A Novel Approach to Underwater Acoustic Target Classification with MelGAN and Audio Spectrogram Transformer

Devichand Budagam

Contents

- Objective
- Dataset
- Data Pre-Processing
- MelGAN
- Audio Spectrogram Transformer

Objective

- Underwater acoustic target classification is a process used in underwater acoustics to identify and categorize objects or targets in the underwater environment using sound signals.

- The underwater environment is highly variable, Current classification methods often struggle to adapt to these variations, leading to reduced accuracy in target identification.

Dataset : ShipsEar Database

- ShipsEar is a database containing underwater recordings of ship and boat sounds, which has 90 recordings of 11 different vessel types.

Category	Type of Vessel
Class A	fishing boats, trawlers, mussel boats, tugboats and dredgers
Class B	motorboats, pilot boats and sailboats
Class C	passenger, ferries
Class D	ocean liners and ro-ro vessels
Class E	background noise recordings

- The amplifier used a 100 Hz high-pass filter to suppress marine background noise, the hydrophone sampling rate is 52,734 Hz, and the AD converter bit depth is 24 bits.

Class	A	B	C	D	E
Duration	1729s	1435s	4054s	2041s	923s

Data Pre-Processing

- The raw Audio signals with different recording lengths were uniformly split into an audio file(WAV format) of 0.5 seconds each with 26,368 samples.
- Data Preprocessing involved computation of mel spectrograms of all the audio samples.

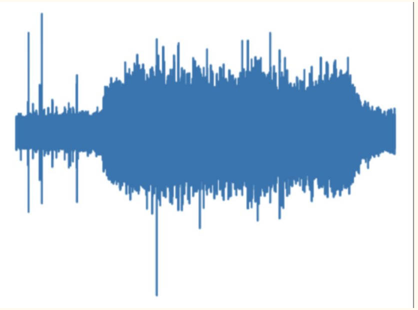
Class	A	B	C	D	E
No.of Samples	416	354	514	376	382

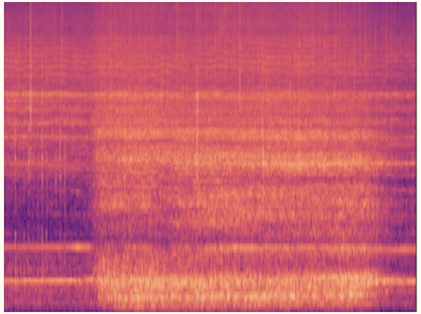
Pre-Processing Configuration:

- For MelGAN : No.of Mel Channels-80; Frame_length-1024; Frame_step=256; SR-52,734 Hz.
- For AST : alpha=0.97; No.of Mel Channels-128; Frame_length-2048; Frame_step=512; SR-52,734 Hz.

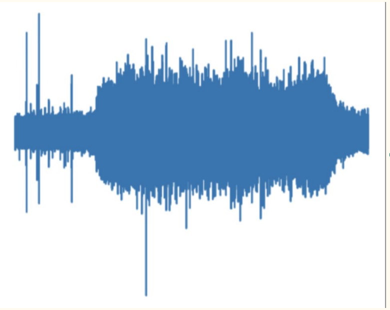
Pre-Processing Pipelines:

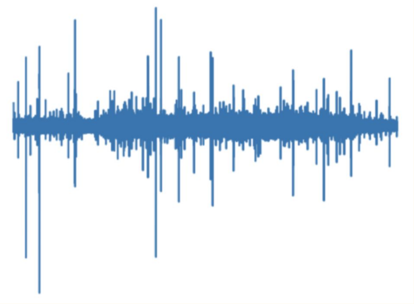
- For MelGAN:

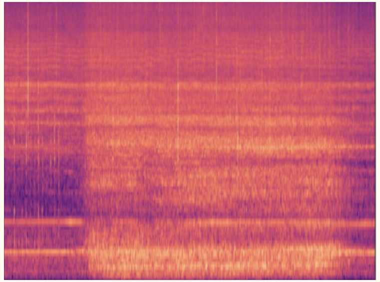

Raw Audio signal

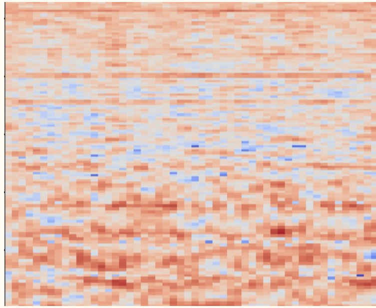

Mel Spectrogram

- For Audio Spectrogram Transformer:


