# DEEP CONVOLUTIONAL NEURAL NETWORK FOR BINAURAL VOICE LOCALIZATION

Gregory D. Hunkins
University of Rochester

## Introduction

The task of 3-D source localization is innate for humans and is essential for every-day functionality. Humans accomplish this task with ease using the binaural cues of interaural level difference (ILD) and interaural time difference (ITD) introduced by the geometry of our head and torso. This ability in machines would allow for advances in robotics, speech enhancement, and other fields due to increased environmental awareness.

## Abstract

A Deep Convolutional Neural Network (DCNN) classification system was designed for the **task of source localization of human voices in 3-D space**. A new dataset, **VoiceBin100K**, is introduced to accomplish this task and for future work in the field. The CNN inputs binaural short-time Fourier Transform (STFT) magnitude and phase features and predicts the location of the speaker's voice according to **168 location classes**. An obtained accuracy score of **99.53%** indicates that this methodology is a highly viable option for this task.

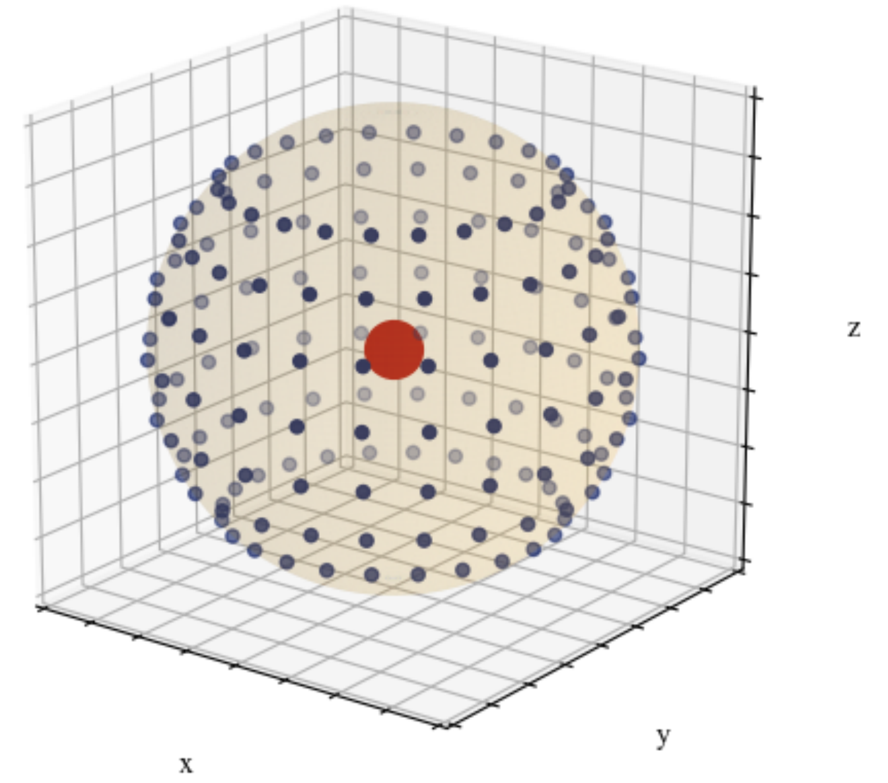


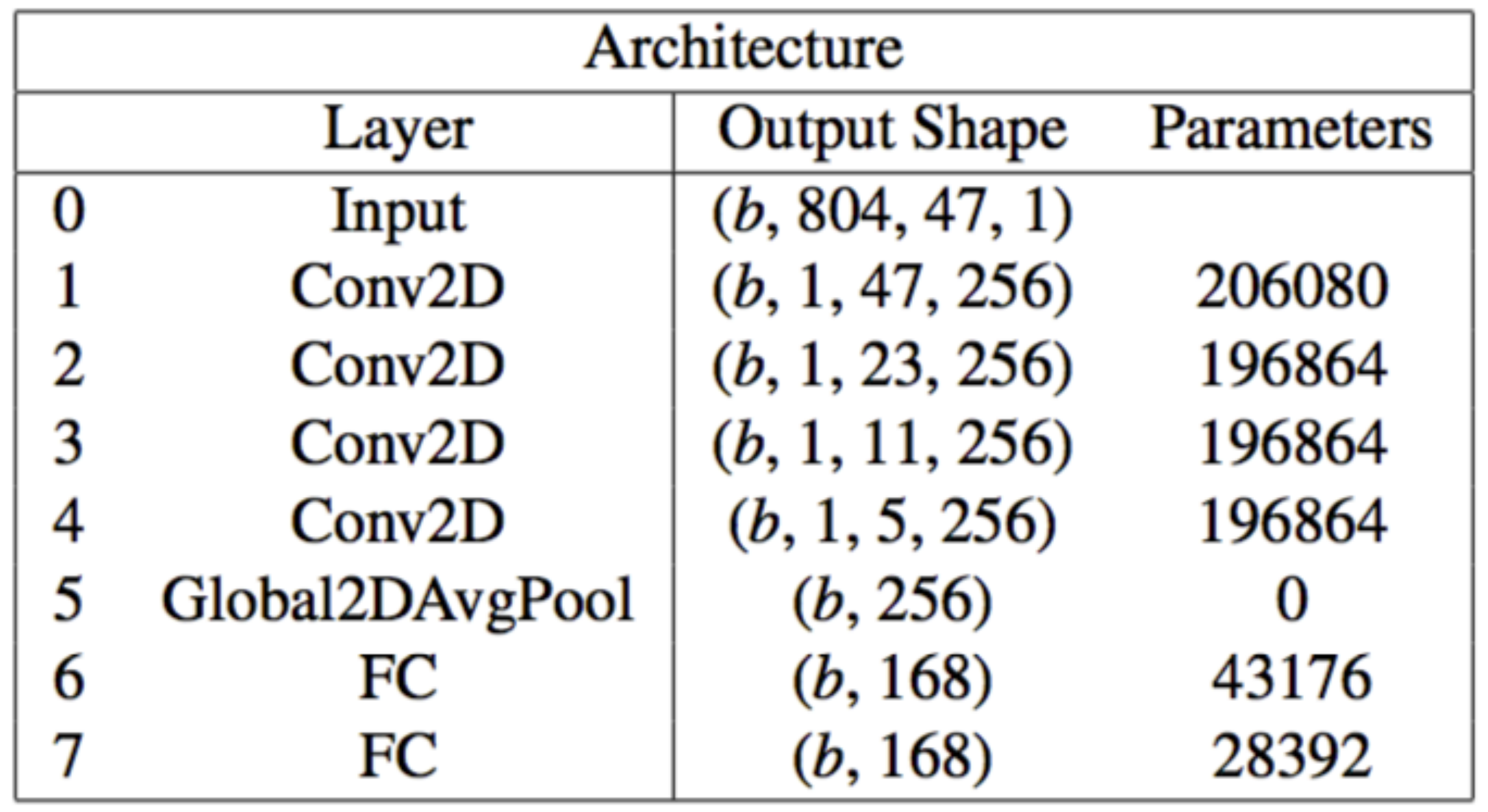
**Figure 1:** 3-D Visualization of 168 Location Classes

## Dataset – VoiceBin100K

- Voice Dataset: TIMIT
- HRIR Dataset: LISTEN
- Noise Dataset: Non-stationary Noise
- Train Set
  - 1825 Distinct Speakers
  - 2200+ instances of each class
- Test Set
  - 518 Distinct Speakers
  - 480+ instances of each class

## Architecture

- Input Feature Vector: Left and Right Channel Magnitude and Phase STFT Features (Figure 2)
- Input Size: (b, 804, 47, 1)



| Architecture | | | |
|---|---|---|---|
| | Layer | Output Shape | Parameters |
| 0 | Input | $(b, 804, 47, 1)$ | |
| 1 | Conv2D | $(b, 1, 47, 256)$ | 206080 |
| 2 | Conv2D | $(b, 1, 23, 256)$ | 196864 |
| 3 | Conv2D | $(b, 1, 11, 256)$ | 196864 |
| 4 | Conv2D | $(b, 1, 5, 256)$ | 196864 |
| 5 | Global2DAvgPool | $(b, 256)$ | 0 |
| 6 | FC | $(b, 168)$ | 43176 |
| 7 | FC | $(b, 168)$ | 28392 |

