CRNN-Based Φ and Θ Estimation with Ambisonics Intensity Vector

...

B.Anshuman [200259]
Avinash Chowdary Pamulapati [202044409]

Overview

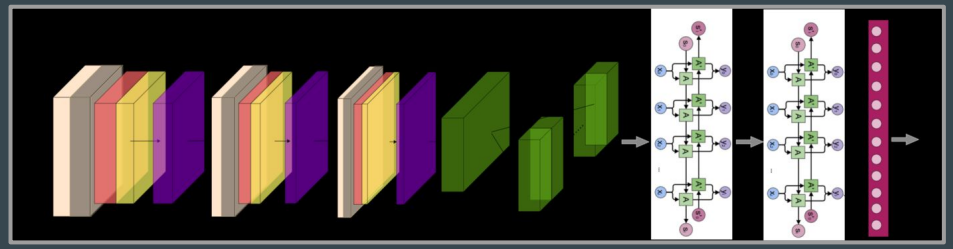
Direction of Arrival estimation has been performed since decades using various techniques including cross-correlation, time delay estimation, and sound intensity vectors. The Paper aimed at using sound intensity vector obtained from a single sound source through First Order Ambisonics in a more apt input format for CRNNs, giving robust results.

Detailed View
Into
the ~~Solution~~ Problem

Direction of Arrival Estimation

Controlling the direction of sound source perceived by an observer is a key factor in simulations and automation. Past decades has seen the involvement of several methods including Delay-and-Sum (DS), Multiple Signal Classifier (MUSIC) subspace, Estimation of Signal Parameters via Rotational Invariance Technique (ESPRIT), and Fourier Transform method (FT-DoA). On the rise of neural networks and their capability to theoretically “learn anything”, it has been applied to this problem statement with significant and robust improvements over methods mentioned before.

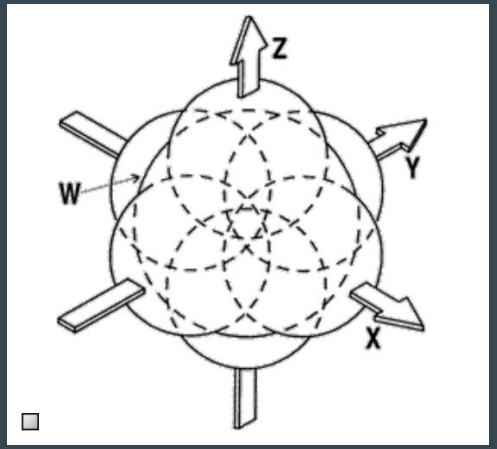
The paper deals with implementing a Convolutional Recurrent Neural Network with non-conventional input representation, namely, using first order ambisonics based 6 component normalized intensity vector for each frame and frequency bins (38 frames x 513 frequency bins). It also approaches the problem as a classification one instead of regression.

