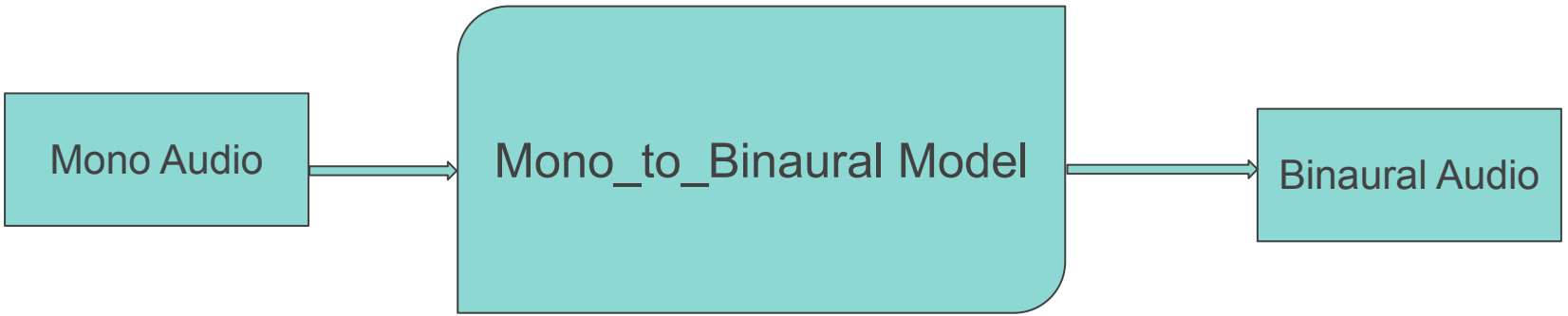
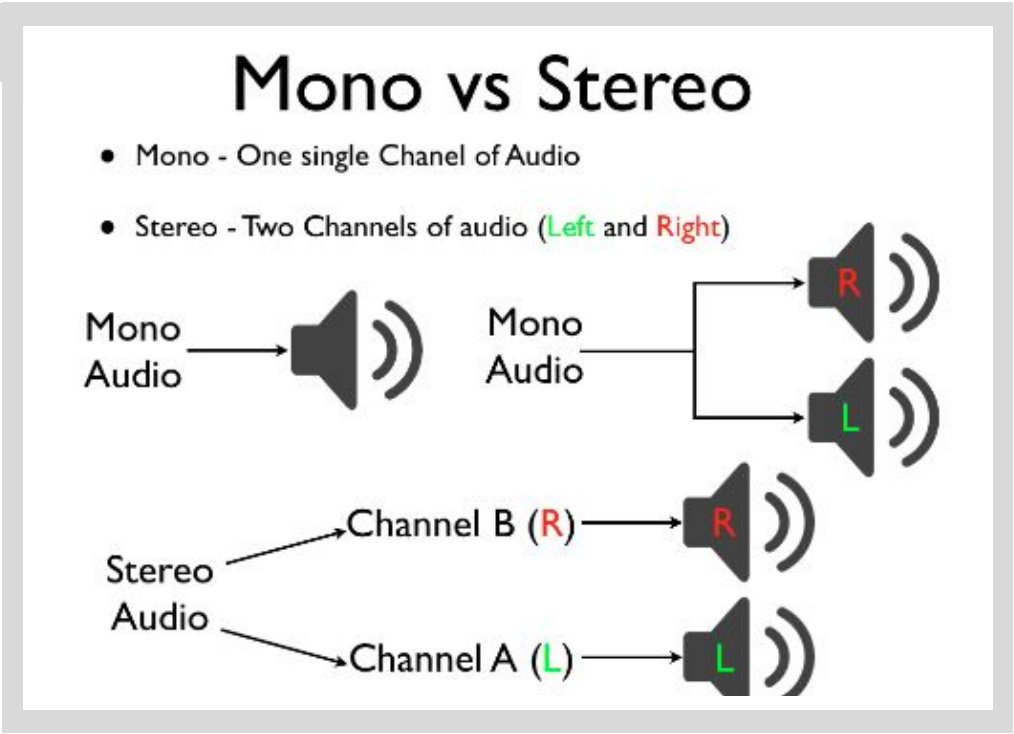
2.5D Visual Sound

