Learning based Approach for Multi-channel Speech Enhancement

Undergraduate Project

Paridhi Maheshwari (14448)
Supervisor: Rajesh M. Hegde

Contents

Introduction

Introduction

- Acoustic beamforming approach for speech enhancement in noisy and reverberant environments.
- Beamformer enhances sound coming from the direction of the steering vector and suppresses other direction sounds. Hence, accurate steering vector estimation is important.
- Using multiple microphones for speech acquisition and enhancement tasks has received considerable attention.
- Inefficiencies of conventional DOA based steering vector estimation
  - Dependent on knowledge other than the signal (eg: array geometry).
  - Rely on the inaccurate plane wave assumption.
  - Fail for reverberant environments.
- Shift to data-driven beamforming: Spectral masks for speech and noise are used to estimate speech covariance matrix, which further provides the beamforming coefficients.

Related Work

Related Work

Model-based Approaches: Signals are parametrized using various generative models [5, 4, 3]. Advantages over these approaches:

- No assumptions about the signal structure.
- All frequencies are processed simultaneously. Hence, reduced computational complexity.

Learning-based Approaches: Training networks to learn spectral masks [2, 1]. They work on a per channel basis and are therefore, indifferent to the array configuration. However, they do not leverage the inter-channel information.

Signal Model

Signal Model

Time domain: clean speech convolved with room impulse response (RIR) followed by AWGN.

In the frequency domain,

$$\mathbf{y}_{f,t,m} = \mathbf{h}_{f,m} \cdot \mathbf{x}_{f,t} + \mathbf{n}_{f,t,m}$$

- $\mathbf{y}_{f,t,m}$ is the STFT of the signal received at the m^{th} microphone.
- $\mathbf{x}_{f,t}$ is the STFT of the clean speech signal.
- $\mathbf{h}_{f,m}$ is the Fourier transform of the RIR of m^{th} microphone.
- $\mathbf{n}_{f,t,m}$ is the STFT of the noise added to the m^{th} microphone.

The complete signal at the (f, t) bin is given by

$$\mathbf{y}_{f,t} = \begin{bmatrix} \mathbf{y}_{f,t,1} & \mathbf{y}_{f,t,2} & \cdots & \mathbf{y}_{f,t,M} \end{bmatrix}^T$$

.

Proposed Algorithm

Feature Extraction

Magnitude v/s Phase: Magnitude variance $<<$ phase variance of $\mathbf{y}_{f,t,m}$ and $\mathbf{h}_{f,m}$. Therefore, only phase information should suffice.

At a given time-frequency bin (f, t), the input features and output labels are of the form

$$\begin{bmatrix} \cos(\angle \mathbf{y}_{f,t,2} - \angle \mathbf{y}_{f,t,1}) \\ \sin(\angle \mathbf{y}_{f,t,2} - \angle \mathbf{y}_{f,t,1}) \\ \vdots \\ \cos(\angle \mathbf{y}_{f,t,M} - \angle \mathbf{y}_{f,t,1}) \\ \sin(\angle \mathbf{y}_{f,t,M} - \angle \mathbf{y}_{f,t,1}) \end{bmatrix} \longrightarrow \begin{bmatrix} \cos(\angle \mathbf{h}_{f,2} - \angle \mathbf{h}_{f,1}) \\ \sin(\angle \mathbf{h}_{f,2} - \angle \mathbf{h}_{f,1}) \\ \vdots \\ \cos(\angle \mathbf{h}_{f,M} - \angle \mathbf{h}_{f,1}) \\ \sin(\angle \mathbf{h}_{f,M} - \angle \mathbf{h}_{f,1}) \end{bmatrix}$$

For a given data sample, the input dimension is $L \times (2M-2) \times (\frac{N}{2}+1)$ where L is the number of time frames, M number of microphones and N number of DFT points. The dimension of the labels is $(2M-2) \times (\frac{N}{2}+1)$.

Learning based RIR Estimation

Aim: Learn a mapping from the 3D input features to the 2D output labels.
Architecture: A recurrent convolutional neural network (RCNN).

1. CNN layers to extract local shift invariant properties of the input spectrogram. Temporal information remains unaffected here.
2. RNN layers to exploit the noise characteristics in the temporal space.
3. A dense connected layer to aggregate the extracted information into the output. Since output is of regression, linear activation is used.

Evaluation: Mean Squared Error (MSE) and R^2 Coefficient of determination were used for quantification.

Beamforming

Minimum variance distortionless response (MVDR) beamformer attempts to null signals coming from any direction other than the the steering vector.

