The signal model will represent the multi channel audio in a reverberant condition. Then, in that chapter we talk about Weighted Prediction Error(WPE) algorithm which is used to remove the convoluted reverberant component from the speech signal We discuss about the commonly used acoustic beamforming approaches namely BeamformIt,Generalized Eigen Value (GEV) Beamforming and Minimum Variance Distortionless

Response(MVDR) Beamforming.

In the Chapter 3 we discuss the Unsupervised Mask Estimation Process. In that chapter first we will going to talk about the state of the art Supervised DNN based Neural Mask Estimator. The drawback with the DNN based mask estimator is that it is trained using a pair of clean and multi channel noisy recordings and the DNN learns the SPP with binary targets. However the requirement of parallel speech recordings in clean and multi-channel reverberant conditions is a key limitation to these neural mask estimation methods.Recently, unsupervised approaches to mask estimation using complex mixture Gaussian model have been attempted [5, 6]. However, they are either computationally expensive or suffer from a degradation in performance compared to the DNN based mask estimation using oracle targets. In this section, we propose a combination of multi-channel Linear Prediction (MCLP) based beamforming method [4] with mask estimation based GEV beamforming for addressing the problem of unsupervised mask estimation and beamforming. The MCLP based algorithm generates a "clean" version of the audio that is bootstrapped to a DNN mask estimation process. Using this simple approach, we show that the model can also be effectively trained on real recordings where there are no parallel clean recordings. With several ASR experiments on CHiME-3 and REVERB challenge dataset, we show that the proposed approach performs on par with the oracle mask estimation methods. In addition, the approach significantly improves over a DNN mask estimator trained on an out-of-domain supervised dataset.

In the Chapter 4, we talk about complex deep learning architecture to estimate the unsupervised mask estimation. Deep Learning has seen a huge interest and change in past decade, however major deep learning models rarely use complex numbers. This problem is especially necessary for speech community because the audio data that we handle is naturally complex valued, after we do spectral decomposition of audio. And this was our motivation for pursuing this topic as consider complex data also for the mask estimation. In this chapter we will talk about first the architecture of complex transformer used for the mask estimation process. We also explore Gaussian-weighted Self-Attention, which is a type of relative attention used in speech signal. The idea behind using relative attention is that in speech signals the current frame is highly correlated with the context frames compare to far away frames, and therefore attention should be based on neighboring frames. In the end of that chapter we discuss experiments that were being performed on CHiME and REVERB data.

In the Chapter 5 we talk about the summary and future work

Chapter 2

Relevant Prior Work

The conventional method of processing the multi-channel audio signal involves the spatial filtering performed via beamforming [7, 8]. The method of beamforming performs a delayed and weighted summation of the multiple spatially separated microphones to provide an enhanced audio signal. The first document on acoustic beamforming was [[9]]. Later with the advancements to the basic beamforming using blind reference-channel selection, two-step time delay of arrival (TDOA) estimation with Viterbi post processing has been proposed to improve the beamforming algorithm [10].

An alternate approach to beamforming using a generalized eigen value(GEV) formulation [2] 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 [2] 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) [3]. The estimate of speech and noise in the given recording thus becomes the key to perform the beamforming in these approaches.

In the work on GEV [2], the speech and noise presence probability or the estimation of power spectral density of speech and noise is done iteratively using Stochastic Gradient Decent(SGD) which is computationally expensive and slow to converge. In the work [11] uses EM Algorithm for localizing and separating sound sources in stereo recordings, but the problem with EM algorithm is same as the previous algorithm i.e. it is slow to converge. Researchers in this field has applied mathematical models and algorithm to efficiently capture the speech in multi channel setting. In one of the work [12], the authors use Gaussian mixture model (GMM) with a Dirichlet prior to localise and estimate the speech source and in the work [13],[14] the author's used clustering approach.

The most successful approach for noise estimation uses a supervised deep neural network (DNN) based speech presence probability (SPP) estimator [15] at every time-frequency bin. The DNN mask estimator is trained using a pair of clean and multi-channel noisy recordings and the DNN learns the SPP with binary targets. The requirement of parallel speech recordings in clean and multi-channel reverberant conditions is a key limitation to these neural mask estimation methods. Recently, unsupervised approaches to mask estimation using complex mixture Gaussian model have been attempted [5, 6]. However, they are either computationally expensive or suffer from a degradation in performance compared to the DNN based mask estimation using oracle targets. Still there is an active research going on in this field, using different deep learning models and using more and more data to estimate better speech and noise mask.

2.1 Signal Model

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, $v^m(k,n)$.

$$y^m(k,n) = \sum_{l=0}^{L_h-1} g^m(k,l)x(k,n-l) + v^m(k,n) \tag{2.1}$$

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.

Now the eq.(2.1) can also be modified as

$$y^m(k,n) = \underbrace{\sum_{l=0}^{D-1} g^m(k,l)x(k,n-l)}_{A} + \underbrace{\sum_{l=D}^{L_h-1} g^m(k,l)x(k,n-l)}_{B} + v^m(k,n) \tag{2.2}$$

where in eq.(2.2), 'A' part models the direct component and early reflections and 'B' part models the late reverb component and 'D' in the equation represents the 'delay' and controls the duration of the early reflection component to be retained. Equation2.2 can also be written as

$$y^m(k,n) = d^m(k,n) + \sum_{l=D}^{L_h-1} g^m(k,l)x(k,n-l) + v^m(k,n) \tag{2.3}$$

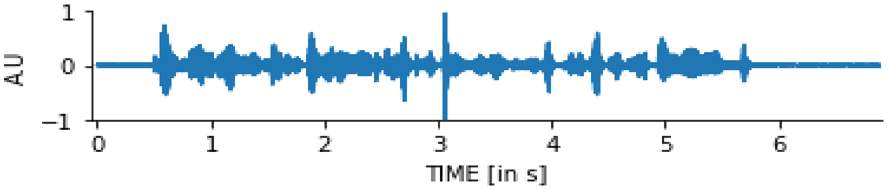

Figure 2.1: Waveform of an audio affected by reverberation

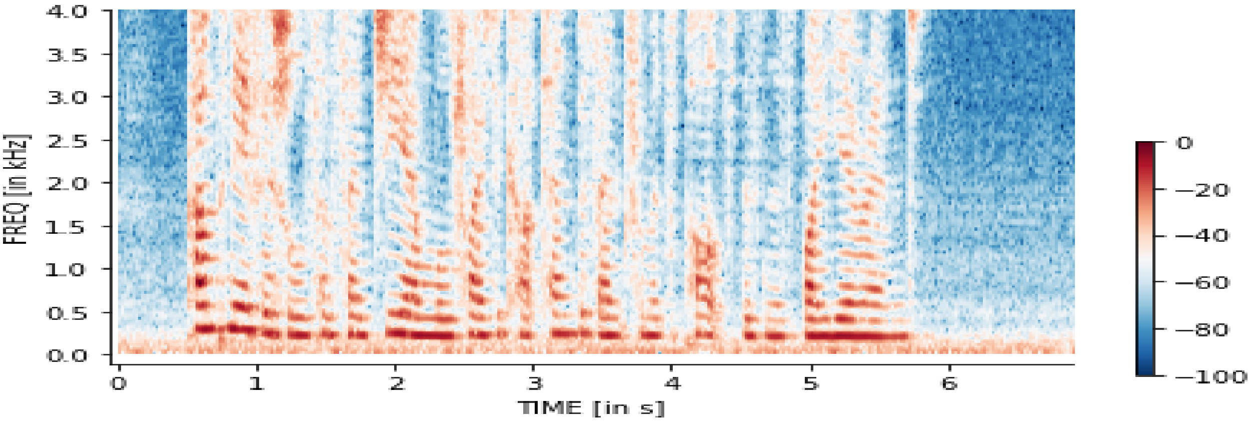

Figure 2.2: Spectrogram of an audio affected by reverberation

2.2 Weighted prediction error (WPE)

For removing the late reflection component or the reverberant component the algorithm that is being used is Normalised Delayed Linear Prediction(NDLP) as mentioned in [16], where the dereverberation is implemented independently on each frequency domain. So the entire problem of NDLP is nothing but a maximum likelihood problem defined separately in individual subbands. With the Short Time Fourier Transform(STFT) decomposition the observation model is defined separately in individual subbands, where the desired signal(mentioned in the previous section) is assumed to be a time varying Gaussian process, i.e it is defined as

$$p(d_{n,f}) = \mathcal{N}_c(d_{n,f}; 0, \rho^2_{n,f}) \tag{2.4}$$

where $\mathcal{N}(.)$ is the pdf of a complex Gaussian random process and variance $\rho^2_{n,f} = E(d_{n,f} d^*_{n,f})$, hence $\rho^2_{n,f}$ is assumed to take different value over different time-frequency bins. The likelihood

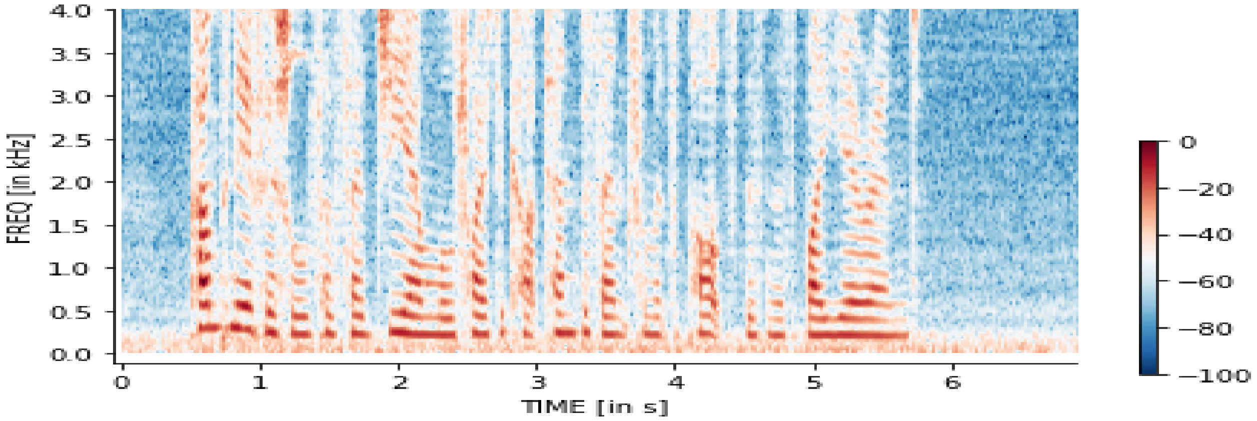

Figure 2.3: Spectrogram of an audio after being applied WPE

function can be represented as:

$$\begin{aligned} \mathcal{L}_f(\theta_f) &= \sum_n \log p(d_{n,f} = x_{n,f} - \overline{c_f}^{*T}\overline{x_{n-D}}; \theta_f) \\ &= -\sum_n \frac{|x_{n,f} - \overline{c_f}^{*T}\overline{x_{n-D}}|^2}{\rho^2_{n,f}} \end{aligned} \tag{2.5}$$

where the parameter $\theta = (\overline{c_f}, \rho^2_{n,f}$ where $\overline{c_f}$ is a regression coefficient and as already mentioned $\rho^2_{n,f})$ is a variance. The eq(2.5) is solved iteratively in which value of these parameters are determined and once we have the value of the parameters we can obtain our desired signal.