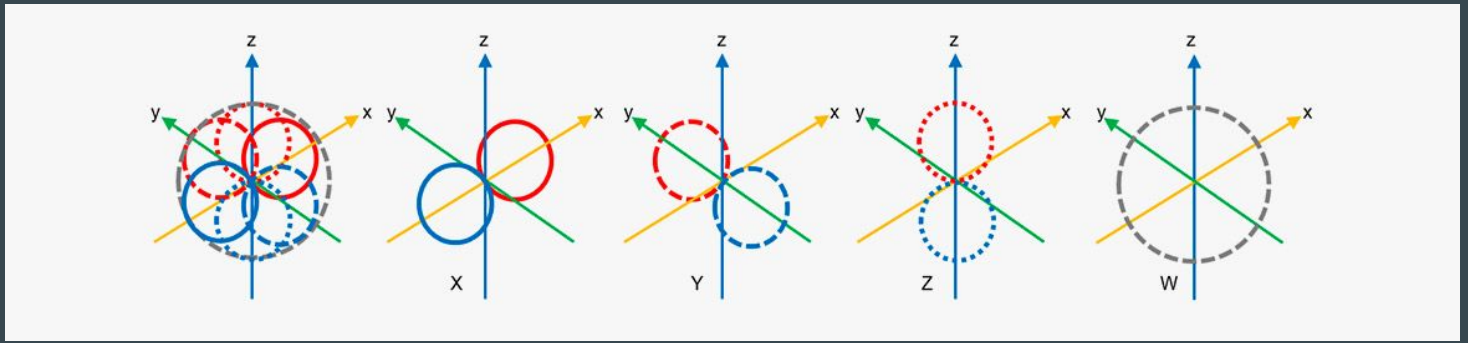


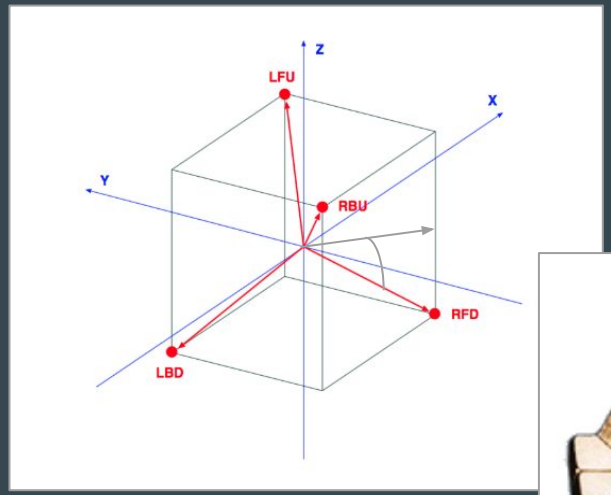
Data Generation and Pre-Processing

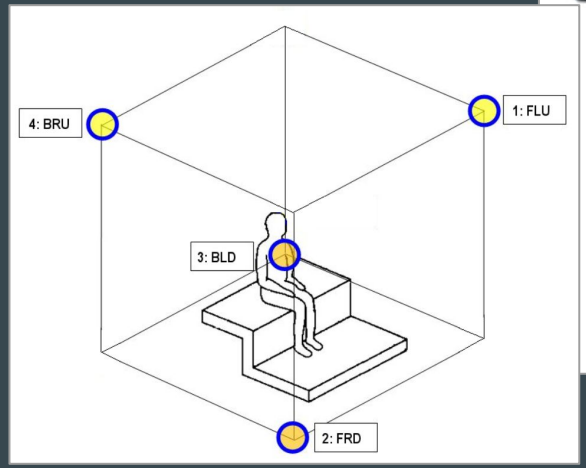
DataLoader : Ambisonics

Ambisonics are representations of spherically received sound into independent components according to spherical harmonic functions. Infinite order Ambisonics would lead to exact description of sound source, but we're using First Order Ambisonics (FOA). It includes three orthogonal and an omnidirectional component, that can approximate the sense of direction with the minimum number of sensors needed.







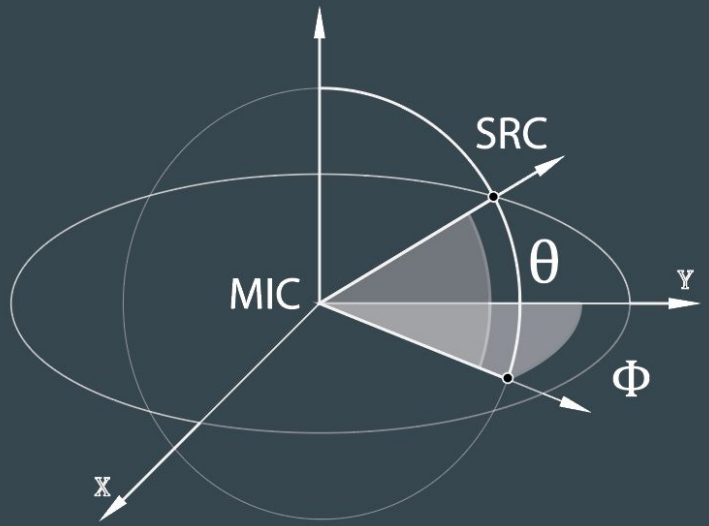


A-Format Microphone

Simulated A-Format Microphone by generating SRIRs with mic orientation according to tetrahedral angles. Each gave a channel audio which are then segregated to B-type format

Algorithm 1 Protocol to generate the SRIRs.