If the beamformer is represented by a linear filter $\mathbf{w}_f$ and steering vector by $\mathbf{r}_f$, the optimization problem:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathbf{w}^H \mathbf{S}_{\mathbf{yy}} \mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^H \mathbf{r}_f = 1 \end{aligned}$$

The closed form solution to this problem is given by

$$\mathbf{w}^H = \frac{\mathbf{r}_f^H \mathbf{S}_{\mathbf{yy}}^{-1}}{\mathbf{r}_f^H \mathbf{S}_{\mathbf{yy}}^{-1} \mathbf{r}_f}$$

And, the spectrum of the enhanced signal is given by

$$\hat{\mathbf{s}}_{f,t} = \mathbf{w}_f^H \mathbf{y}_{f,t}$$

Experiments and Results

RIR Estimation for Certain Frequencies

We first experimented on a small frequency range, to validate the proposed architecture.

Setup: The model was trained on $15K$ recording samples, each consisting of 30 time frames and the frequency set $101 - 110$. A 2048 point DFT was computed in all the experimentation. For validation, a different set of $1K$ recordings were used with similar specifications.

The best model yielded:

Mean square error	0.2840
R^2 score	40.21%

RIR Estimation for Certain Frequencies

The model roughly captures the pattern structure. We believe that increasing the number of frequencies will provide more inter-frequency information and the network will be able to learn more intricate patterns as well.

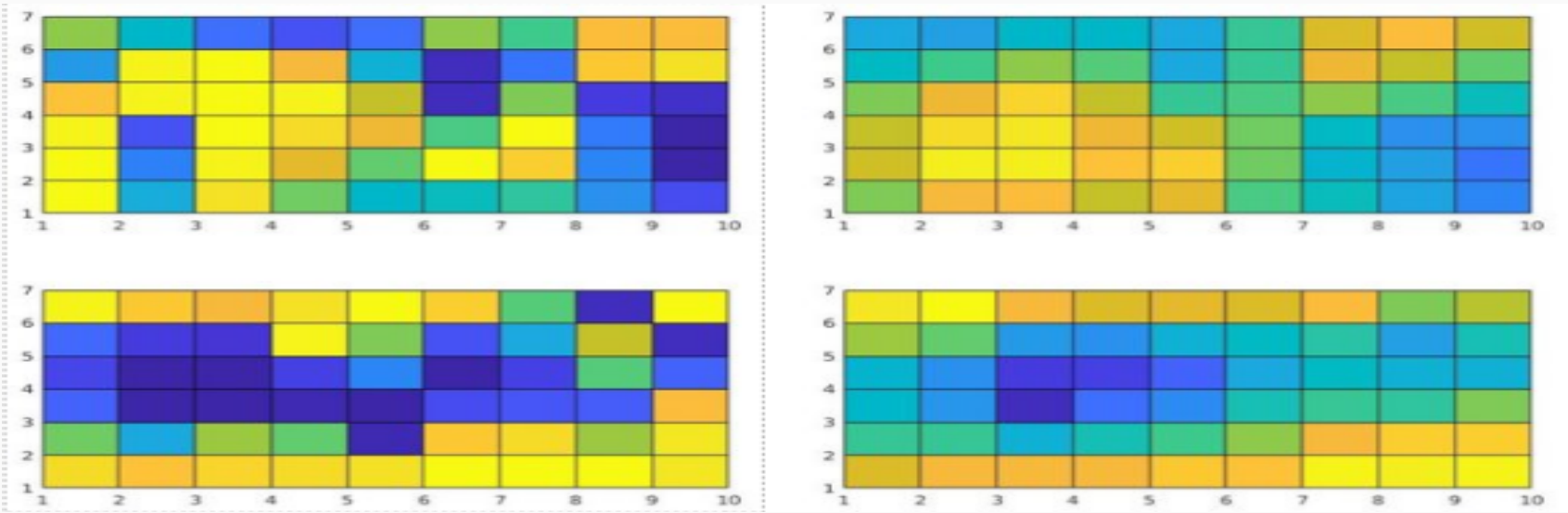

Figure 1: Results of trained model for one testing recording (a) Cosine of ideal labels (b) Cosine of predicted labels (c) Sine of ideal labels (d) Sine of predicted labels

Comparison with Model-based Approach

CGMM Approach: Model the speech and noise STFT signals using a Complex Gaussian Mixture Model [3] and estimate parameters iteratively via Expectation-Minimization (EM) algorithm.

Approach	MSE
CGMM Approach	1.0120
Proposed Approach	0.2840

Table 1: Mean square error of the experimental conditions defined above

Magnitude v/s Phase: CGMM method predicts both magnitude and phase. We observed that the variance of magnitude was 10^{-3} times that of phase. This justifies our hypothesis of neglecting the magnitude.