**Table 1**. Architecture Summary

## Experiments

- Speaker Invariance
- Log vs. Linear Spectrogram
- Small Training Set Size
- Noise Robustness
- HRIR Invariance (Results forthcoming)

## Results

- Linear STFT: 95.53% (Figure 3)
- **Log STFT: 99.53%**
- 20 dB Noise: 97.29%
- 10 dB Noise: 92.38%
- **0 dB Noise: 90.92%**

## Conclusion

From the obtained results, it is clear that this architecture accomplishes the task of binaural voice source localization even under challenging non-stationary noise conditions. Speaker and noise invariance is displayed, and final classification accuracies of **99.53%** under ideal conditions and **90.92%** under 0 dB noise are achieved.

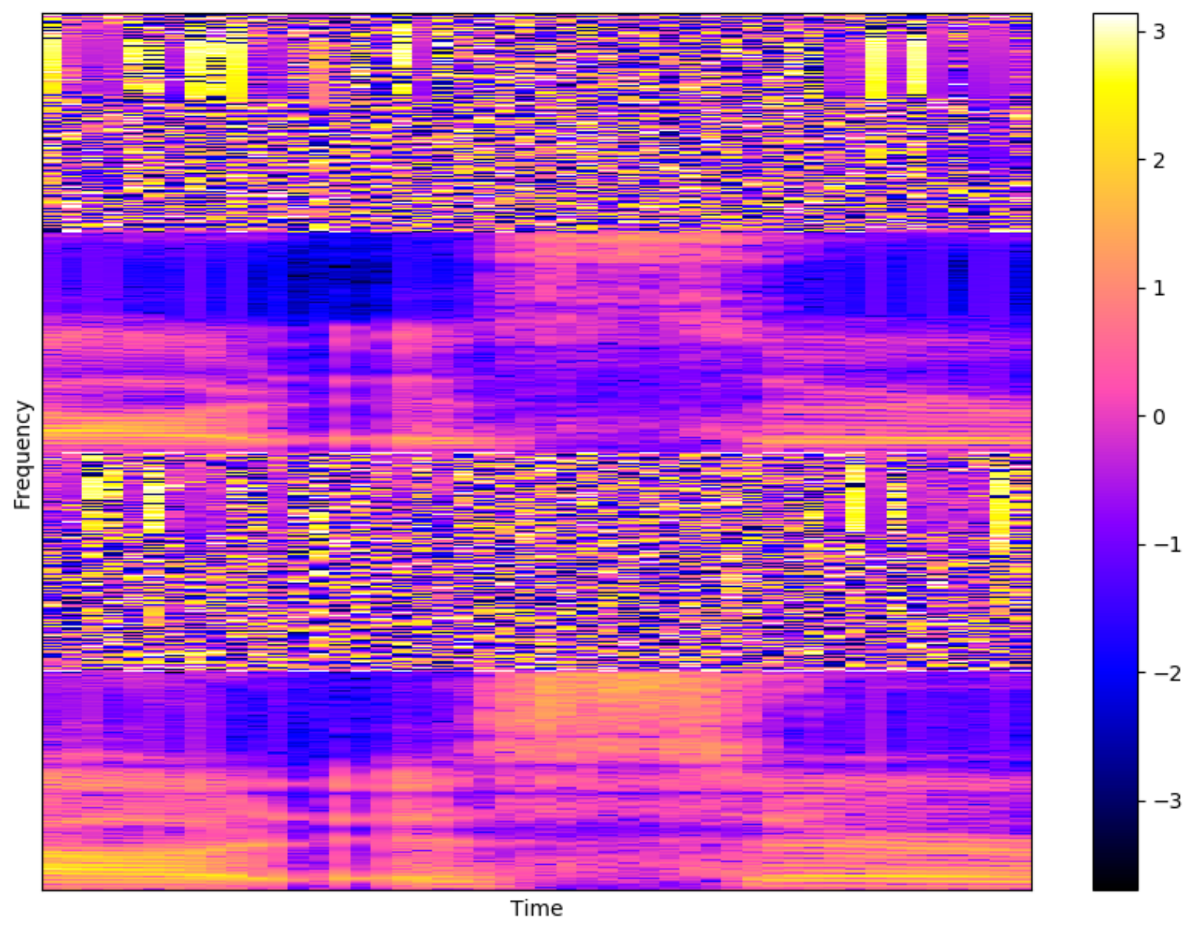


**Figure 2:** Input Feature Representation of Phase and Magnitude Spectrograms

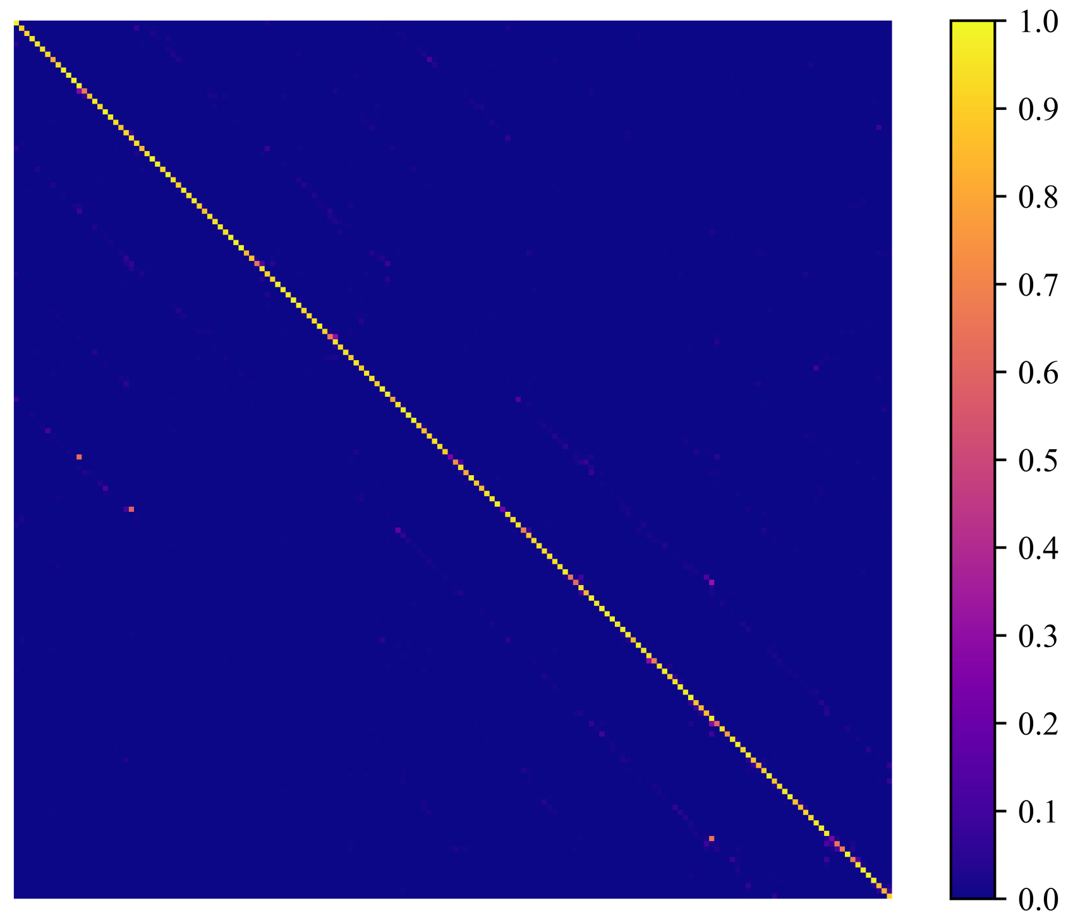


**Figure 3:** Confusion Matrix