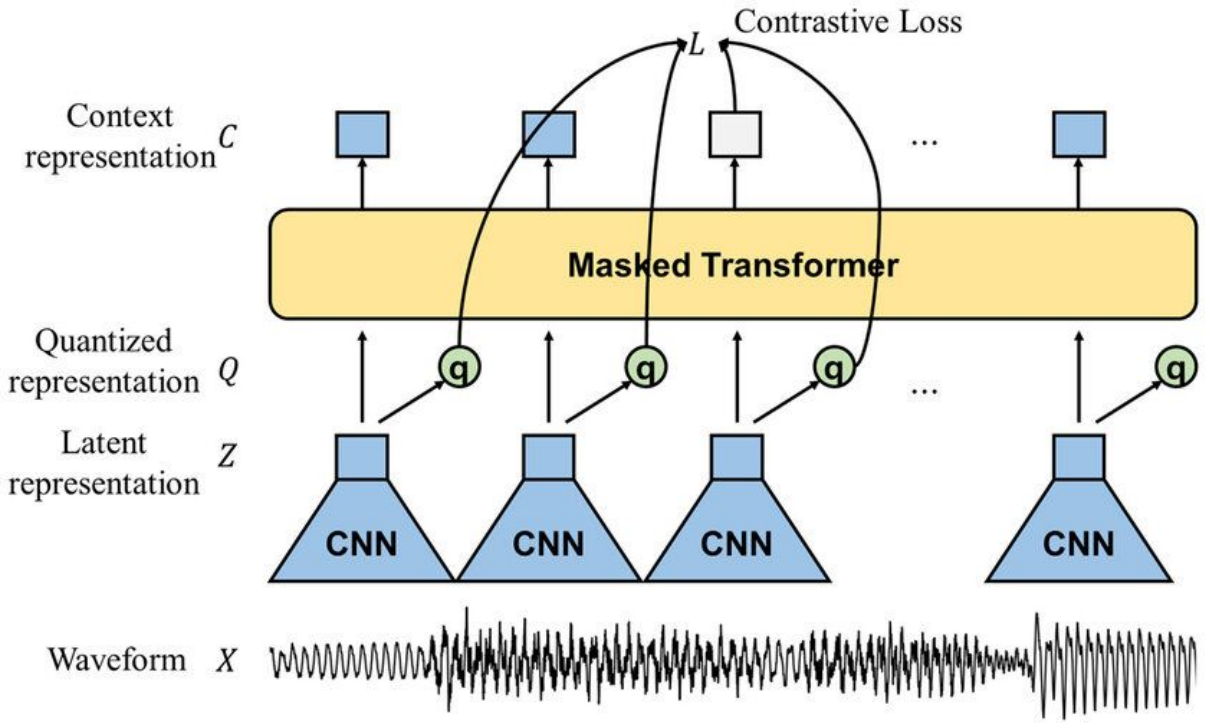
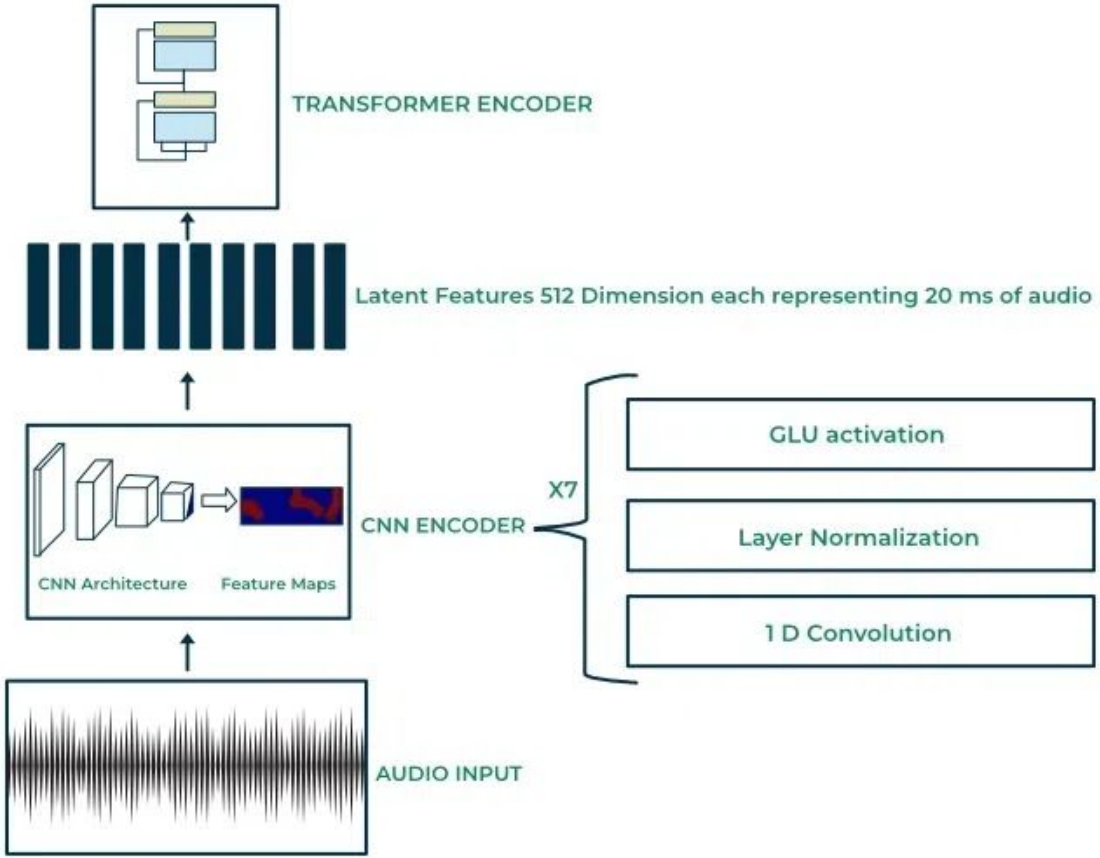