Dristanta Das• 24.06.2021

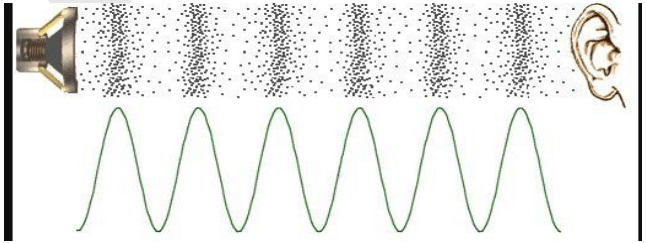
Objective:

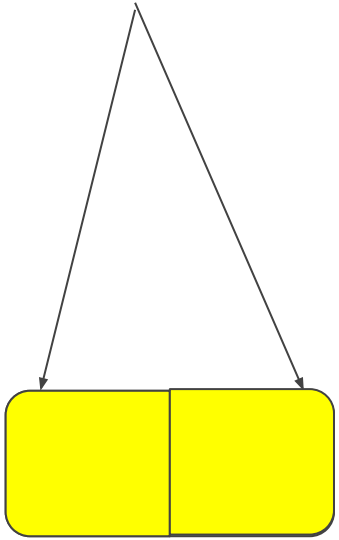


What is Mono And Stereo audio!