Raw Audio signal

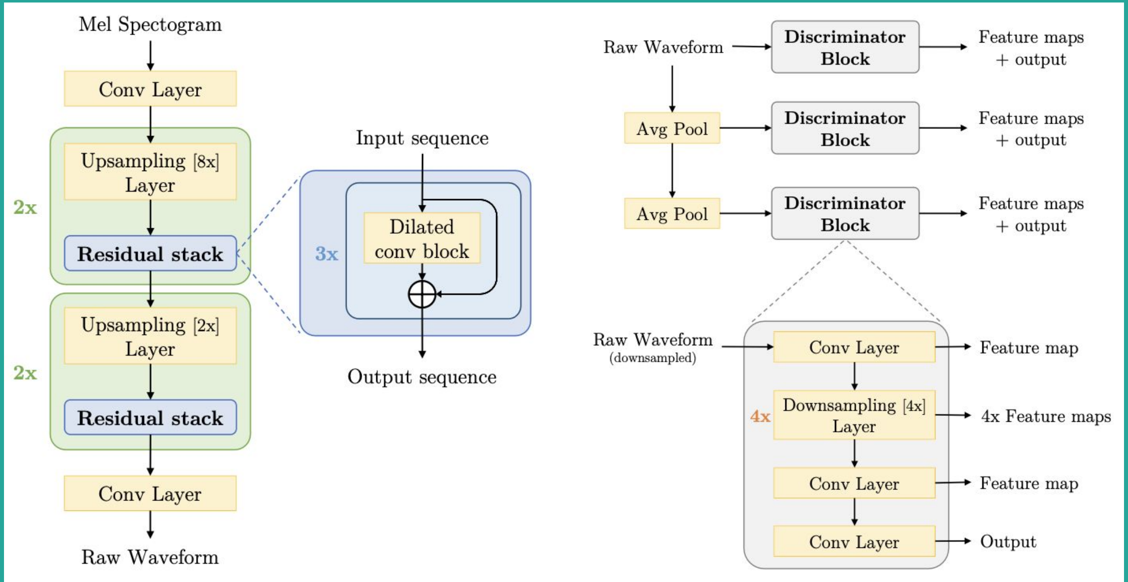

Pre Emphasis


Mel Spectrogram


Normalization

MelGAN

- MelGAN is a fast and efficient architecture for conditional generation of raw audio signals from corresponding mel spectrograms.
- It achieved excellent MOS score compared to other large models on various audio processing tasks.
- Generator has 47M parameters being one of the smallest models in the audio generation and discriminator has 16M parameters.



Generator **Discriminator**

Training Objective :

- The loss function involves both the general discriminator loss(Adversarial Learning) along with Feature matching loss for conditional generation of audio samples.

- Feature Matching Loss : $$\mathcal{L}_{\mathrm{FM}}(G, D_k) = \mathbb{E}_{x,s\sim p_{\mathrm{data}}}\left[\sum_{i=1}^{T}\frac{1}{N_i}||D_k^{(i)}(x) - D_k^{(i)}(G(s))||_1\right]$$

- Loss function : $$\min_G\left(\mathbb{E}_{s,z}\left[\sum_{k=1,2,3} -D_k(G(s,z))\right] + \lambda\sum_{k=1}^{3}\mathcal{L}_{\mathrm{FM}}(G, D_k)\right)$$

λ, being a hyperparameter, determines the effect of conditionality on the generator.