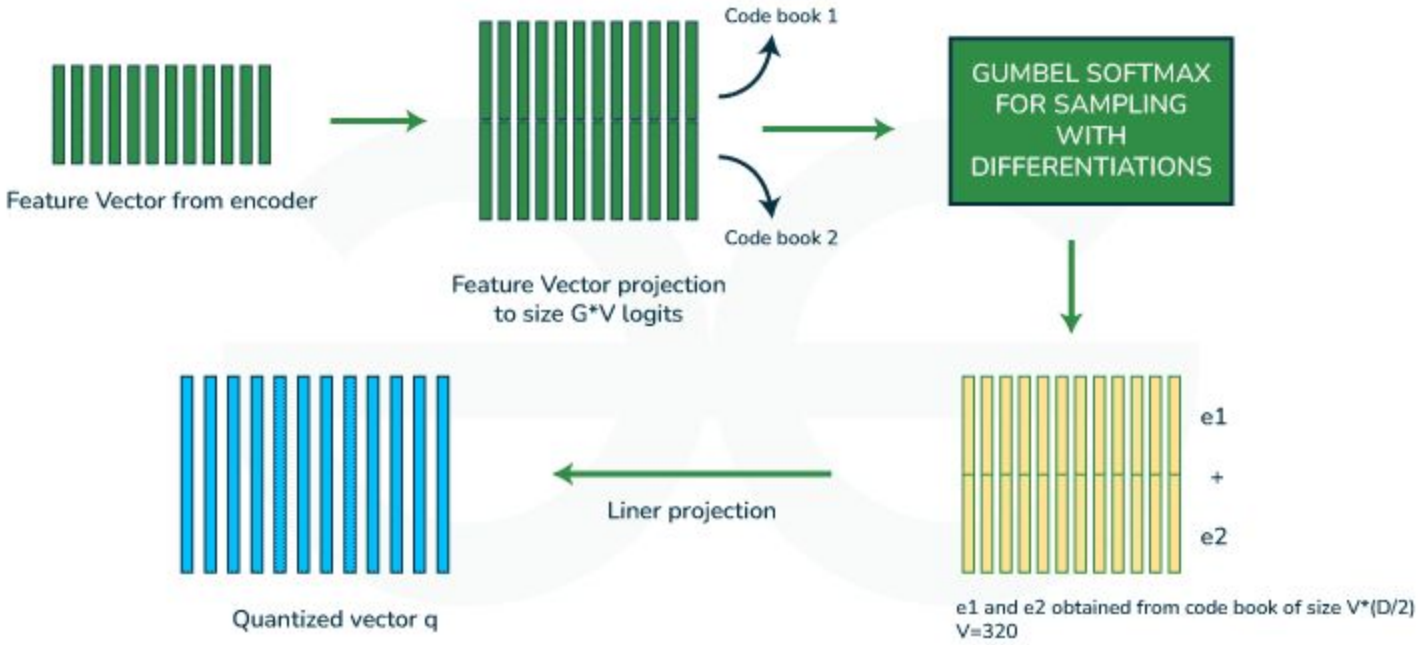
Endterm Evaluation