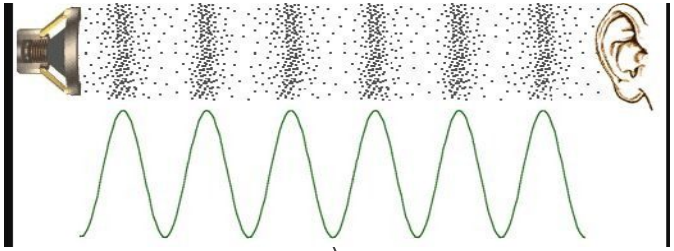


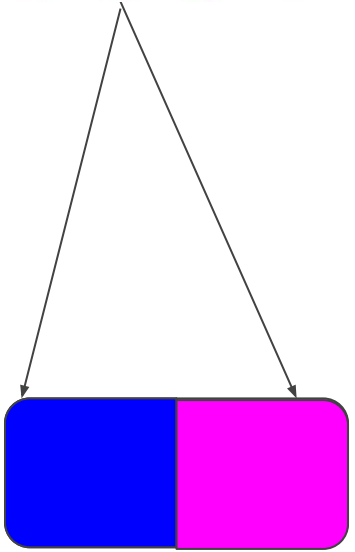
Mono



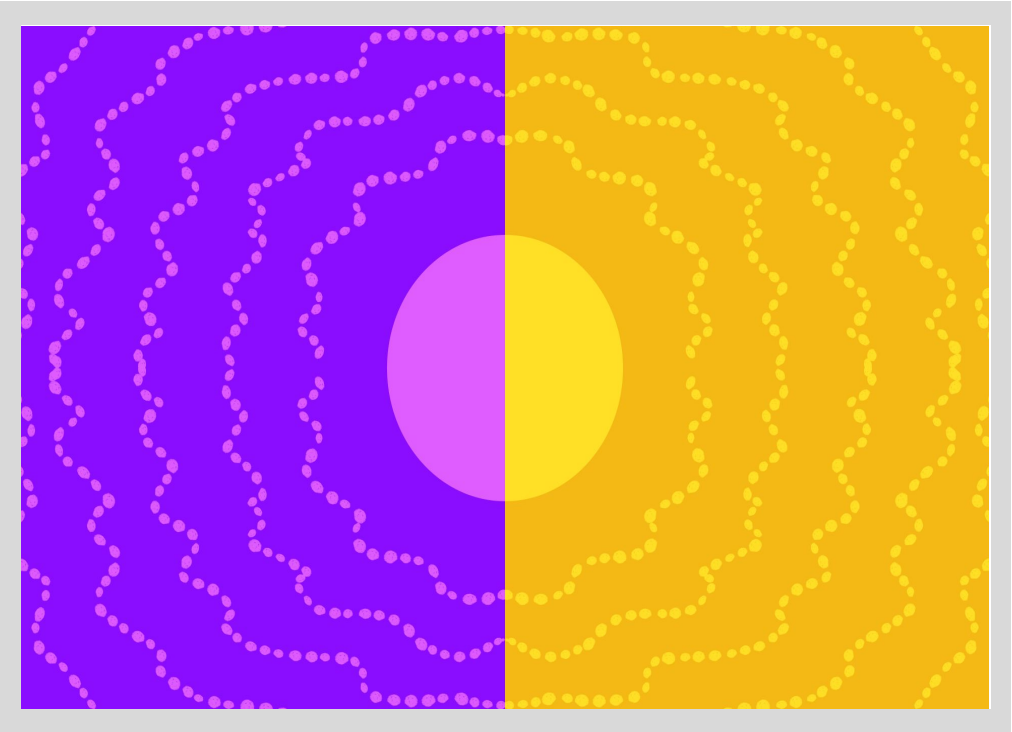


Binaural

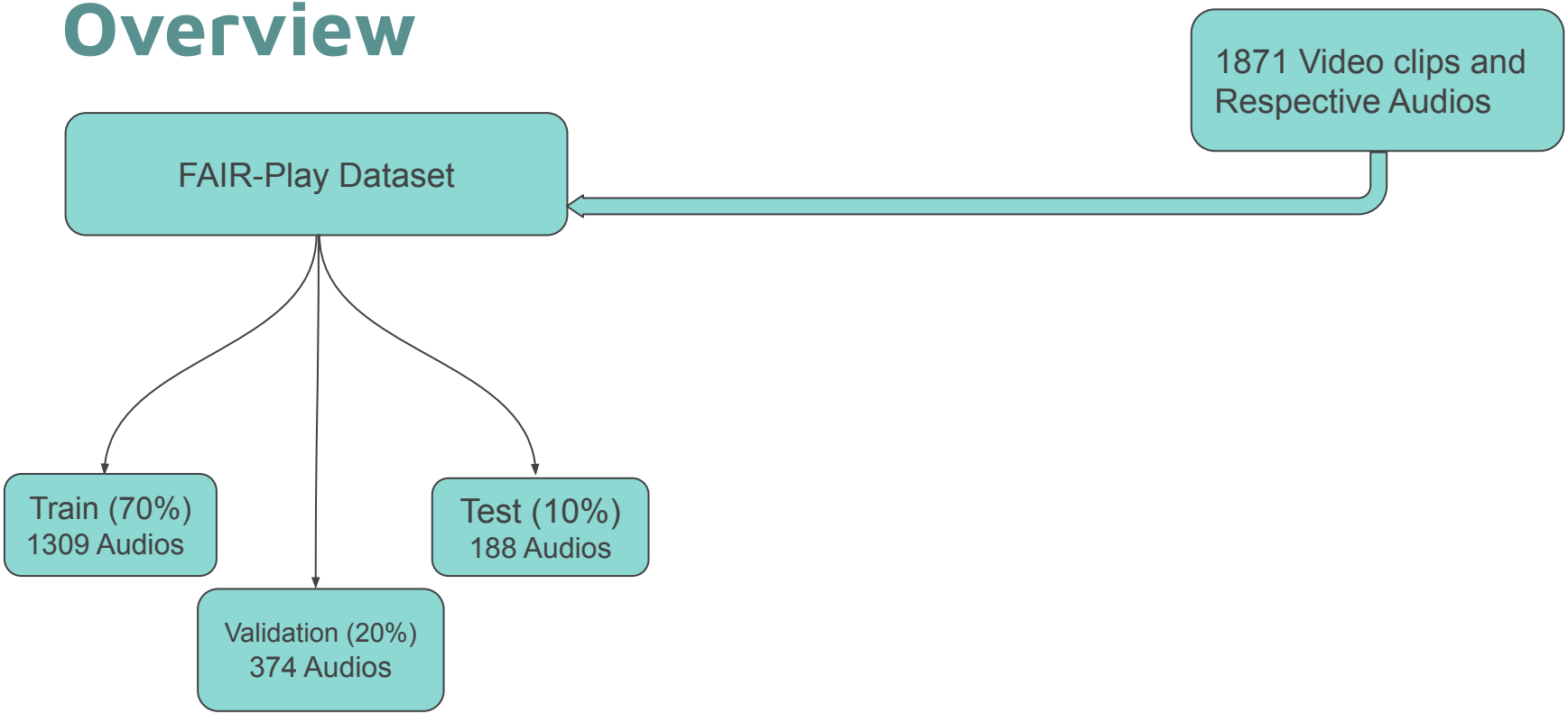




How this helps us to get the spatial knowledge of the audio!