$$\hat{c_f} = \left(\sum_n \frac{\bar{x}_{(n-D,f)}\bar{x}^*_{n-D,f}T}{\rho^2_{n,f}}\right)^+ \left(\sum_n \frac{\bar{x}_{(n-D,f)}x^*_{n,f}}{\rho^2_{n,f}}\right) \tag{2.6}$$

$$\hat{\rho^2_{n,f}} = \max(|\hat{d}_{n,f}|^2, \epsilon_f) \tag{2.7}$$

where ϵ_f is very small number and eq(2.6) and eq(2.7) is solved iteratively to obtain the value of parameters, and than from these we solve the equation:

$$\hat{d}_{n,f} = x_{n,f} - c_f^T \bar{x}_{n-D,f} \tag{2.8}$$

2.3 BeamformIt

The Beamformit algorithm as suggested in the paper[10] , is based on weighted-delay and sum techniques where the aim is to create a single enhanced signal from an unknown number of multiple microphone channels.

$$z[n] = \sum_{m=1}^{M} W_m[n] y_m[n - TDOA^{m,ref}[n]] \tag{2.9}$$

where $z[n]$ is the beam formed signal, $W_m[n]$ is the relative weight for each microphone for instant n(with the sum of all weights equal to 1), $y_m[n]$ 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.

The main goal of this algorithm revolves around finding the value of $TDOA^{(m,ref)}$, where there are some preprocessing steps and than some post processing steps. The preprocessing steps involve individual channel signal enhancement which is generally achieved with the help of Wiener filtering, and than to estimate the reference channel. Average of the cross correlation between each channel i and all of others $j = 1...M, j \neq i$ is computed on segments of 1 seconds, as

$$\overline{ycorr_i} = \frac{1}{K(M-1)} \sum_{k=1}^{K} \sum_{j=1, j \neq i}^{M} ycorr[i, j; k] \tag{2.10}$$

where M is the total number of microphone and $K = 200$ indicate the number of one second blocks used in the average. The $ycorr[i, j; k]$ indicates a standard cross correlation measure between channels i and j for each block k. The channel i with the highest average cross correlation is chosen as the reference channel. Steps that are involved are determining the overall channels weighing factor and estimating the skewness in the dataset(in the paper [10] they were using ICSI Meeting and in that slight skewness was present).

The $TDOA^{m,ref}[n]$ is estimated via cross correlation techniques, calculated on the several acoustic frames(in the paper [10], they have taken the window of 250msec). Rather than using the conventional correlation the paper[10] suggest to use the GCC-PHAT(Generalized Cross Correlation with Phase Transform). The reason for not using the conventional correlation is that it is said to be sensitive to noise and reverberation. To compute the GCC-PHAT, given two signal $y_i[n]$ and $y_{ref}[n]$ is done as follows:

$$R_{PHAT}^{i,ref}(d) = F^{-1}\left(\frac{Y_i(f)[Y_{ref}(f)]*}{|Y_i(f)[Y_{ref}(f)]*|}\right) \tag{2.11}$$

where $Y_i(f)$ and $Y_{ref}(f)$ are the Fourier transform of the two signals, F^{-1} indicated the inverse Fourier transform, $[]*$ denoted the complex conjugate and $|.|$ denotes the modulus. The resulting $R^{i,ref}_{PHAT}$ is the correlation function between signals i and ref. The Time delay of arrival (TDOA) for these two micro- phones (i and ref) is estimated as

$$TDOA^i_1 = \arg\max_d R^{i,ref}_{PHAT} \tag{2.12}$$

However rather than calculating just the the max value, in this algorithm we select the N(in the experiment N=4) max values. The idea behind selecting 'N' max value is that, a maximum can occur not due to the source but some spurious noise or laughter from the background and therefore not the select that TDOA value this step is taken. After the calculation of TDOA values, in the post processing step a Viterbi algorithm is run to select the most appropriate value of TDOA for each segment and than used in [2.9] to obtain the single enhanced signal.

2.4 Generalized Eigen Value (GEV) Beamforming

The Generalized Eigen Value based Beamforming can be formulated as follows, given that we have an array of M microphones, the aim is to apply the beamforming operation in frequency domain the spatial filter coefficients $w(m,k)$ to obtain the enhanced signal. The problem can be formulated mathematically as follows

$$z(k,n) = \sum_{m=0}^{M-1} w(m,k)\ y^m(k,n) \tag{2.13}$$

where $z(k,n)$ is the beamformed signal and $y^m(k,n)$ represent the observed signal at the m^{th} microphone which consist of two components, one is desired speech signal $x(k,n)$ and another one is stationary noise component of signal represented by $v^m(k,n)$.

Where the power spectral density(PSD) of the beamformed output can be written as

$$\phi_{ZZ}(k) = \mathbf{W^H(k)\hat{\Phi}_{YY}(k)W(k)} \tag{2.14}$$

$$\phi_{ZZ}(k) = \mathbf{W^H(k)\hat{\Phi}_{XX}W(k)} + \mathbf{W^H(k)\hat{\Phi}_{VV}W(k)} \tag{2.15}$$

And therefore 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 [2], i.e.,

$$\mathbf{w}_{GEV}(k) = \arg\max_{\mathbf{w}(k)} \frac{\mathbf{w}^H(k)\hat{\mathbf{\Phi}}_{XX}(k)\mathbf{w}(k)}{\mathbf{w}^H(k)\hat{\mathbf{\Phi}}_{VV}(k)\ \mathbf{w}(k)} \tag{2.16}$$

where $\hat{\mathbf{\Phi}}_{XX}$ and $\hat{\mathbf{\Phi}}_{VV}$ are power spectral density (PSD) estimates of the clean signal and noise respectively, assuming the clean signal component and the noise signal component are uncorrelated to each other. As in the equation 2.16, the value of filter coefficient $w_{GEV}(k)$ is the eigenvector corresponding to the largest eigenvalue of $\phi_{VV}^{-1}(k)\phi_{XX}(k)$. This is the reason that this method or the beamformer obtained by the Max-SNR criterion is known as *GEV beamformer*(GEV:generalized eigenvalue) based beamformer.

Compared to MVDR(discussed in the next section), GEV beamformer can introduced the speech distortions. In other words both beamformer differ only in a scalar constant r(k), that tries to "normalise" the filter coefficient value $w_{GEV}(k)$. That single channel post filter is given by:

$$\mathbf{r}_{BAN} = \frac{\sqrt{\mathbf{w}_{GEV}^{H}(k)\hat{\mathbf{\Phi}}_{VV}(k)\hat{\mathbf{\Phi}}_{VV}(k)\mathbf{w}_{GEV}(k)/K}}{\mathbf{w}_{GEV}^{H}(k)\hat{\mathbf{\Phi}}_{VV}(k)\ \mathbf{w}_{GEV}(k)} \tag{2.17}$$

where **K** is a normalising factor. This filter performs a Blind Analytic Normalisation(BAN) to obtain the distortion less response in the direction of the speaker. As the name suggest "BAN(Blind Analytic Normalisation)", its closed form expression can be computed. In the conventional paper on GEV i.e [2], the generalized eigenvalue problem is solved either by the deterministic gradient descent method or Stochastic Gradient Ascent, by finding the value of $\phi_{XX}(k)$ and $\phi_{VV}(k)$ iteratively. But in the next chapter we discuss how the deep learning framework is used to determine the value of $\phi_{XX}(k)$ and $\phi_{VV}(k)$, which will provide us the faster and efficient way to estimate these power spectral density (PSD) matrices.

2.5 Minimum Variance Distortionless Response(MVDR) Beamforming

Minimum Variance Distortionless Respose(MVDR) Beamforming is the data adaptive algorithm which is widely used in the acoustic beamforming with many variants that exist like Minimum Power Distortionless Respose(MPDR) Beamforming. The idea of MVDR can be traced back to 1969, which is proposed by Capon. Thus it is also known as Capon Beamformer. So the idea of MVDR is it minimizes the residual noise with the constraint, that any signal arriving from the source direction remain distortion less and hence it received the name 'Distortionless

Response', the objective equation is as follows:

$$\mathbf{w}_{GEV}(k) = \arg\min_{\mathbf{w}(k)} \mathbf{w}^H(k)\hat{\mathbf{\Phi}}_{VV}(k)\mathbf{w}(k) \quad \ni \mathbf{w}^H d = 1 \tag{2.18}$$

where $\mathbf{d}$ is the DOA vector, where vector $\mathbf{d}$ is principal component of the estimated power spectral density matrix of speech i.e. $\mathbf{d} = P(\hat{\mathbf{\Phi}}_{XX}(k))$.

Solving the equation 2.18 will give us the following objective function:

$$\mathbf{w}_{MVDR}(k) = \frac{\hat{\mathbf{\Phi}}_{VV}^{-1}(k)\mathbf{d}}{\mathbf{d}^H\hat{\mathbf{\Phi}}_{VV}^{-1}(k)\ \mathbf{d}} \tag{2.19}$$

where $\hat{\mathbf{\Phi}}_{XX}$ and $\hat{\mathbf{\Phi}}_{VV}$ are power spectral density (PSD) estimates of the clean signal and noise respectively, and similar to like the

The most successful approach to the estimation of clean and noise PSD is through the use of a neural mask estimator (will be described in the next chapter). Once the PSD matrices are estimated, we can easily solve the equations 2.16 and equation 2.19, and obtain the value of spatial filter coefficient.

Chapter 3

Unsupervised Mask Estimation

3.1 Neural Mask Estimator

As proposed in [15, 17], the neural mask estimators are deep feed-forward/recurrent networks that are trained to predict the speech presence probability in each time-frequency bin. Figure 3.1 and 3.2 show the network architecture. 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 threshold different for voiced and unvoiced regions of the audio [15].