for each DoA_0 **do**
  repeat
    procedure ROOM
      $l = rand(2.5, 10)$
      $L = rand(2.5, 10)$ ▷ in meters
      $h = rand(2, 3)$
      $RT_{60} = rand(0.2, 0.8)$ ▷ in seconds
    end procedure
    procedure MICPOS
      $x_{mic}, y_{mic}, z_{mic} \in room$
      ▷ at least 0.5 m from walls
      $d_{mic-src} = rand(1, 3)$ ▷ in meters
    end procedure
    procedure SRCPOS
      Pick $DoA_{1,2}$
    end procedure
  until a compatible configuration is found
end for



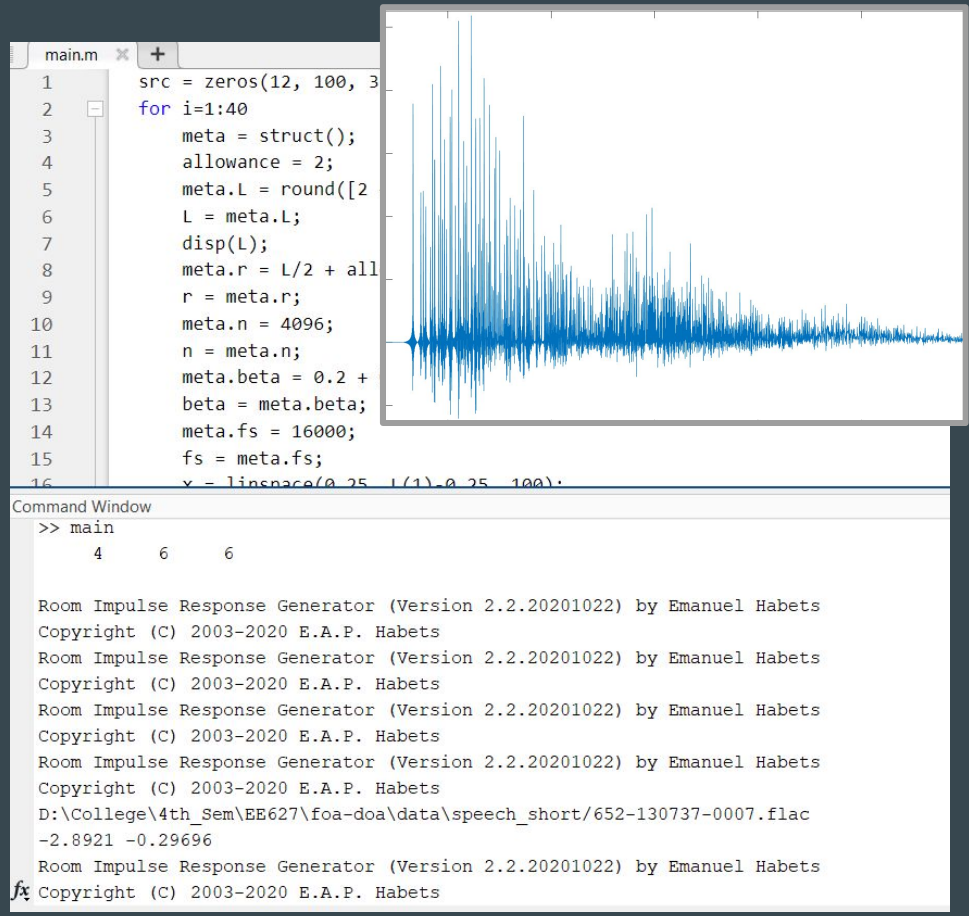
Microphone and Source Positions were generated in nearly equal number of classes for all possible combination of azimuth and elevation angles.

The above algorithm was used with the Image method to generate SRIRs for 4 different orientations of the microphone.

DataLoader : SRIRs

Raw data generation

- Generated 4 Spatial Room Impulse Responses for FOA through the Image Method[1], that considers relative vector of source-receiver, reverberation and dimensions of the enclosure.
- Convolved them with the plain speech obtained from LibreSpeech[2], getting Ambisonics-A format.
- Converted to Ambisonics-B format, finally obtaining 4x16384 per sample.