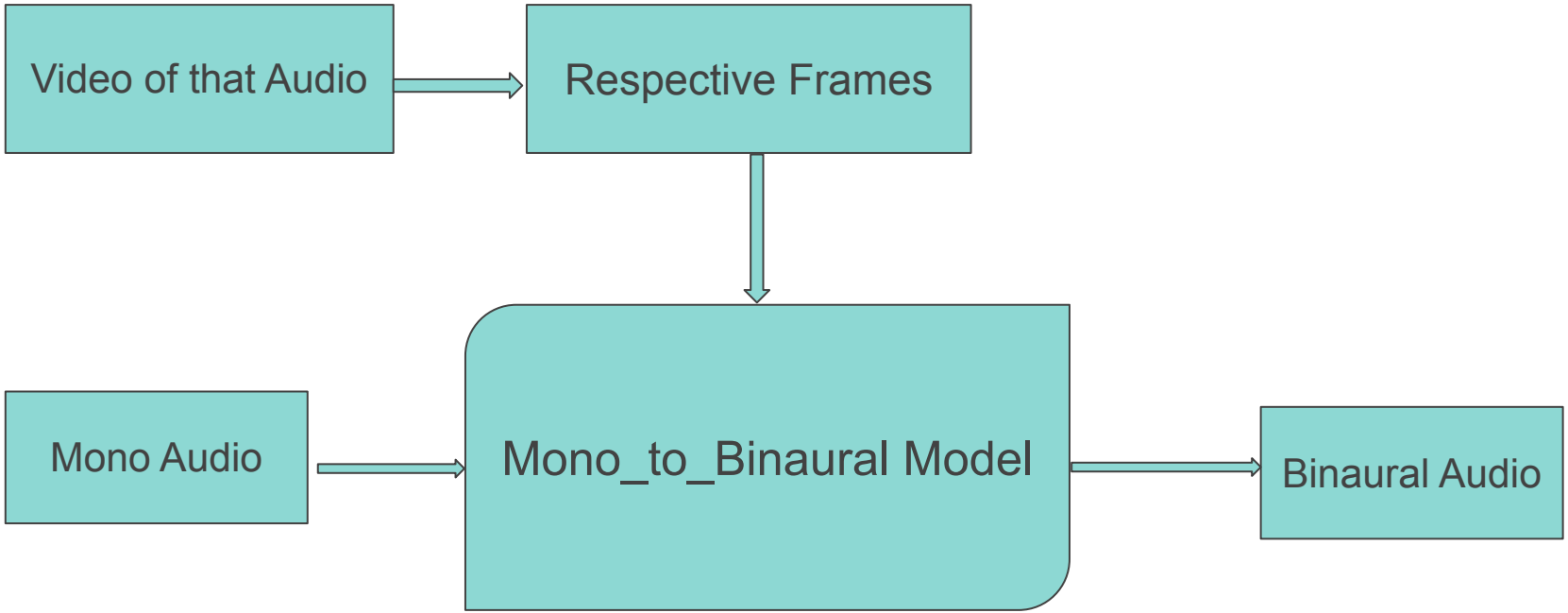
Overview


1871 Video clips and
Respective Audios
FAIR-Play Dataset
Train (70%)
1309 Audios
Validation (20%)
374 Audios
Test (10%)
188 Audios