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). The masks for each channel are then condensed to a single speech and single noise mask using a median operation. The median is preferred over a mean computation because of its resilience to outliers. Once the mask estimator is trained, the PSD matrices needed in Eq. (2.16) for $\mathbf{y}(k,n) = [y^0(k,n), .., y^{M-1}(k,n)]^T$ is,

$$\hat{\mathbf{\Phi}}_{XX}(k) = \frac{\sum_n s(k,n)\mathbf{y}(k,n)(\mathbf{y}(k,n))^H}{\sum_n s(k,n)} \tag{3.1}$$

$$\hat{\mathbf{\Phi}}_{NN}(k) = \frac{\sum_n u(k,n)\mathbf{y}(k,n)(\mathbf{y}(k,n))^H}{\sum_n u(k,n)} \tag{3.2}$$

As mentioned in the paper [15] for the BLSTM layer, the weights are drawn from a uniform distribution ranging from-0.04 to 0.04 and biases are initialized with zeros. RMSProp with

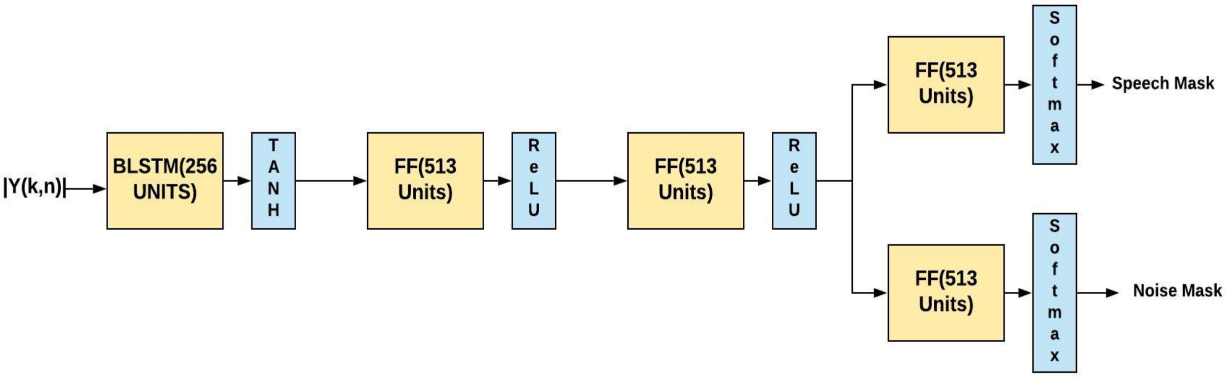

Figure 3.1: BLSTM network configuration for mask estimation

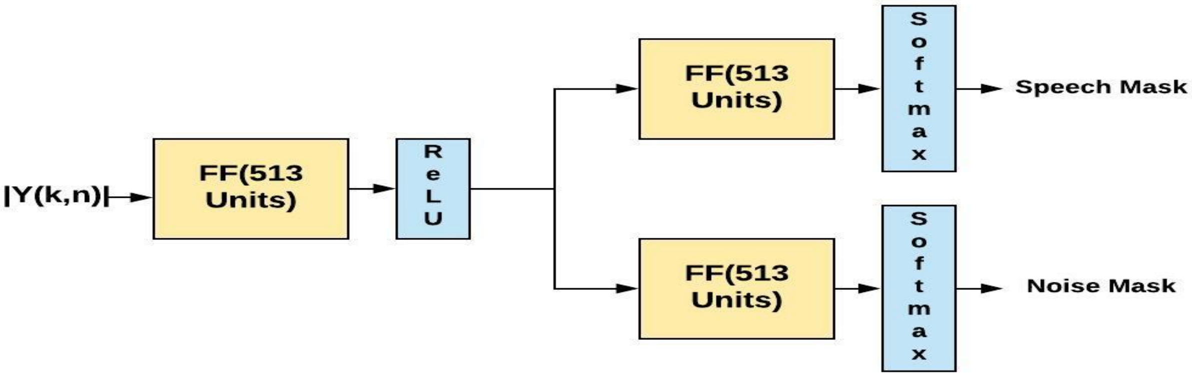

Figure 3.2: Feed Forward network configuration for mask estimation

momentum of 0.9 is being employed for learning with the learning rate of 0.001. Also to achieve the better generalization, dropout with the value of $p = 0.5$ is being used.0.5 is chosen as empirically it has shown to give better word error rate. Batch normalisation is also used. The loss function that is used to train the model is binary cross entropy i.e estimated speech and noise mask are compared with the ideal speech and ideal noise mask.

One of the limitations of the neural mask estimation described 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 this section, we propose to move away from the requirement of having simulated settings by generating unsupervised pseudo targets for the real (and simulated) multi-channel recordings.

3.2 Joint Spatial Filtering and Multi Channel Linear Prediction(MCLP)

We use the joint spatial filtering and multi-channel linear prediction (MCLP) approach with Bayesian inference proposed in [4, 18] to derive the unsupervised targets for the neural mask estimator. For a single reference signal characterized by STFT coefficients $d_1(k,n)$, the MCLP model for a m^{th} microphone signal [4] is given as,

$$y^m(k,n) = a^m(k)d_1(k,n) + (\mathbf{g}^m(k))^H \boldsymbol{\phi}(k,n) \tag{3.3}$$

where $y^m(k,n)$ is the STFT of the mth microphone signal, $a^m[k]$ is the relative gain of the desired signal collected at the mth microphone. The late reflection components in the multichannel signal are modeled as a linear prediction with $\mathbf{g}^m(k)$ denoting a vector of LM prediction coefficients and $\boldsymbol{\phi}(k,n) = [y^1(k,n-D-1),..,y^1(k,n-D-L),...,y^M(k,n-D-1),..,y^M(k,n-D-L)]^T$ is the LM dimensional vector containing the delayed STFT components from all the M microphones for L previous lags. In vector form,

$$\mathbf{y}(k,n) = \mathbf{a}(k)d_1(k,n) + \mathbf{G}(k)^H \boldsymbol{\phi}(k,n) \tag{3.4}$$

where $\mathbf{G}[k]$ is the $LM \times M$ MCLP filter coefficients, $\mathbf{a}[k]$ is the relative transfer function (RTF) of each microphone with respect to the reference signal.

A spatial filter $\mathbf{w}(k)$ is constructed such that $\mathbf{w}^H(k)\mathbf{a}(k) = 1$. This gives,

$$\mathbf{w}^H(k)\mathbf{y}(k,n) = d_1(k,n) + \mathbf{w}^H(k)\mathbf{G}(k)^H \boldsymbol{\phi}(k,n) \tag{3.5}$$

By assuming a complex circular Gaussian prior on the desired source signal, $p(d_1(k,n)) \sim N_c(d_m(k,n); 0, \gamma_{kn})$, a maximum likelihood (ML) approach to parameter estimation can be pursued [18]. This ML problem can be solved using a coordinate ascent method where the parameters of the model (Θ) containing the MCLP prediction coefficients $\mathbf{G}(k)$, the RTF $\mathbf{a}(k)$, the spatial filter $\mathbf{w}(k)$ and the unknown variance γ_{kn} are iteratively estimated.

The solution to the ML estimation problem [18] is given below. The ML problem can be equivalently stated as,

$$maximize \quad \sum_{n=0}^{N-1} \gamma_{kn}^{-1} \left| \mathbf{w}[k]^H \left[\mathbf{y}[k,n] - \mathbf{G}[k]^H \boldsymbol{\phi}[k,n] \right] \right|^2, \tag{3.6}$$

$$\textit{subject to } \mathbf{w}[k]^H \mathbf{a}[k] = 1. \tag{3.7}$$

The prediction filter $\mathbf{G}[k]$, spatial filter $\mathbf{w}[k]$ and the RTF $\mathbf{a}[k]$ are estimated sequentially in an iterative scheme, using the equations given below

$$\hat{\mathbf{G}}(k) = \mathbf{R}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{-1}(k)\mathbf{R}_{\boldsymbol{\phi}\boldsymbol{y}}(k) \tag{3.8}$$

where

$$\mathbf{R}_{\boldsymbol{\phi}\boldsymbol{\phi}}(k) = \sum_{n=0}^{N-1} \gamma_{kn}^{-1} \boldsymbol{\phi}(k,n)\boldsymbol{\phi}^H(k,n) \tag{3.9}$$

$$\mathbf{R}_{\boldsymbol{\phi}\boldsymbol{y}}(k) = \sum_{n=0}^{N-1} \gamma_{kn}^{-1} \boldsymbol{\phi}(k,n)\mathbf{y}^H(k,n) \tag{3.10}$$

Once the MCLP prediction coefficients $\hat{\mathbf{G}}(k)$ are estimated, the RTF vector $\mathbf{a}(k)$ can estimated as the first column of the prediction residual, i.e., $\mathbf{y}[k,n] - \hat{\mathbf{G}}[k]^H \boldsymbol{\phi}[k,n]$.

Let $\mathbf{R}_{\hat{r}\hat{r}}$ denote the spatial correlation matrix of the predicted reverberation component $\hat{\mathbf{r}} = \hat{\mathbf{G}}^H(k)\boldsymbol{\phi}(k,n)$. Then, the spatial filter can be estimated as,

$$\hat{\mathbf{w}}(k) = \frac{\mathbf{R}_{\hat{\mathbf{r}}\hat{\mathbf{r}}}^{-1}\hat{\mathbf{a}}}{\hat{\mathbf{a}}^H \mathbf{R}_{\hat{\mathbf{r}}\hat{\mathbf{r}}}^{-1}\hat{\mathbf{a}}}. \tag{3.11}$$

Finally, the desired signal variance γ_{kn} is estimated using an AR modeling approach on the estimate of the desired signal $d_1(k,n)$ [4]. More details on the ML estimation can be found in [18, 4]

Using the iterative procedure outlined above, the estimation of the late reflection components and beamforming of the desired source signal are jointly performed. The output estimate of $d_1(k,n)$ is used as the estimate of the clean signal in the GEV beamforming (Fig. 3.3).