Dataloader : Intensity Vector

Data Refinement

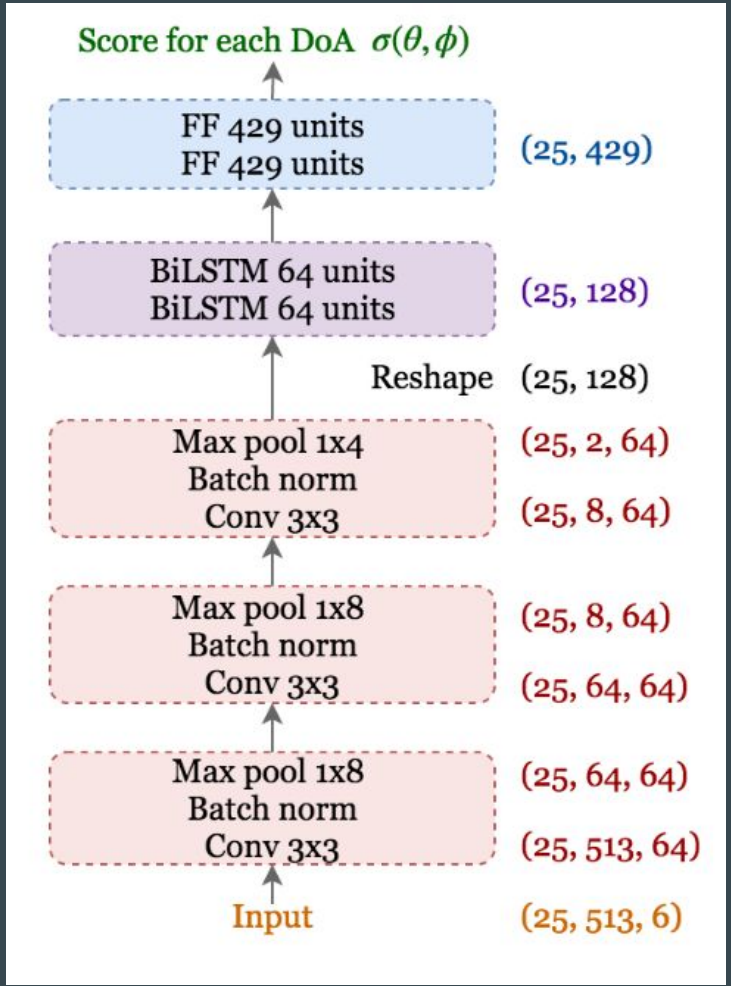
- Took STFT of the 4 channels obtained, giving size of 4x512x38.
- If STFT was plotted on a spectrogram, then each point on the plane would correspond to a vector of 4 channels of imaginary numbers.
- Each of these vectors would be mapped to the intensity vector for that (f, t), and then normalized by total intensity. This is reshaped from 6x512x38 to 38x512x6, essentially an apt form of an image.

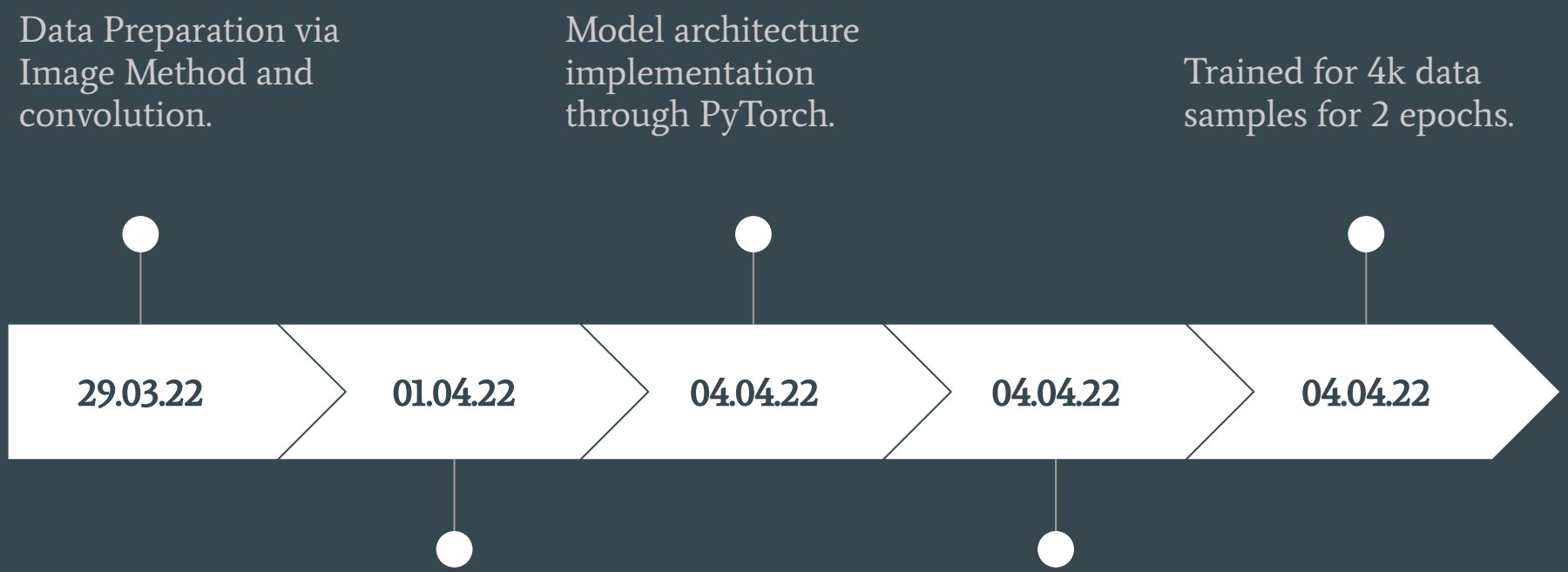
$$\mathbf{I}_{\mathrm{a}}(t,f) = \begin{bmatrix} \mathcal{R}\{W(t,f)^* X(t,f)\} \\ \mathcal{R}\{W(t,f)^* Y(t,f)\} \\ \mathcal{R}\{W(t,f)^* Z(t,f)\} \end{bmatrix}$$