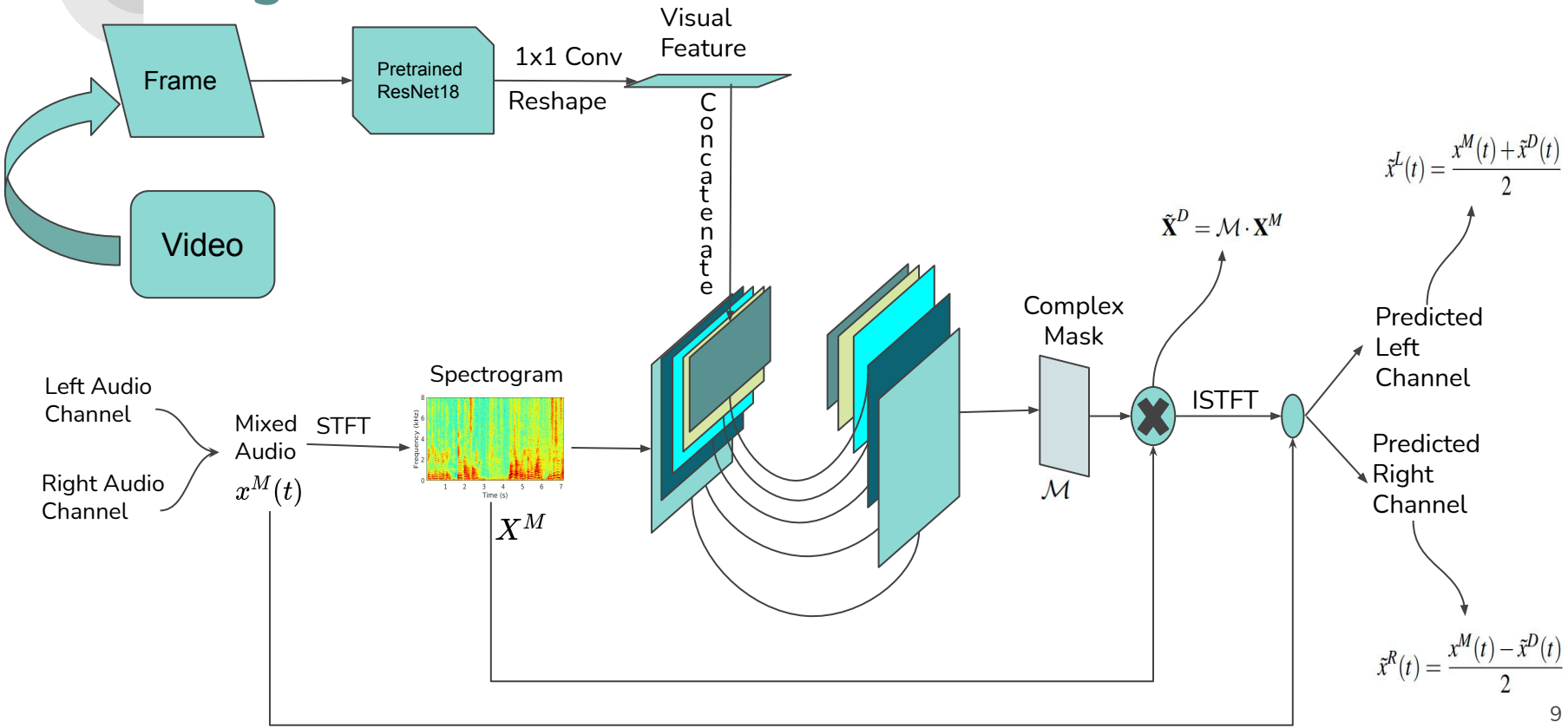
Key Idea

The key idea is that the visual frames reveals important spatial cues.

Suggestion:


Video of that Audio
Respective Frames
Mono Audio
Mono_to_Binaural Model
Binaural Audio