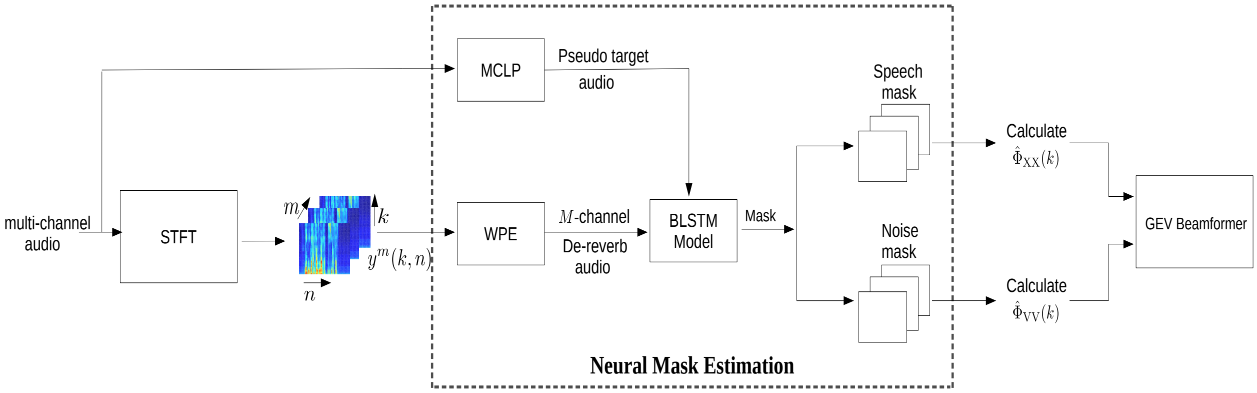

Figure 3.3: Block schematic of the unsupervised neural network based mask estimation.

3.3 Mask Estimation Procedure

[1]

The experimental setup is show in the fig(3.3), and the process is as follows. A multi-channel audio signal is taken and its 512 point Short Time Fourier Transform(STFT) is computed with the window(hann window) of 30 msec and and the overlap of 15 msec to form a 3 dimensional ($F \times T \times M$) tensor where length, height and width represents number of frequency bins F, time frames T and number of channels M respectively. By segregating voiced and unvoiced section in each frequency bin, an ideal binary mask (IBM) is estimated for the 3D input using the MCLP based beamformed audio as the target [15].

The model architecture for the mask estimation uses a Bi-directional Long Short-term memory (BLSTM) network followed by two fully connected layers. We use Rectified linear unit (ReLU) activation function for the first two layers and Sigmoid for the last layer. A dropout regularization is used with dropout parameter of 0.5 after every layer. For training the unsupervised model, the targets are derived from the audio beamformed using the method of multi-channel linear prediction as described in previous section. The speech and noise masks are estimated using the model for all the channels jointly. A single speech mask and noise mask (complimentary to the speech mask) are generated by taking the median of all the masks from the multiple channels. The $\hat{\mathbf{\Phi}}_{XX}$ and $\hat{\mathbf{\Phi}}_{VV}$ are calculated and the beamformed STFT estimate is then converted back to the audio signal using overlap synthesis. These audio signals are converted to acoustic features for ASR training and testing.

[1] 'UNSUPERVISED NEURAL MASK ESTIMATOR FOR GENERALIZED EIGEN-VALUE BEAMFORMING BASED ASR',Rohit Kumar, Anirudh Sreeram, Anurenjan Purushothaman, Sriram Ganapathy,ICASSP 2020

Table 3.1: CLSTM model architecture.

Layer	Configuration
Conv 2D (ReLU)	filters = 128, kernel (3,3)
Conv 2D (ReLU)	filters = 128, kernel (3,3)
Maxpooling 2D	size = (2,2)
Conv 2D (ReLU)	filters = 64, kernel (3,3)
Conv 2D (ReLU)	filters = 64, kernel (3,3)
LSTM (ReLU)	1024 units (frequency recurrence)
DNN (ReLU)	1024 units
DNN (ReLU)	1024 units
DNN (Softmax)	senone posteriors

3.4 Experiments and Results

3.4.1 ASR setup

The ASR system uses filter-bank (FBANK) features that are 40 log-mel spectrogram features extracted every 25ms windows with a shift of 10ms on multi-channel audio signals that are enhanced with WPE [19]. We use the Kaldi toolkit [20] for deriving the senone alignments used in the PyTorch deep learning framework. A hidden Markov model - Gaussian mixture model (HMM-GMM) system is initially trained to generate the alignments. The acoustic model used in this work is a convolutional long short term memory (CLSTM) model where the LSTM recurs over frequency. The configuration of the CLSTM model is given in Table 3.1. A dropout of 20% and batch normalization is used after every layer for regularization. For the ASR decoding, an initial tri-gram model is used to generate a lattice rescored with a recurrent neural network (RNN) [21]. The proposed method of beamforming using the psuedo mask estimates from a multi-channel linear prediction based beamformer is compared with the beamforming using delay-sum and Viterbi algorithm (BeamformIt [10]), a 3-D CNN based neural acoustic model which jointly performs beamforming and ASR [22] and the generalized eigen-value (GEV) based beamforming with supervised mask estimation on the simulated data [15].

3.4.2 CHiME-3 ASR

The CHiME-3 corpus for ASR contains multi-microphone tablet device recordings from everyday environments, released as a part of 3rd CHiME challenge [23]. Four varied environments are present, cafe (CAF), street junction (STR), public transport (BUS) and pedestrian area (PED). For each environment, two types of noisy speech data are present, real and simulated. The real data consists of 6-channel recordings of sentences from the WSJ0 corpus spoken in

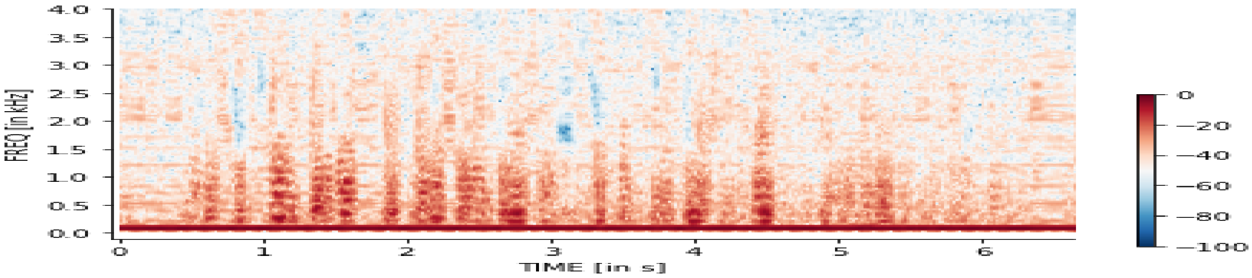

Figure 3.4: Spectrogram of an audio after being applied GEV

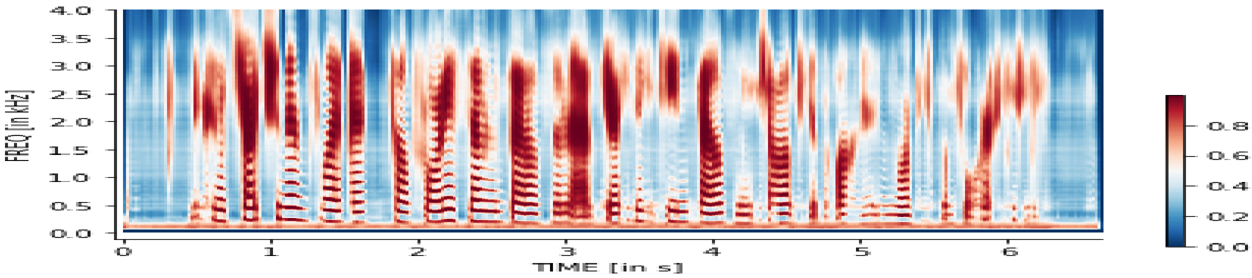

Figure 3.5: Speech Presence Mask that is estimated by the model for CHiME Data

Table 3.2: Word Error Rate (%) for CHiME-3 dataset.

Training	Dev			Eval		
	Real	Sim	Avg	Real	Sim	Avg
BeamformIt [10]	6.1	8.4	7.3	13.0	12.7	12.9
3-D CNN [22]	7.2	7.2	7.2	15.4	9.1	12.2
Sup. Out-of-dom. GEV	7.1	8.8	7.9	11.2	10.7	10.9
Unsup. MVDR	4.9	6.2	5.5	9.4	7.4	8.4
Unsup. GEV	**4.9**	**5.8**	**5.3**	**9.0**	**7.3**	**8.1**
Sup. oracle MVDR [17]	5.1	6.5	5.8	9.1	7.5	8.3
Sup. oracle GEV [15]	4.9	6.1	5.5	9.4	7.2	8.3

the environments listed above. The simulated data was constructed by artificially mixing clean utterances with environment noises. The training data has 1600 (real) noisy recordings and 7138 simulated noisy utterances. The development (dev) and evaluation (eval) data consists of the 410 and 330 utterances respectively. For each set, the sentences are read by four different talkers in the four CHiME-3 environments. This results in 1640 (410 × 4) and 1320 (330 × 4) real development and evaluation utterances in total.

The results for the CHiME-3 ASR system with various beamforming methods are given in

Table 3.3: WER (%) for each noise condition in CHiME-3 dataset with supervised and unsupervised GEV beamforming methods.

Cond.	**Dev Data**				**Eval Data**			
	Sim		Real		Sim		Real	
	Sup	Unsup	Sup	Unsup	Sup	Unsup	Sup	Unsup
BUS	4.9	4.9	6.0	6.0	5.9	6.1	12.8	11.9
CAF	8.1	7.6	4.8	4.8	8.3	8.1	8.8	8.5
PED	5.6	5.3	4.2	4.2	7.1	7.0	9.2	8.9
STR	5.7	5.4	4.6	4.7	7.3	7.9	6.8	6.6

Table 3.2. The ASR results for the BeamformIt [10] are similar to the 3-D CNN model [22]. The GEV based on out-of-domain set consists of training the neural mask estimation on Reverb Challenge dataset (described next) and using the mask estimator outputs for GEV beamforming on CHiME-3 dataset. While there is a domain mis-match, this approach provides the best baseline beamforming system, particularly on the evaluation data of CHiME-3.

The supervised GEV/MVDR using oracle mask estimates on the in-domain CHiME-3 dataset provides the upper bound in terms of the performance of the unsupervised methods. The proposed unsupervised GEV beamforming using the MCLP based source signal targets provides very similar results to the supervised oracle mask estimation based GEV. In terms of relative improvements over the BeamformIt method and the out-of-domain mask estimation based GEV, the proposed approach yields about 27 % and 35 % respectively on the development data and 37 % and 25 % respectively on the evaluation data.