$$\mathbf{I}_{\mathrm{r}}(t,f) = \begin{bmatrix} \mathcal{I}\{W(t,f)^* X(t,f)\} \\ \mathcal{I}\{W(t,f)^* Y(t,f)\} \\ \mathcal{I}\{W(t,f)^* Z(t,f)\} \end{bmatrix}.$$

Pattern-Recognition Model

Convolutional Recurrent Neural Net with the following architecture was used for classification of the FOA audio samples into one of the 684 output classes possible (paper used 429[3]).
The input during implementation is of the order 38x513x6 instead, as 1.024s of speech was used from LibreSpeech[2] and the paper directed to split the waveform into 2 sequences, which was not done due to implementation and understanding reasons.




Data Preparation via Image Method and convolution.
Model architecture implementation through PyTorch.
Trained for 4k data samples for 2 epochs.
29.03.22
01.04.22
04.04.22
04.04.22
04.04.22
Feature extraction in form of ambisonics intensity vector.
5 layers and Cross Entropy Loss with NAdam optimizer.

RESULTS

Comparison

The paper includes scores for multiple sources as :

Room	Simulated SRIR			Real SRIR		
Angular error	<5°	<10°	<15°	<5°	<10°	<15°
Baseline [13]	27.5	56.6	70.2	24.6	55.0	70.7
CRNN + (1)	45.9	85.1	92.7	23.9	66.0	87.0
CRNN + (4) (proposed)	**51.6**	**91.1**	**95.2**	**28.6**	**70.2**	**89.6**

Due to our implementation including a single source for DoA estimation as well as training for 1 epoch only (No GPU used), we obtained an accuracy of 20.9%, i.e. correctly classifying ~20% sources to its correct class out of 684 classes.

```
Num epochs : 2
Number of correct DoA predictions within threshold of 2 neighboring classes : 839
Number of total datasamples : 4000
```

Simplifications Assumed

S1 : DoA of a single source is to be estimated but it can be expanded to multiple sources for the same implementation, only by increasing epochs for increasing accuracy and using Binary Cross Entropy Loss for each classes instead of Cross Entropy as a whole vector.

S2 : Regularization by stacking 2 such architectures and forcing the intermediate output to correspond to match to MUSIC “powermap” [4] was not done, which would have generalised the model, but it was also seen to hinder performance because MUSIC isn’t accurate in noisy or reverberant environment.

S3 : Diffuse babble noise of SNR 0-20dB wasn’t added because SRIR generated already introduced noise in the plain speech used.

References

[1] J. B. Allen and D. A. Berkley “Image method for efficiently simulating small-room acoustics”

[2] LibreSpeech Dataset : https://www.openslr.org/12

[3] R. Schmidt “Multiple emitter location signal parameter estimation”

[4] A. Politis, and T. Virtanen, “Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network”

[5] V. Varanasi, R. Serizel, and E. Vincent, “DNN based robust DOA estimation in reverberant, noisy and multi-source environment”