EE392A: UGPII (Audio Processing)

Wav2Vec2.0 Architecture




TRANSFORMER ENCODER
Latent Features 512 Dimension each representing 20 ms of audio
CNN Architecture
Feature Maps
CNN ENCODER
X7
GLU activation
Layer Normalization
1 D Convolution
AUDIO INPUT



Wav2Vec2 Quantization Process

Results

- Total 7 types of noises used on SNR ranging from 0-25
- 1000 Audio files from Libreespeech test dataset used to obtain the codebook pairs corresponding to clean and noisy versions
- Types of comparison
  - Offset + noise
  - Reverb + offset
  - Reverb + offset + noise

- Average accuracy with various reverb

RIR	0.1	0.2	0.4	0.5	0.7	0.8
Avg	8.04	5.13	4.53	2.1	2.65	1.12

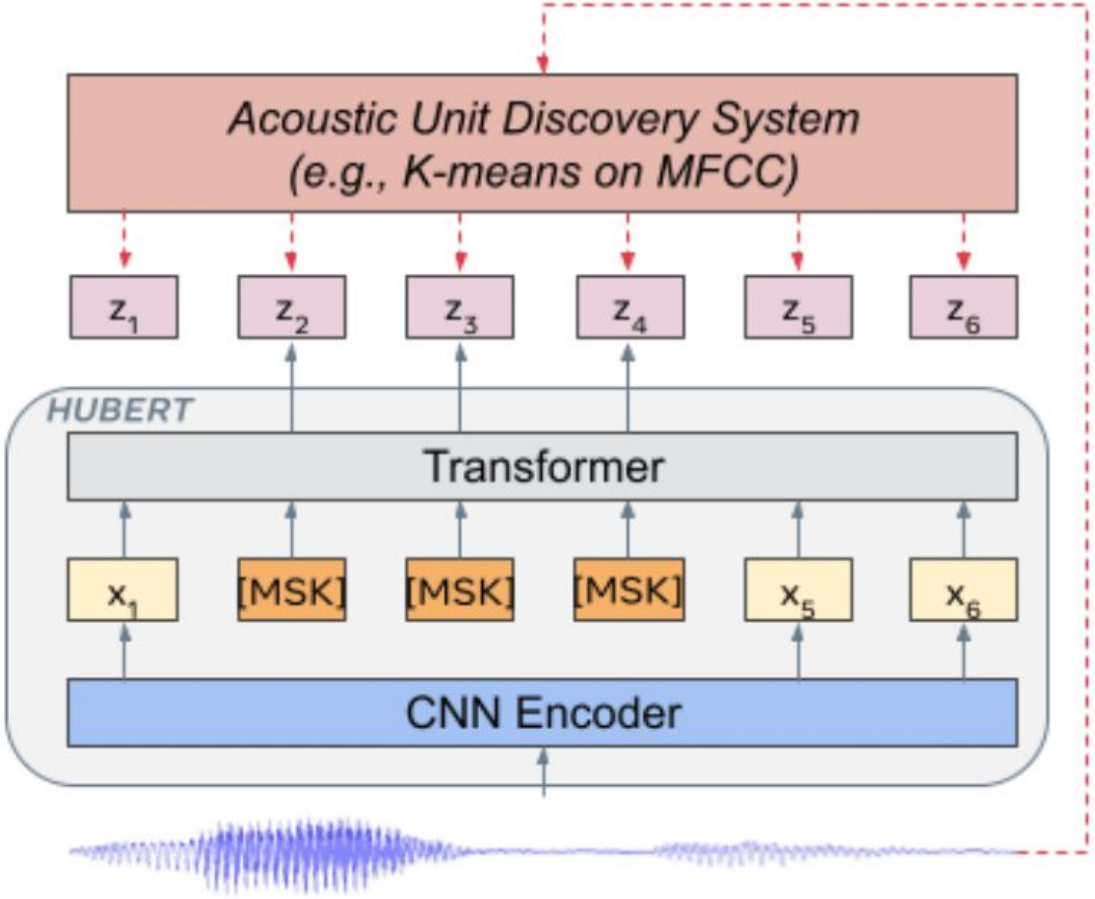
- Average accuracy for different type of Noise + offset