The comparison of the supervised and unsupervised approaches on the different noise conditions of the CHiME-3 dataset are shown in Table 3. As seen in this Table, for most of the noise conditions, the unsupervised method compares well with the supervised mask estimation approach. A degradation is seen in the unsupervised case for “Street” noise in simulated conditions. However, a good improvement in ASR performance is seen for “Bus” noise in real evaluation conditions for the unsupervised approach as well.

3.4.3 Reverb Challenge

The Reverb Challenge dataset [24] contains recordings with real and simulated reverberation condition, recorded using 8 channels for the ASR task. 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 real data consists of utterances spoken by human speakers in a noisy reverberant room, with utterances from the multi-channel Wall Street Journal audio-

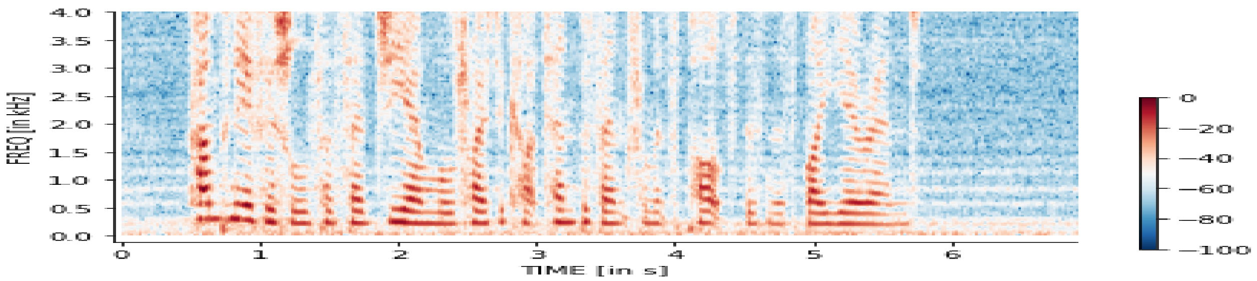

Figure 3.6: Spectrogram of an audio after being applied GEV

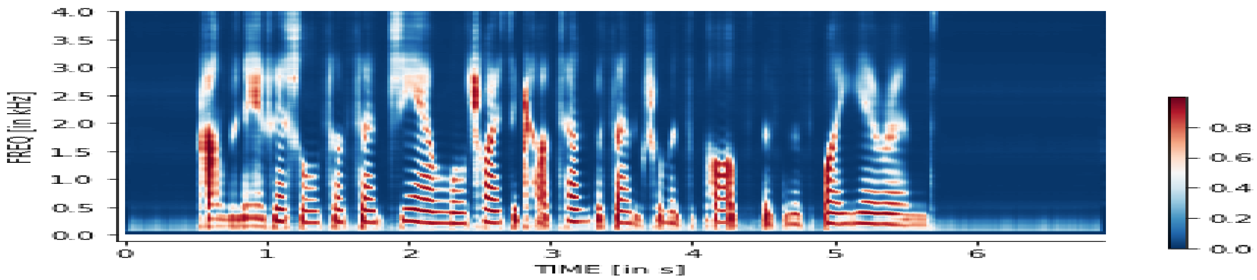

Figure 3.7: Speech Presence Mask estimated by the model for REVERB14 audio

Table 3.4: Word Error Rate (%) for REVERB Challenge dataset using various beamforming methods.

Training	Dev			Eval		
	Real	Simu	Avg	Real	Simu	Avg
BeamformIt [10]	19.7	6.2	12.9	22.2	6.5	14.4
3-D CNN [22]	20.4	6.7	13.5	21.2	6.6	13.9
Unsup. MVDR	17.2	**5.1**	11.2	14.9	5.6	10.3
Unsup. GEV	**15.6**	5.6	**10.6**	**13.5**	**5.3**	**9.4**
Sup. oracle MVDR [17]	17.5	5.2	11.3	13.0	5.3	9.2
Sup. oracle GEV [15]	17.0	5.6	11.3	13.0	5.3	9.2

visual (MC-WSJ-AV) corpus [3]. 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 results for the various beamforming methods on the Reverb Challenge dataset are shown in Table 3.4. The unsupervised beamforming method improves significantly over the BeamformIt method and the 3-D CNN approach. On the average, the unsupervised mask estimation approach performs similar to the supervised mask estimation approach in the GEV/MVDR beamforming. The unsupervised approach improves the BeamformIt approach relatively by 18% on the development data and 35 % on the evaluation data. The ASR results on the Reverb Challenge dataset are seen to be consistent with those for the CHiME-3 dataset.

3.4.4 CHiME v/s Reverb Challenge Data

The two dataset that we are dealing with, have subtle difference between them.

- The first difference is the corpus, in REVERB Challenge where simulated data utterances are taken from the WSJCAM0 and real data is taken from the MC-WSJ-AV corpus. In CHiME, utterances are taken from the WSJ0.

- The second and the most prominent difference between the two corpus is type of noise that is there in both the dataset. In REVERB we have reverberant noise(with the RT60 of 0.25sec,0.50sec and 0.75sec) and along with that stationary background noise, which is caused mainly by air conditioning systems in a room in real scenario, and in simulated scenario we add a noise at an SNR of 20 dB. Whereas in CHiME dataset, there is no reverberant noise that is present in data, only the four different type of environment noises are present, as discussed in the previous section.

- The third difference that is present in both is dataset is the number of microphone we have. In the REVERB dataset we have 8 channel audio file is there. Whereas in CHiME we have 6 channel audio file is there.

- The fourth difference between both the dataset is arrangement or geometry of microphones. In the REVERB dataset, circular array of omnidirectional microphone with diameter of 20 cm is being used. Whereas in the CHiME dataset, all microphone face forward (i.e. toward the speaker holding the tablet) apart form the top-center microphone(mic 2) which is faces backwards.

3.5 Summary

In summary, we have proposed an unsupervised mask estimation approach for the GEV beamforming. The mask estimation is based on the joint estimation of the late reverberation component and a spatial filter that performs the beamforming to identify the clean source signal. This estimation is based on a maximum likelihood framework in a multi-channel linear prediction setting. The estimate of the clean source signal is used in the neural mask estimator to generate the speech presence probability which is in turn used in the generalized eigen value beamforming. Several ASR experiments on the CHiME-3 and the Reverb Challenge datasets confirm that the proposed approach of unsupervised mask estimation achieves performance similar to the supervised oracle mask estimation using paired clean and noisy audio recordings.

Chapter 4

Complex Mask Estimation

4.1 Introduction

Deep Learning has seen a huge interest and change in past decade, however major deep learning models rarely use complex numbers. This problem is especially important for speech community because the audio data that we handle is naturally complex valued after we do spectral decomposition of audio. In the previous chapter, we have trained our NN mask estimator using the real data. However recently the Transformer [25] was presented as new sequence learning architecture with significant improvements over RNNs in machine translation and many other natural language processing tasks. The transformer uses the attention mechanism to compute the symbol-by-symbol correlations in parallel, over the entire input sequence such that we can find the similarity or dependence of symbol over the other symbol over the time.

There are mainly two reason for using Transformer in this section as our architecture compare to any sequential neural network. First the Transformer can process an input sequence in parallel, which can significantly reduce training and inference times. And the second reason is somewhat important i.e if we take any sequential network even LSTM or GRU which claim to have long term dependencies, in practical scenario the history that they can look backward is still limited and it is difficult to learn long-range dependencies between symbols. However Transformer has resolved this issue with the self attention mechanism and can even model those long term dependencies.

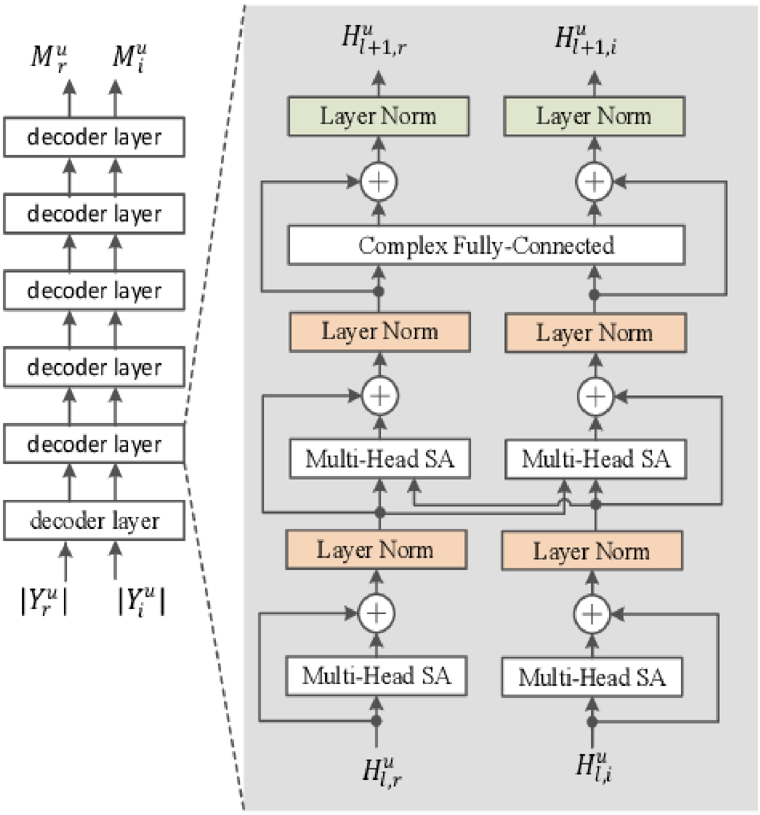

Figure 4.1: Complex Transformer Architecture

4.2 Complex Transformer Based Architecture

The model architecture that we use is proposed in the paper [26] and shown in figure 4.1. The complex transformer architecture for mask estimation unlike the previous section mask estimation(Bi-LSTM based) takes two inputs that is one input represent the real part of the input STFTs and other input represent the imaginary part of the input STFTs. The network inputs, Y_r^u and Y_i^u, are the real and imaginary parts of the input noisy spectrum. However unlike the normal Transformer which has encoder and decoder architecture, in our mask estimation process we consider only the encoder part to predict the speech and noise mask.

