EE627

CRNN-Based Joint Azimuth and Elevation Localization with Ambisonics Intensity Vector

PAPER REPORT

Problem Statement

Estimating the Direction of Arrival (DoA) of audio sources is key to many applications, in particular for source separation and speech recognition. The paper presents a source localization system through First Order Ambisonics (FOA) format of recordings, which undergoes preprocessing procedures that's then fed to the pattern recognition model based on a Convolutional and Recurrent Neural Network. When dealing with the spatial properties of a sound field, the Ambisonics format is particularly well suited. This format is based on the decomposition of the sound field based on Spherical Harmonic Functions. It is isotropic and enables easy spatial manipulation of the signal. First Order Ambisonic recording involves the use of special four-capsule microphones; these are arranged in a tetrahedral array to maintain face coherence and directional markers based on the pickup pattern and angle of the capsules in relation to each other. From this we get {W, X, Y, Z} axes. By taking the individual output of each capsule on the microphone we get A Format recording which is later converted to B Format in post production. This is further processed through STFT and intensity vector as shown below is extracted for every frequency-time bin, giving us a tensor for the CRNN implemented.

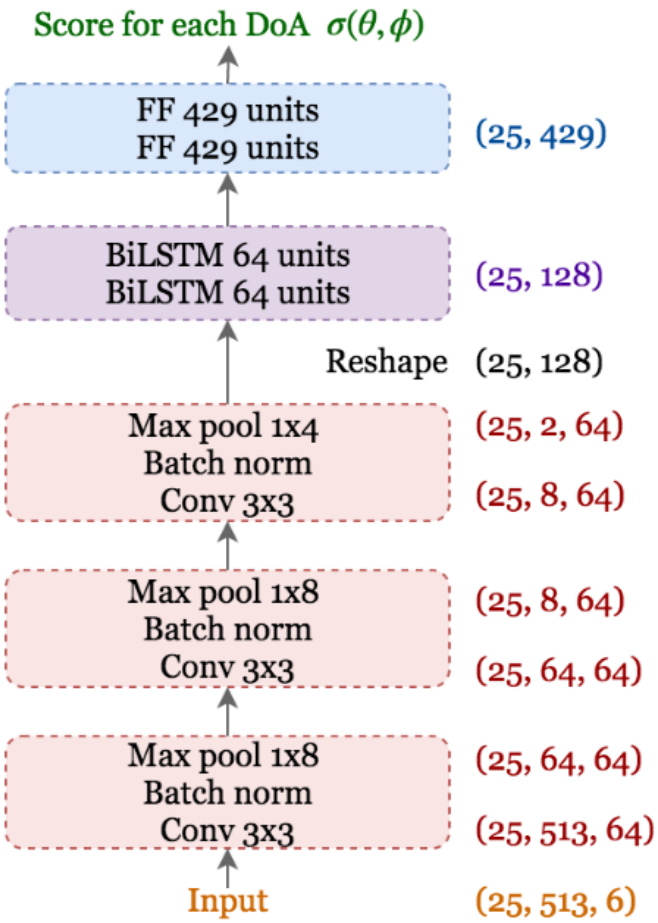
$$I_a(t,f) = \begin{bmatrix} R\{W(t,f)^* \times X(t,f)\} \\ R\{W(t,f)^* \times X(t,f)\} \\ R\{W(t,f)^* \times X(t,f)\} \end{bmatrix} \quad I_r(t,f) = \begin{bmatrix} R\{W(t,f)^* \times X(t,f)\} \\ R\{W(t,f)^* \times X(t,f)\} \\ R\{W(t,f)^* \times X(t,f)\} \end{bmatrix}$$

The DoA of the sound source can be estimated in each (t, f) bin by the opposite direction of the Active Intensity Vector $I_a(t, f)$. The main direction is then recovered by feeding its value in all (t, f) bins to a neural network that will predict the DoA as a classification problem.

PROBLEM FORMULATION

Problem was to be approached by first generating A-format ambisonics data, by generating SRIRs for 4 tetrahedrally oriented microphone channels, convolving each of them from one of the speech files downloaded from LibriSpeech. This was then transformed to B-Format track whose STFT was taken of window size 1024 and half-overlap with hamming window. This was followed by intensity vector calculation as shown above, giving a size of 38x512x6 matrix. This was saved with the meta-data, that was further used to create a label file used for loss calculation. Now for implementing the architecture, I used PyTorch's flexible superclasses and implemented the following :

This was followed by using CrossEntropy Loss and NAdam optimizer, and then trying out a few hyperparameters values, but none gave good results.



RESULTS

Our simplified model achieved an accuracy of 20.9% after just 1 epoch of training, classifying 839 sample files correctly in 4000 files, with 684 total number of classes. With no use of GPU, I was able to train for just 1 epoch, and wasn't able to tune the hyperparameters, which weren't given in the paper, due to time constraints. But the paper itself achieved state of the art results as follows :

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**

CONCLUSION

This paper's key features were making DoA estimation into a classification problem and using intensity vectors which turned out to be a compact and robust representation for such a problem statement. Approaching this paper seemed easy, but implementation was tough, with no prior knowledge of writing DNNs. Simulating and generating data through image method also gives near realistic room impulse responses for each of the tetrahedrally placed oriented microphones.

Finally, the paper used a score metric while having multiple sources to be estimated, which wasn't necessary for our simplification of only one source per audio sample.