Extension to Speech Enhancement

Beamforming applied to CGMM Approach

Signal	STOI Score	OPS Score
Microphone Recording	0.6226	22.8711
`DOA_MVDR` Output	0.6452	19.8771
`DOA_TimeDelay` Output	0.6784	24.2194
`CGMM` Output	0.7064	29.5826

Table 2: STOI and OPS Scores for original recordings, and signals enhanced using conventional and CGMM techniques

Extension to Speech Enhancement

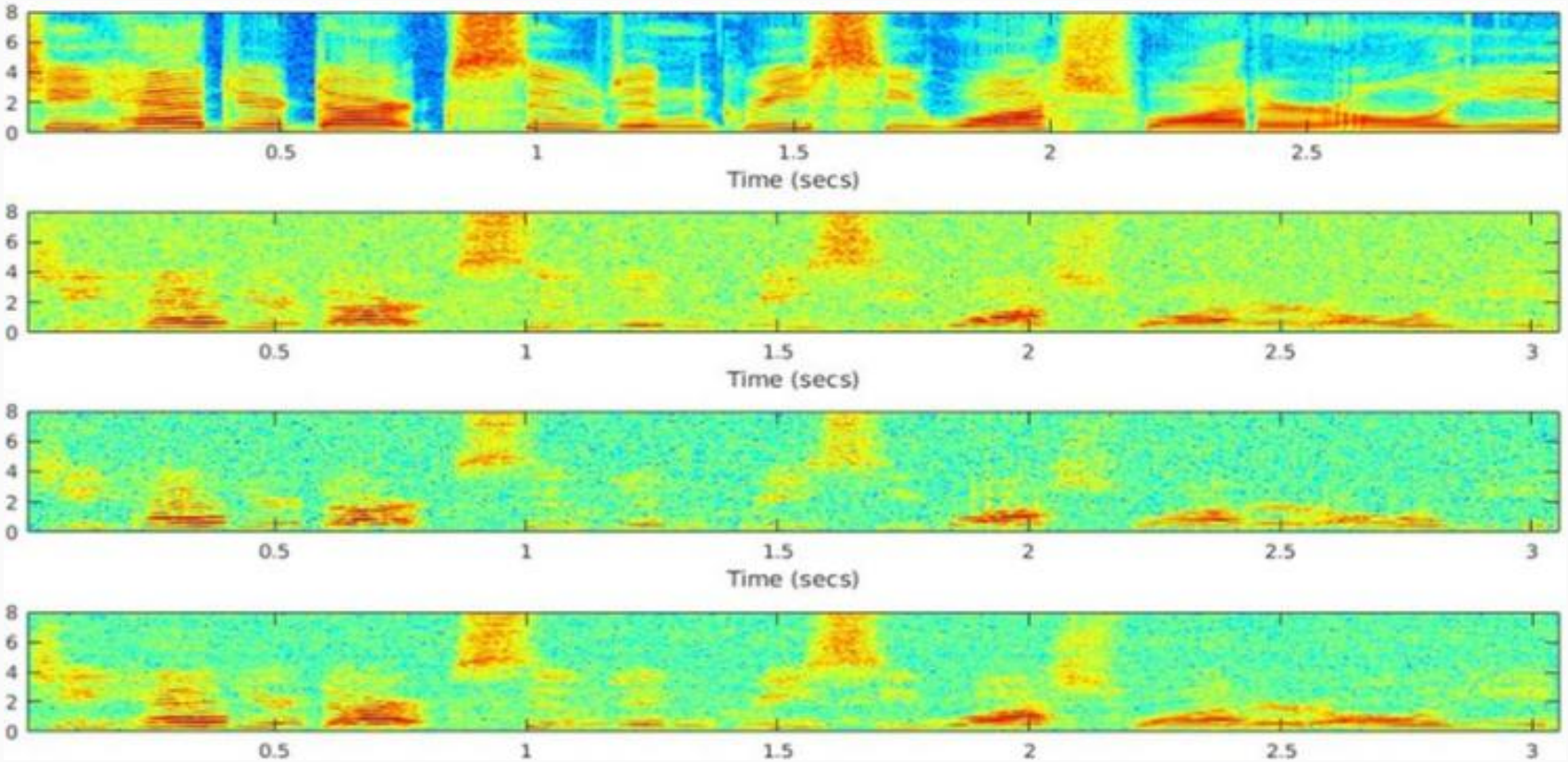

Figure 2: (a) Clean speech signal (b) Recorded signal (c) Enhanced signal using conventional DOA algorithm (d) Enhanced signal using proposed CGMM approach

Thank you.

Questions?

References i

H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux.
Improved mvdr beamforming using single-channel mask prediction networks.
In *INTERSPEECH*, pages 1981–1985, 2016.

J. Heymann, L. Drude, and R. Haeb-Umbach.
Neural network based spectral mask estimation for acoustic beamforming.
In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 196–200. IEEE, 2016.

References ii

T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani.
Robust mvdr beamforming using time-frequency masks for online/offline asr in noise.
In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5210–5214. IEEE, 2016.

M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada.
A multichannel mmse-based framework for speech source separation and noise reduction.
IEEE Transactions on Audio, Speech, and Language Processing, 21(9):1913–1928, 2013.

References iii


D. H. T. Vu and R. Haeb-Umbach.
Blind speech separation employing directional statistics in an expectation maximization framework.
In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 241–244. IEEE, 2010.