Also unlike the paper [26], where they will spit out two output, one output represent the real part of speech and other output imaginary part of speech, in our architecture we concatenate those real and imaginary mask and than pass it through a linear layer and sigmoid non linearity to convert to a single speech presence probability mask and the same procedure is done for getting the noise presence probability mask. After getting those estimated mask for training we compare them with ideal speech and noise mask similar to what we have done in the previous section. The loss function that is used for training the model is binary cross entropy

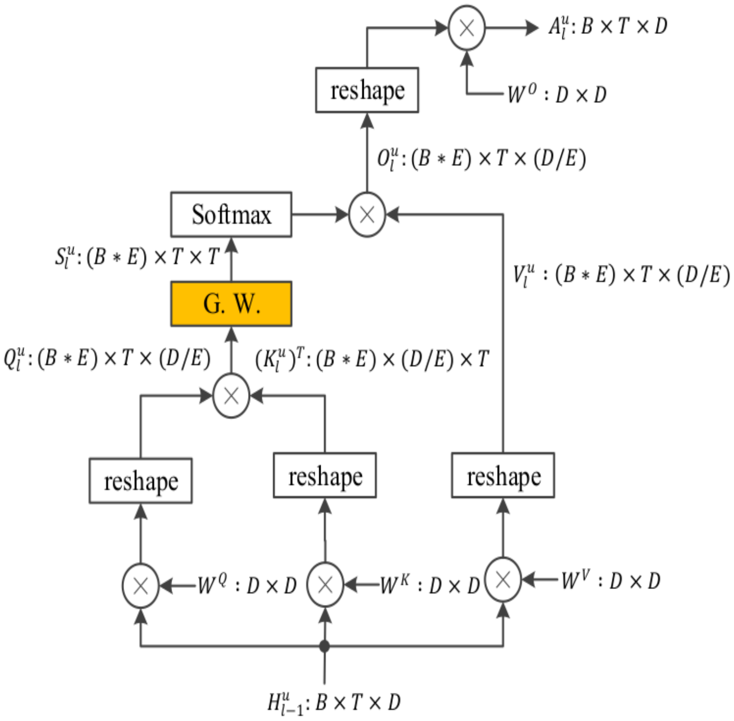
$A_l^u: B \times T \times D$
reshape
$W^O: D \times D$
$O_l^u: (B * E) \times T \times (D/E)$
Softmax
$S_l^u: (B * E) \times T \times T$
G. W.
$V_l^u: (B * E) \times T \times (D/E)$
$Q_l^u: (B * E) \times T \times (D/E)$
$(K_l^u)^T: (B * E) \times (D/E) \times T$
reshape
reshape
reshape
$W^Q: D \times D$
$W^K: D \times D$
$W^V: D \times D$
$H_{l-1}^u: B \times T \times D$

Figure 4.2: Block Diagram of Gaussian weighted based self attention

and optimizer that is being used is Adam. Also unlike the NLP, the speech data that we deal is highly correlated with the closer components. Keeping this thing in mind, as mentioned in the paper [26] a positional encoding is required to penalize attention weights according to the acoustic signal characteristics, such that less attention is provided to more distant symbols. This is similar to idea of relative attention that is first used in paper [27] and will be described in detail in the next section.

4.2.0.1 GSA: Gaussian-weighted Self-Attention

As mentioned in the previous paragraph to apply relative attention we have used the Gaussian weighted mask with the learn-able variance as shown in the figure 4.2 where B, T and D are the batch size,sequence length and input dimension. E is the number of self attention units. Since we are using a complex architecture there is little bit difference compare to conventional transformer.

When we compute the query matrix $\mathbf{Q} = \mathbf{XW_Q}$, the key matrix $\mathbf{K} = \mathbf{XW_K}$ and the value matrix $\mathbf{V} = \mathbf{XW_V}$,the $\mathbf{Q}$,$\mathbf{K}$, and $\mathbf{V}$ are complex values and the complex attention as mentioned in paper [28] is given by:

$$\begin{aligned}
\mathbf{QK^TV} &= (\mathbf{XW_Q})(\mathbf{XW_K})^T(\mathbf{XW_V}) \\
&= (\mathbf{AW_Q} + \mathbf{iBW_Q})(\mathbf{W_K^TA^T} + \mathbf{iW_K^TB^T})(\mathbf{AW_V} + \mathbf{iBW_V}) \\
&= (\mathbf{AW_QW_K^TA^TBW_V} - \mathbf{AW_QW_K^TB^TBW_V} \\
&- \mathbf{BW_QW_K^TA^TBW_V} - \mathbf{BW_QW_K^TB^TAW_V}) \\
&+ i(\mathbf{AW_QW_K^T}A^TBW_V + \mathbf{AW_QW_K^TB^TAW_V} \\
&+ \mathbf{BW_QW_K^TA^TAW_V} - \mathbf{BW_QW_K^TB^TBW_V}) \\
&= \mathbf{A}^{'} + \mathbf{iB}^{'}
\end{aligned}$$

where $\mathbf{A}^{'}$ and $\mathbf{B}^{'}$ represent the real and the imaginary part of the complex attention result respectively.The tensors $\mathbf{W_Q}$,$\mathbf{W_K}$ and $\mathbf{W_V}$ are the learn-able parameters

Now to apply the relative multi head attention, the Gaussian Mask or Gaussian weighting matrix is applied to the score matrix, which is computed from the key and query matrix multiplication as follows:

$$\mathbf{S_l^u} = \mathbf{G_l} \odot \frac{\mathbf{Q_l^u}(\mathbf{K_l^u})^\mathbf{T}}{\sqrt{d}} = \mathbf{G_l} \odot \mathbf{C_l^u} \tag{4.1}$$

where $\mathbf{l}$ represents the hidden layer output, where the Gaussian Matrix in a simple term is nothing but a Hermitian matrix where the entries of vary like a Gaussian distribution with mean zero and variance σ, which is also the learn-able parameter, initialised randomly. Each entry of a Gaussian matrix is given by $\exp(\frac{-(i-j)^2}{\sigma^2})$, where i represent the target frame index and and j represent the context frame index. The diagonal term will always be 1 and the value of non diagonal elements vary depending upon the value of σ i.e. their values will vary inversely proportional to the distance between the target and context frame therefore more importance will be given to context compare to far away speech in the acoustic data. Also the absolute value of the score matrix should be considers, as this imply that negative correlations are as important as the positive correlation and we are looking at the both side of the target frame with equal importance.

After applying with Gaussian weight matrix,as shown in the figure it is passed through the softmax non linearity and than multiplied with the value matrix .

4.3 Experiments and Results

4.3.1 Mask Estimation

The experimental setup for all the experiments are as follows. A 512 point Short Time Fourier Transform (STFT) of the multi-channel audio signal is computed jointly to form a 3 dimensional $(T \times M \times F)$ tensor where length, height and width represents number of frequency bins F, time frames T and number of channels M respectively. And similar to like unsupervised mask estimation the targets are obtained by segregating voiced and unvoiced section in each frequency bin, an ideal ratio mask (IRM) is estimated for the 3D input using the MCLP based beamformed target [15].

The network architecture is as follows, we are using 6 encoder layer Transformer with 8 multi-head attention. In each encoder layer we have a self attention unit and a feed forward network. As mentioned in the previous paragraph we have two input tensors to our transformer network each of dimension $(T \times M \times F)$, where one input represent the real part of the input and other input represent the imaginary part of the input. After passing through the transformer we will obtain two outputs, we will going to concatenate those tensors along columns such that the dimension of concatenating tensor is $(T \times M \times 2 * F)$. Now this tensors is passed through two independent linear layer network with sigmoid non linearity that will convert these concatenated tensors into a size of $(T \times M \times F)$ tensor and output represent the speech and noise mask respectively. The model is trained for 25 epochs using Adam as our optimiser. The loss function that is being used is same as that used in previous chapter and that is binary cross entropy. A dropout regularization is used with dropout parameter of 0.5 after every layer. For training the model, the targets are derived from the audio beamformed using the method of multi-channel linear prediction described in section 3.2. The speech and noise masks are estimated using the model for all the channels jointly. A single speech mask and noise mask (complimentary to the speech mask) are generated by taking the median of all the masks from the multiple channels. The $\hat{\mathbf{\Phi}}_{XX}$ and $\hat{\mathbf{\Phi}}_{VV}$ are estimated and from those the beamformed STFT is obtained which is then converted back to the audio signal using overlap synthesis. These audio signals are converted to acoustic features for ASR training and testing.

4.3.2 ASR Setup

The ASR system uses filter-bank (FBANK) features that are 40 log-mel spectrogram features extracted every 25ms windows with a shift of 10ms on multi-channel audio signals that are enhanced with WPE [19]. We use the Kaldi toolkit [20] for deriving the senone alignments used in the PyTorch deep learning framework. A hidden Markov model - Gaussian mixture model

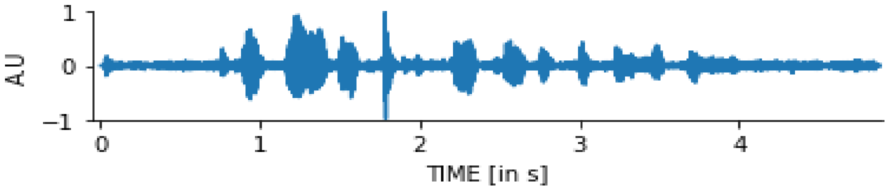

Figure 4.3: Waveform of an audio after being applied Complex Mask estimation GEV Beamforming

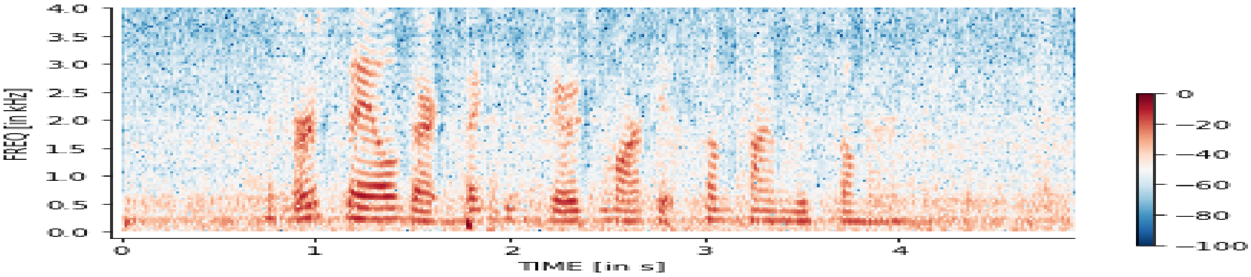

Figure 4.4: Spectrogram of an audio after being applied Complex Mask estimation GEV Beamforming

(HMM-GMM) system is initially trained to generate the alignments. The acoustic model used in this work is a convolutional long short term memory (CLSTM) model where the LSTM recurs over frequency. The configuration of the CLSTM model is given in Table 3.1. A dropout of 20% and batch normalization is used after every layer for regularization. For the ASR decoding, an initial tri-gram model is used to generate a lattice rescored with a recurrent neural network (RNN) [21]. The proposed method of beamforming using the psuedo mask estimates from a multi-channel linear prediction based beamformer is compared with the beamforming using delay-sum and Viterbi algorithm (BeamformIt [10]), a 3-D CNN based neural acoustic model which jointly performs beamforming and ASR [22] and the generalized eigen-value (GEV) based beamforming with supervised mask estimation on the simulated data [15].