The Model Diagram



Some Of Test Results:-

Result 1

Input Audio with Video

Some Of Test Results:-

Result 1

Predicted

Some Of Test Results:-

Result 2

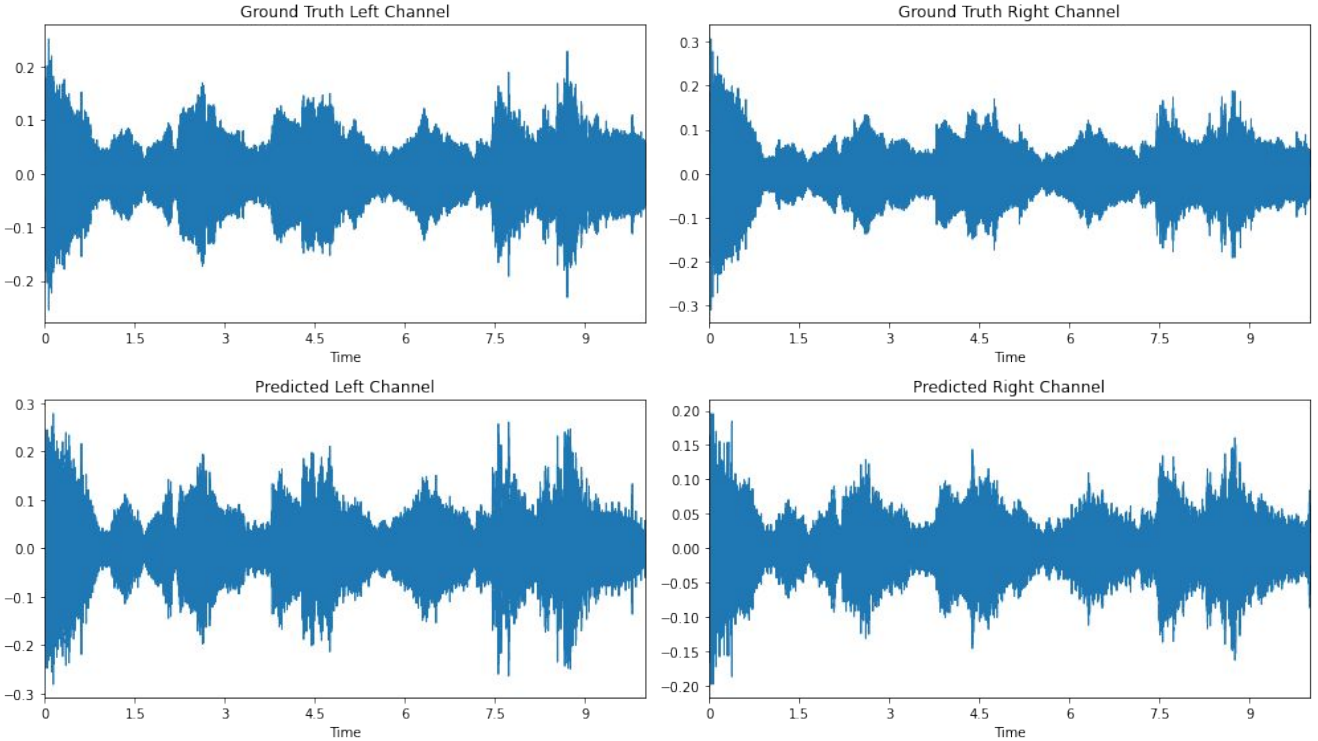
Input audio with Video

Some Of Test Results:-

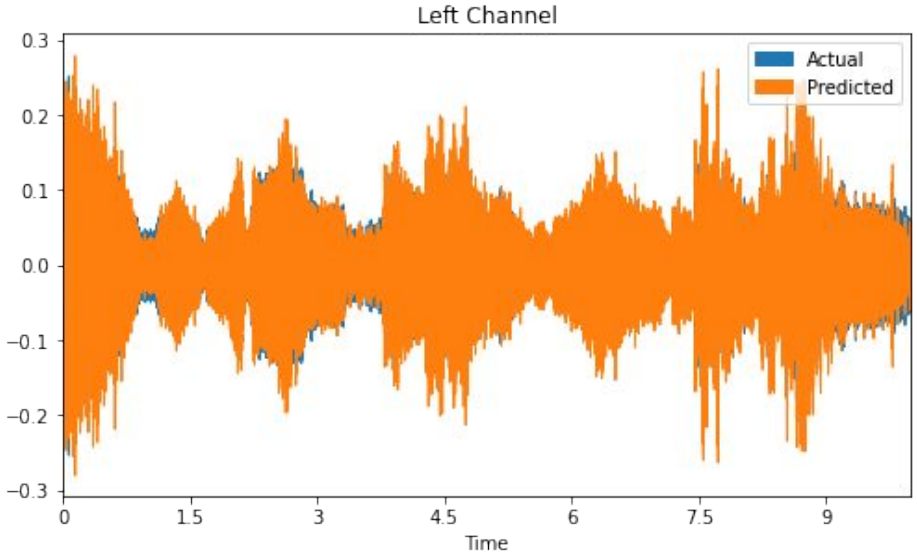
Result 2

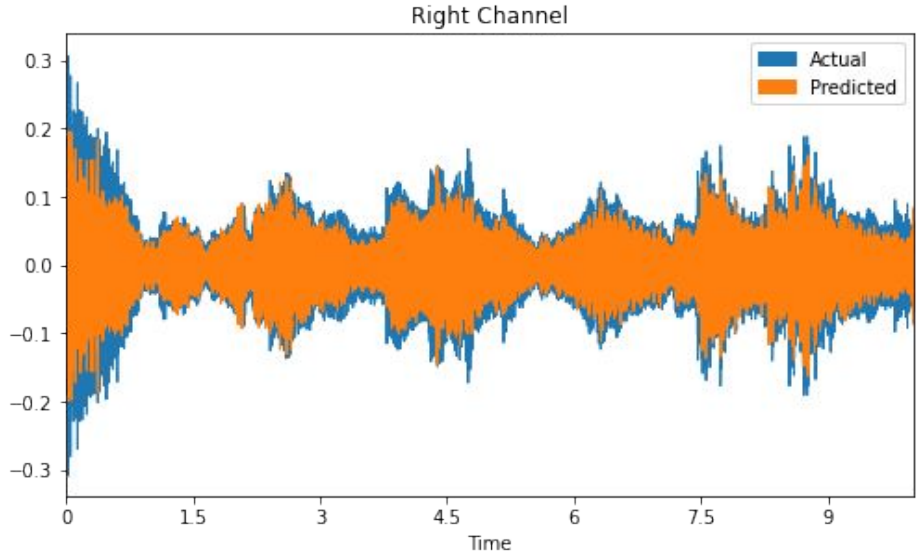
Predicted

How the previous audio “looks” like!



How the previous audio “looks” like!

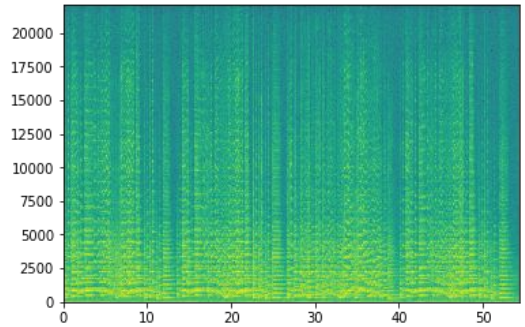




Training and Validation loss

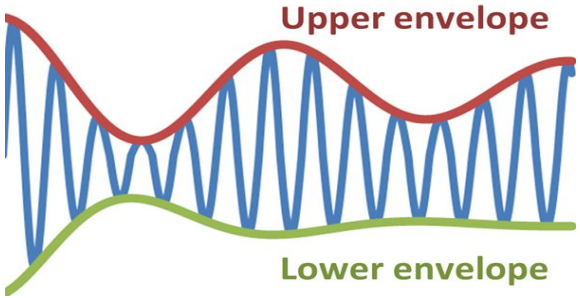
Used Metrics for Accuracy Generation

STFT Distance



$$\mathcal{D}_{\{\text{STFT}\}} = ||\mathbf{X}^L - \tilde{\mathbf{X}}^L||_2 + ||\mathbf{X}^R - \tilde{\mathbf{X}}^R||_2.$$

Envelope Distance



Let E[x(t)] denote the envelope of signal x(t).
The envelope distance is defined as:

$$\mathcal{D}_{\{\text{ENV}\}} = ||E[x^L(t)] - E[\tilde{x}^L(t)||_2 + ||E[x^R(t)] - E[\tilde{x}^R(t)||_2.$$

Accuracy

TABLE I

QUANTITATIVE RESULT OF BINAURAL AUDIO PREDICTION

	FAIR-PLAY	
Methods	STFT	ENV
Audio-Only	0.966	0.141
Flipped-Visual	1.145	0.149
Mono-Mono	1.155	0.153
MONO2BINAURAL(Original Paper)	0.836	0.132
MONO2BINAURAL(My training)	**1.020**	**0.146**

Thanking All