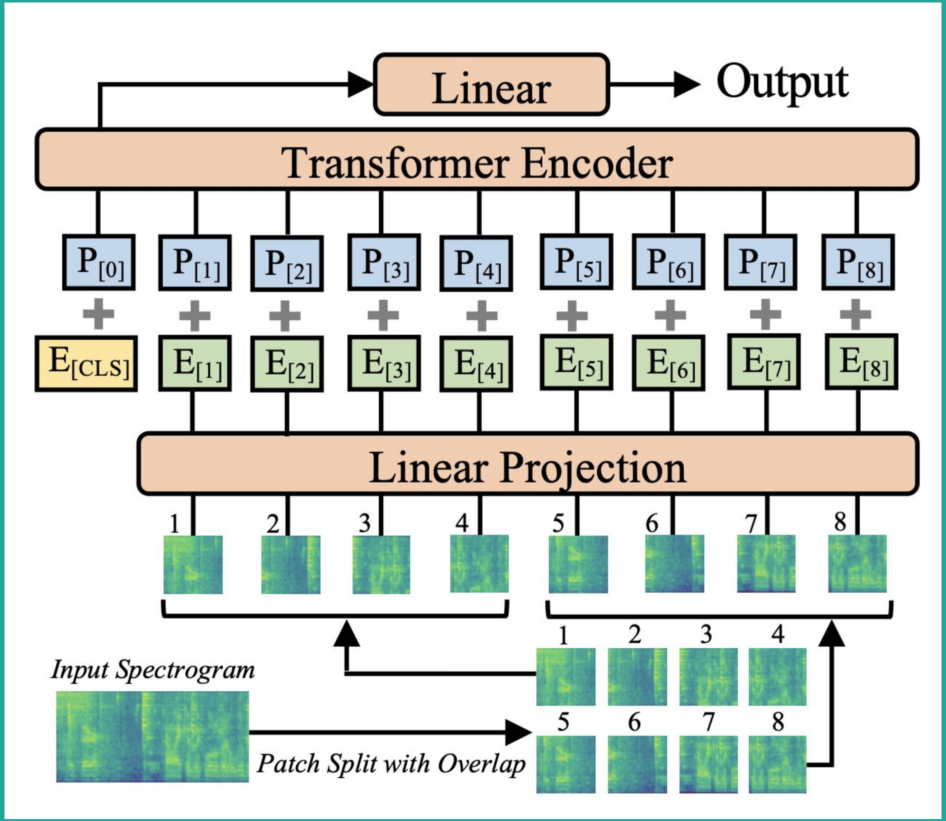
- MelGAN generates as many synthetic audio samples as real audio samples and is trained specifically for each class.
- Training Parameters : Adam Optimizer (Clipnorm=1)
  Learning rate (Generator): 1e-5
  Learning rate (Discriminator): 1e-6
  λ=10
  Batch Size =16
  Epochs=(80-100) Depending on number of audio samples in the class.

AST:Audio Spectrogram Transformer

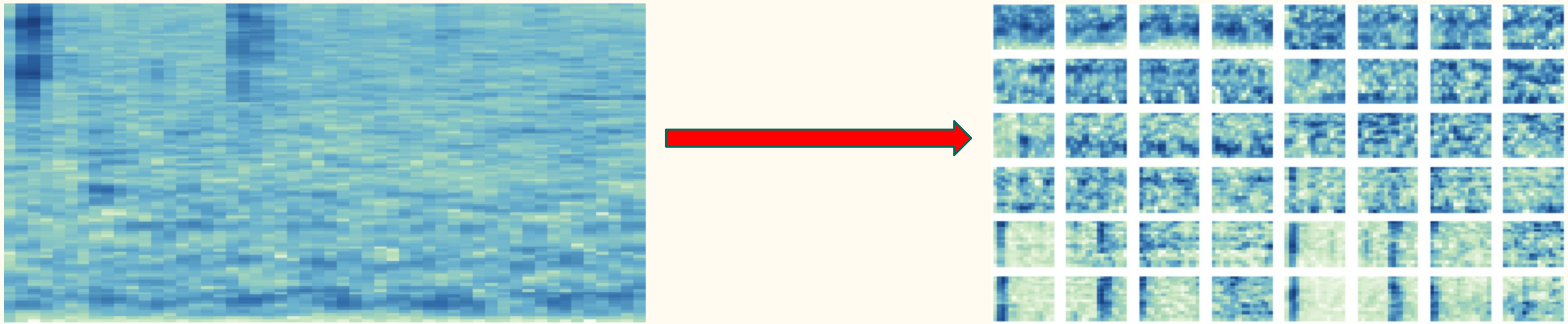
- AST is a *convolution-free*, *purely* attention- based model that is directly applied to an audio spectrogram and can capture long-range global context even in the lowest layers.

- AST is implementation of Vision Transformer(ViT) to Audio Spectrograms.

- Input mel spectrogram is of 128 × 52 dimensions. We then split the spectrogram into a sequence of N 16×16 patches with an overlap of 6 in both time and frequency dimension, hence 48 patches are the input for the Transformer. We flatten each 16×16 patch to a 1D patch embedding of size 384 using a linear projection layer.



- Since the Transformer architecture does not capture the input order information and the patch sequence is also not in temporal order, we add a trainable positional embedding to each patch embedding to allow the model to capture the spatial structure of the 2D audio spectrogram.



- Real audio samples along with synthetic audio samples are used for training and evaluation purposes with 75% as training(and validation data) and 25% as test data.Data is splitted with equal class distribution to construct balanced train and evaluation datasets.
- Train data samples : 2660; Validation data samples : 400; Test data samples : 1024

Training Parameters :

- Adam optimizer with a learning rate=1e–6; weight decay=1e-5
- Loss Function : Categorical Cross Entropy
- Batch size =32
- Epochs=50
- Number of Encoder Blocks=12
- Number of Multi-Attention Heads=6

Evaluation :

- Accuracy=83%
- Precision=91%
- Recall=90%