4.3.3 Chime-3 Data

The CHiME-3 corpus for ASR contains multi-microphone tablet device recordings from everyday environments, released as a part of 3rd CHiME challenge [23]. Four varied environments are present, cafe (CAF), street junction (STR), public transport (BUS) and pedestrian area (PED). For each environment, two types of noisy speech data are present, real and simulated. The real data consists of 6-channel recordings of sentences from the WSJ0 corpus spoken in the environments listed above. The simulated data was constructed by artificially mixing clean utterances with environment noises. The training data has 1600 (real) noisy recordings and

Table 4.1: Word Error Rate (%) for CHiME-3 dataset.

Training	Dev			Eval		
	Real	Sim	Avg	Real	Sim	Avg
BeamformIt [10]	6.1	8.4	7.3	13.0	12.7	12.9
3-D CNN [22]	7.2	7.2	7.2	15.4	9.1	12.2
Sup. Out-of-dom. GEV	7.1	8.8	7.9	11.2	10.7	10.9
Unsup. MVDR	4.9	6.2	5.5	9.4	7.4	8.4
Unsup. GEV	**4.9**	**5.8**	**5.3**	**9.0**	**7.3**	**8.1**
Sup. oracle MVDR [17]	5.1	6.5	5.8	9.1	7.5	8.3
Sup. oracle GEV [15]	4.9	6.1	5.5	9.4	7.2	8.3
Complex Transf. Unsup MVDR	6.39	8.65	7.52	14.94	16.66	15.8
Complex Transf. Unsup GEV	6.23	8.48	7.35	14.62	16.29	15.45

7138 simulated noisy utterances. The development (dev) and evaluation (eval) data consists of the 410 and 330 utterances respectively. For each set, the sentences are read by four different talkers in the four CHiME-3 environments. This results in 1640 (410 × 4) and 1320 (330 × 4) real development and evaluation utterances in total.

The results for the CHiME-3 ASR system with various beamforming methods are given in Table 4.1.In the Table 4.1 we can see that the ASR results for Complex Transformer based beamforming were worse than the baseline model i.e. unsupervised based GEV beamforming.

4.3.4 Reverb Data

The Reverb Challenge dataset [24] contains recordings with real and simulated reverberation condition, recorded using 8 channels for the ASR task. 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 real data consists of utterances spoken by human speakers in a noisy reverberant room, with utterances from the multi-channel Wall Street Journal audio-visual (MC-WSJ-AV) corpus [3]. 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 results for the various beamforming methods on the Reverb Challenge dataset are shown in Table 4.2. As consistent with the CHiME data here also the complex transformer

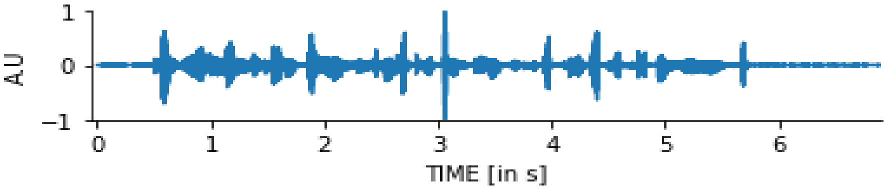

Figure 4.5: Waveform of an audio after being applied Complex Mask estimation GEV Beamforming

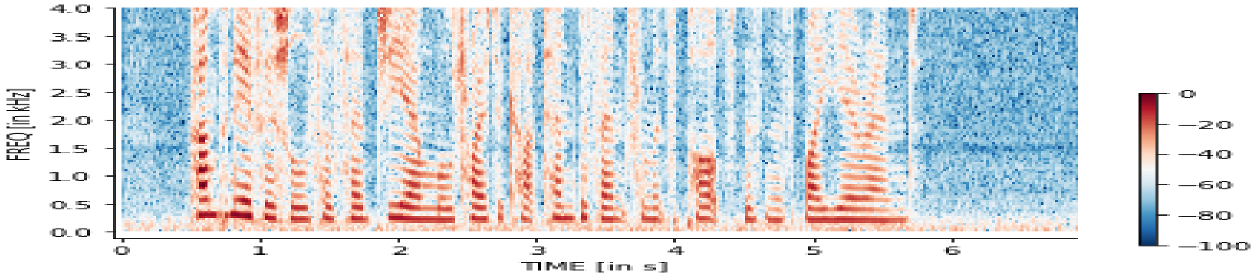

Figure 4.6: Spectrogram of an audio after being applied Complex Mask estimation GEV Beamforming

Table 4.2: Word Error Rate (%) for REVERB Challenge dataset using various beamforming methods.

Training	Dev			Eval		
	Real	Simu	Avg	Real	Simu	Avg
BeamformIt [10]	19.7	6.2	12.9	22.2	6.5	14.4
3-D CNN [22]	20.4	6.7	13.5	21.2	6.6	13.9
Unsup. MVDR	17.2	**5.1**	11.2	14.9	5.6	10.3
Unsup. GEV	**15.6**	5.6	**10.6**	**13.5**	**5.3**	**9.4**
Sup. oracle MVDR [17]	17.5	5.2	11.3	13.0	5.3	9.2
Sup. oracle GEV [15]	17.0	5.6	11.3	13.0	5.3	9.2
Complex Transf. Unsup MVDR [17]	19.79	6.5	13.14	17.83	6.1	11.96
Complex Transf. Unsup GEV [15]	18.6	6.1	12.35	16.2	6.2	11.2

based beamforming on the basis of WER is performing not as good as our baseline that is unsupervised mask estimation based beamforming. But when we have listened those audio we found to perceive the audio beamformered by Complex Transformer Based mask estimation to be better and that's why compared both the audio with some other parameters also like PESQ(Perceptual Evaluation of Speech Quality), SRMR(Speech-to-Reverberation modulation energy ratio) and MUSHRA(MUltiple Stimuli with Hidden Reference and Anchor) AB Tests.

Table 4.3: PESQ Score for REVERB Challenge dataset using various beamforming methods.

Training	Dev	Eval
BeamformIt [10]	3.07	3.075
MCLP based Beamforming [4]	3.32	3.3
Unsup. MVDR	3.05	3
Unsup. GEV	2.96	2.91
Complex Transf. Unsup MVDR	**3.12**	**3.1**
Complex Transf. Unsup GEV	**3.12**	**3.13**

Table 4.4: SRMR Scores for REVERB Challenge dataset using various beamforming methods.

Training	Dev			Eval		
	Real	Simu	Avg	Real	Simu	Avg
BeamformIt [10]	5	3.77	4.38	5.07	2.73	3.90
MCLP based Beamforming [4]	6.12	4.74	5.43	5.31	4.88	5.09
Unsup. MVDR	6.4	4.65	5.525	5.41	4.9	5.15
Unsup. GEV	5.68	4.46	5.07	5.02	4.5	4.76
Complex Transf. Unsup MVDR	**6.6**	**4.83**	**5.71**	**5.6**	**4.94**	**5.27**
Complex Transf. Unsup GEV	**6.45**	**4.66**	**5.55**	**5.4**	**4.89**	**5.27**

Where Perceptual Evaluation of Speech Quality (PESQ) is a family of standards comprising a test methodology for automated assessment of the speech quality as experienced by a user of a telephony system. It is standardized as ITU-T recommendation P.862 (02/01). In the table 4.3, we have compared the audio beam formed by our proposed approach that is Complex Transformer based beamforming to the unsupervised mask estimation based beamforming, and other commonly used beam formed method. In the table 4.3, the MCLP based beamforming [4] was the highest PESQ score. Also our proposed method improves the PESQ score over the unsupervised neural mask estimation based beamforming.

To solidify our statement we have also performer SRMR evaluation on our data which have been beamformed by the proposed method and there also we have seen consistency of results as seen in PESQ evaluation. Compared to beamformit both our method gave better results in terms of SRMR metric and if we compare results between the unsupervised mask estimation based beamforming and Complex transformer based beamforming, the later performed better.

Table 4.5: MUSHRA Scores for REVERB Challenge dataset(in percent)

Training	Dev		Eval	
	REAL	SIMU	REAL	SIMU
Complex Transf. Unsup GEV(A audio)	67.18	93.75	81.25	79.68
Unsup. GEV(B audio)	32.81	6.25	18.75	20.31

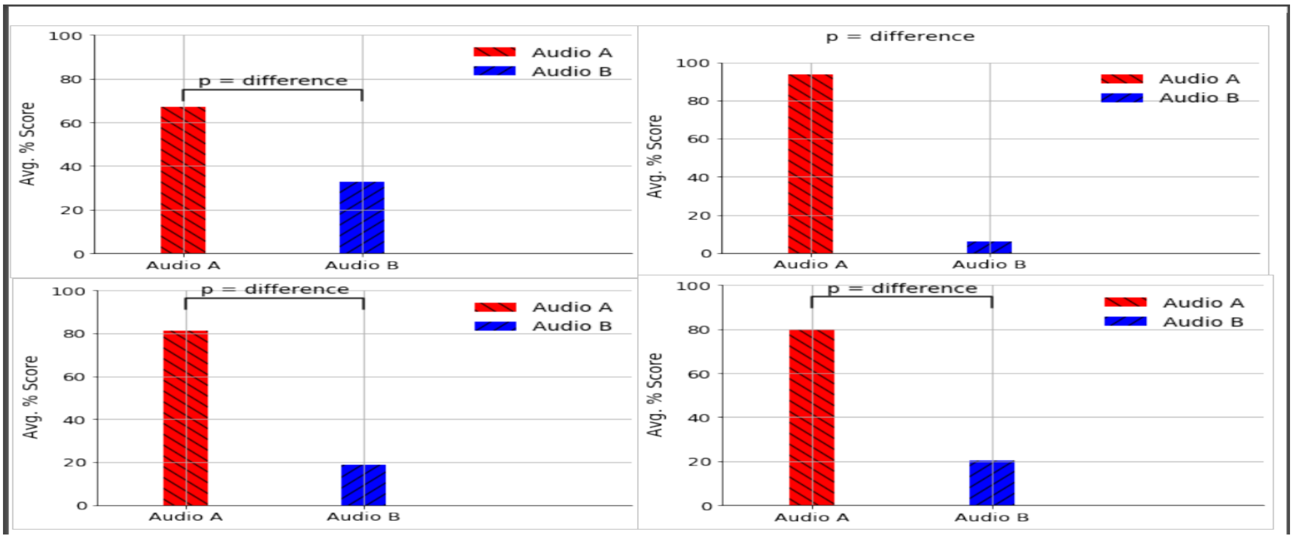

Figure 4.7: These are REVERB-14 MUSHRA AB test percentage graph where red bar indicate audio A(Complex architecture beamformed audio) and blue bar indicate Audio B(Unsupervised Baselin model) (a)topmost left-dev real,(b)topmost right-dev simu,(c)bottom most left-eval real,(d)bottom right-eval simu

4.3.5 MUSHRA(MUltiple Stimuli with Hidden) AB Test

In this section we report the results of subjective evaluation done on our audios beam formed by Unsupervised Mask Estimation method v/s Complex Mask Estimation Method. For that we have employed/conducted the MUSHRA((MUltiple Stimuli with Hidden) AB Test that was published by ITU-R Recommendations BS.1534-2 [29]. In the MUSHRA AB test, the test subject is asked to listen two audios and he/she has to select that audio which he/she considers to be comparatively cleaner than the other one. In setting up the experiment, we have hosted the site using the google cloud platform.

[1]

Now if we talk about the analysis of the results in this experiments **8** people had participated, and they were asked to perform a listening test of around 10 minutes where they were asked

[1] for the link of the experiment : http://rohit-ab.el.r.appspot.com/

Table 4.6: MUSHRA Scores for CHiME-3 Challenge dataset(in percent)

Training	Dev		Eval	
	REAL	SIMU	REAL	SIMU
Complex Transf. Unsup GEV(A audio)	40.62	4.68	26.56	10.93
Unsup. GEV(B audio)	59.37	95.31	73.43	89.06

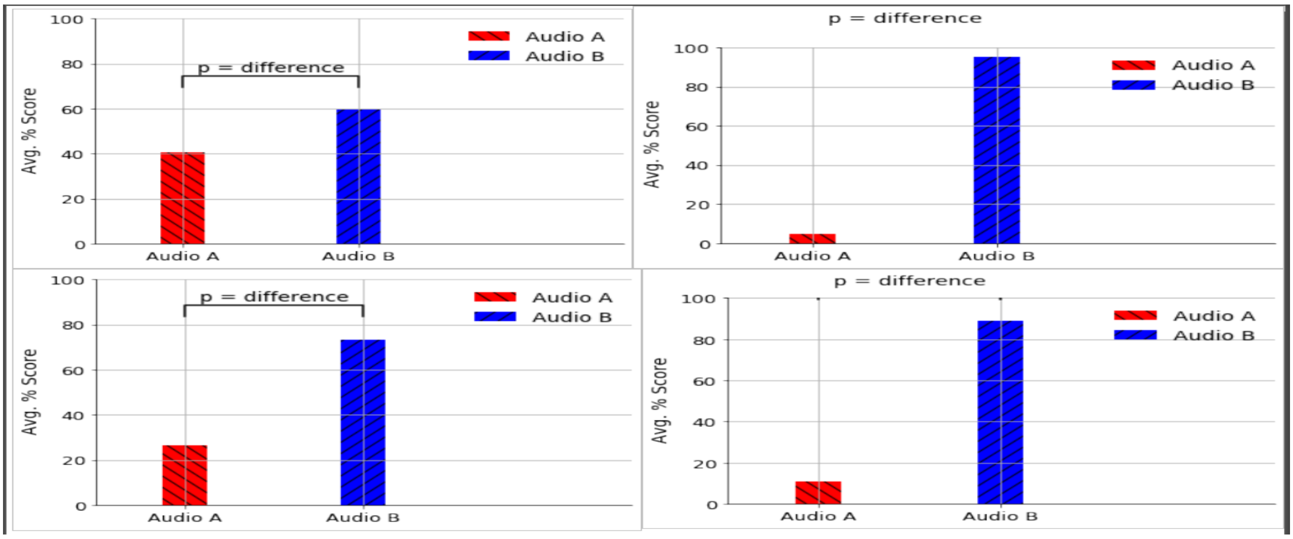

Figure 4.8: These are CHiME-3 MUSHRA AB test percentage graph where red bar indicate audio A(Complex architecture beamformed audio) and blue bar indicate Audio B(Unsupervised Baselin model)(a)topmost left-dev real,(b)topmost right-dev simu,(c)bottom most left-eval real,(d)bottom right-eval simu

to listen to the audios from both CHiME-3 dataset and REVERB-14 dataset. The experiment was designed such that a listener will be asked to listen every possible condition that exist in both dataset. Also this point is kept in mind that in stimuli there will be no skewness/biasness present towards one gender.

The results of REVERB-14 and CHiME-3 challenge result are given in the Table 4.5 and 4.6 respectively. And in the figure 4.3 and fig 4.4 we can see the percentage representation of the results. In the REVERB-14 dataset, the audio which are beamformed by the mask estimated through complex architecture (Audio A) gave far more better results compare to the mask that has been estimated by the unsupervised mask estimation algorithm (Audio B). However in the CHiME dataset, the result are totally different, in CHiME-3 unsupervised mask estimation is performing far better than the complex mask estimation algorithm. And this actually agrees with the ASR results where the the difference between the ASR results of both the results differ by 8% to 9%.

Chapter 5

Summary and Future Extension

5.0.1 Summary

The following are the novel contributions from the current work,

- The conventional state of the art neural mask estimator, as discussed require clean target audio to train the model, In chapter 4 we have proposed an unsupervised mask estimation approach for the GEV beamforming.

- The mask estimation is based on the joint estimation of the late reverberation component and a spatial filter that performs the beamforming to identify the clean source signal. This estimation is based on a maximum likelihood framework in a multi-channel linear prediction setting.

- The estimate of the clean source signal is used in the neural mask estimator to generate the speech presence probability which is in turn used in the generalized eigen value beamforming.

- For showing the effectiveness of our proposed approach, we have performed several ASR experiments on the CHiME-3 and the Reverb Challenge datasets. . This confirms that the proposed approach of unsupervised mask estimation achieves performance similar to the supervised oracle mask estimation using paired clean and noisy audio recordings.

- In the chapter 5, we have proposed a novel method to generate the unsupervised mask by considering the complex data. Several experiments are performed to show the effectiveness of the proposed approach.

- We have also conducted one listening test, to show that beamforming done by considering the method propsed in chapter 5, will give is better quality audio.

5.0.2 Future Work

- Till now we have trained the beamforming model separately and ASR model separately, the main our objective would be now to train both the model jointly, i.e now we will do beamforming keeping the objective of ASR i,e to reduce the WER(word error rate).
- Train the complex mask estimation model with different loss function. Also to tune the hyper parameter of the model like learning rate,or number of hidden layer or changing the number of heads to see if we can get some better quality mask.
- The idea of using Gaussian Mask in Chapter 4 incorporates the relative attention idea in the Transformer architecture. This idea was proposed in [27]. Therefore we can try other relative attention ideas here or can use this same architecture but different distribution like student's t distribution.
- Also till now all the masks are being estimated independently of each other, in the future work this can also be done that to estimate a single mask jointly by seeing all the multi-channel data.

References

[1] Paolo Chiariotti, Milena Martarelli, and Paolo Castellini, "Acoustic beamforming for noise source localization–reviews, methodology and applications," *Mechanical Systems and Signal Processing*, vol. 120, pp. 422–448, 2019.

[2] Ernst Warsitz and Reinhold Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[3] 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)*. IEEE, 2016, pp. 5210–5214.

[4] Srikanth Raj Chetupalli and Thippur V Sreenivas, "Joint spatial filter and time-varying mclp for dereverberation and interference suppression of a dynamic/static speech source," *arXiv preprint arXiv:1910.09782*, 2019.

[5] Efthymios Tzinis, Shrikant Venkataramani, and Paris Smaragdis, "Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 81–85.

[6] Lukas Drude, Daniel Hasenklever, and Reinhold Haeb-Umbach, "Unsupervised training of a deep clustering model for multichannel blind source separation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 695–699.

[7] Barry D Van Veen and Kevin M Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP magazine*, vol. 5, no. 2, pp. 4–24, 1988.

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

[9] John Billingsley and R Kinns, "The acoustic telescope," *Journal of Sound and Vibration*, vol. 48, no. 4, pp. 485–510, 1976.

[10] Xavier Anguera, Chuck Wooters, and Javier Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[11] Michael I Mandel, Daniel P Ellis, and Tony Jebara, "An em algorithm for localizing multiple sound sources in reverberant environments," in *Advances in neural information processing systems*, 2007, pp. 953–960.

[12] Shoko Araki, Tomohiro Nakatani, Hiroshi Sawada, and Shoji Makino, "Blind sparse source separation for unknown number of sources using gaussian mixture model fitting with dirichlet prior," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 33–36.

[13] Hiroshi Sawada, Shoko Araki, and Shoji Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2010.

[14] Nobutaka Ito, Shako Araki, Takuya Yoshioka, and Tomohiro Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2014, pp. 268–272.

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

[16] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[17] 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)*. IEEE, 2017, pp. 286–290.

[18] Srikanth Raj Chetupalli and Thippur V Sreenivas, "Late Reverberation Cancellation Using Bayesian Estimation of Multi-Channel Linear Predictors and Student's t-Source Prior," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1007–1018, 2019.

[19] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[20] 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*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[21] 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.

[22] 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)*. IEEE, 2018, pp. 5499–5503.

[23] 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)*. IEEE, 2015, pp. 504–511.

[24] Keisuke Kinoshita et al., "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *2013 IEEE WASPAA*. IEEE, 2013, pp. 1–4.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[26] Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee, "T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6649–6653.

[27] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck, "Music transformer," *arXiv preprint arXiv:1809.04281*, 2018.

[28] Muqiao Yang, Martin Q Ma, Dongyu Li, Yao-Hung Hubert Tsai, and Ruslan Salakhutdinov, "Complex transformer: A framework for modeling complex-valued sequence," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4232–4236.

[29] Gregory Sell and Daniel Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 413–417.