Noise v/s SNR	Babble	Cafeteria	Car	Livingroom	Shopping	Traffic	Train station	Avg
0	1.45	2.25	10.35	3.63	2.86	3.1	4.33	4
5	3.01	4.19	13.05	6.55	5.03	5.39	7.2	6.35
10	5.25	6.4	15.7	9.26	7.88	8.58	10.75	9.12
15	8.13	9.08	17.12	11.72	10.75	11.67	13.59	11.72
20	11.05	11.77	17.86	13.88	13.34	14.23	15.71	13.98
25	13.75	11.22	19.99	15.73	15.46	16.14	17	16.04
Avg	7.11	7.99	15.68	10.13	9.22	9.85	11.43	10.2

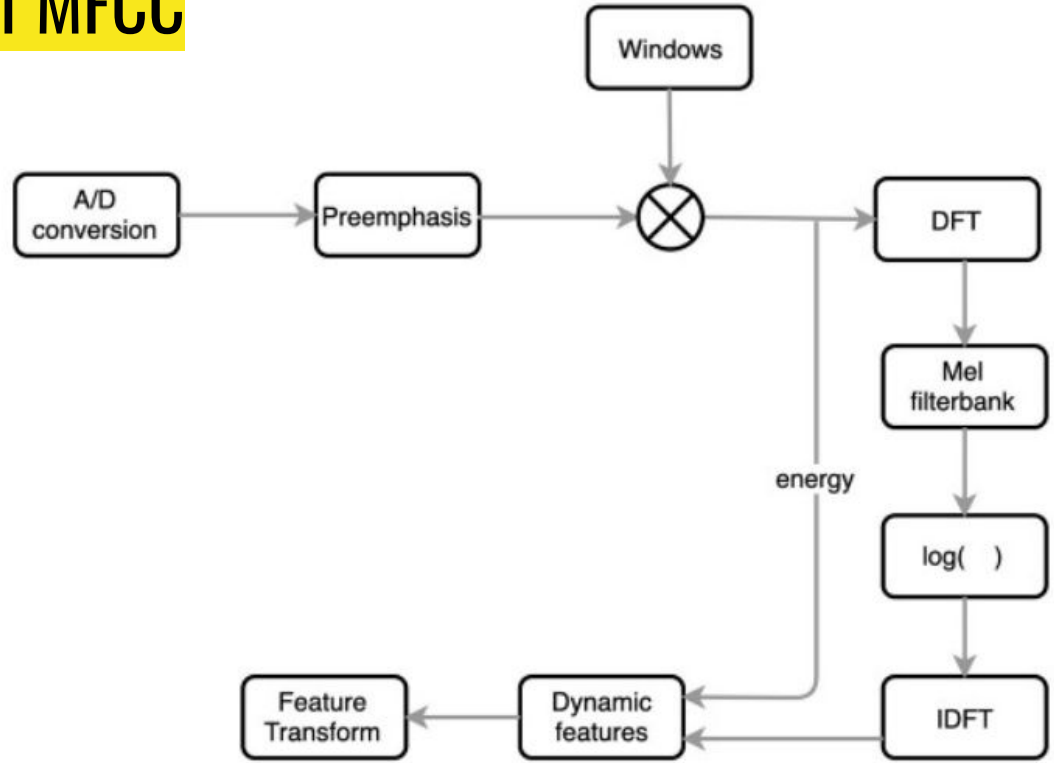
Percentage Change for the SNRs and given noise file with reverb = 0.5

SNR	Babble	Car	Cafeteria	Livingroom	Shopping	Traffic	Train station	Average
0	0.36	0.89	0.41	0.55	0.53	0.45	0.5	0.54
5	0.57	1.13	0.63	0.81	0.77	0.8	0.74	0.78
10	0.8	1.32	0.86	1.06	1.01	1.05	0.99	1.01
15	1.07	1.42	1.07	1.23	1.22	1.29	1.25	1.23
20	1.3	1.56	1.26	1.41	1.4	1.45	1.44	1.4
25	1.45	1.66	1.44	1.56	1.57	1.61	1.58	1.55
Avg	0.93	1.34	0.95	1.1	1.09	1.13	1.08	1.09

HuBERT Architeture



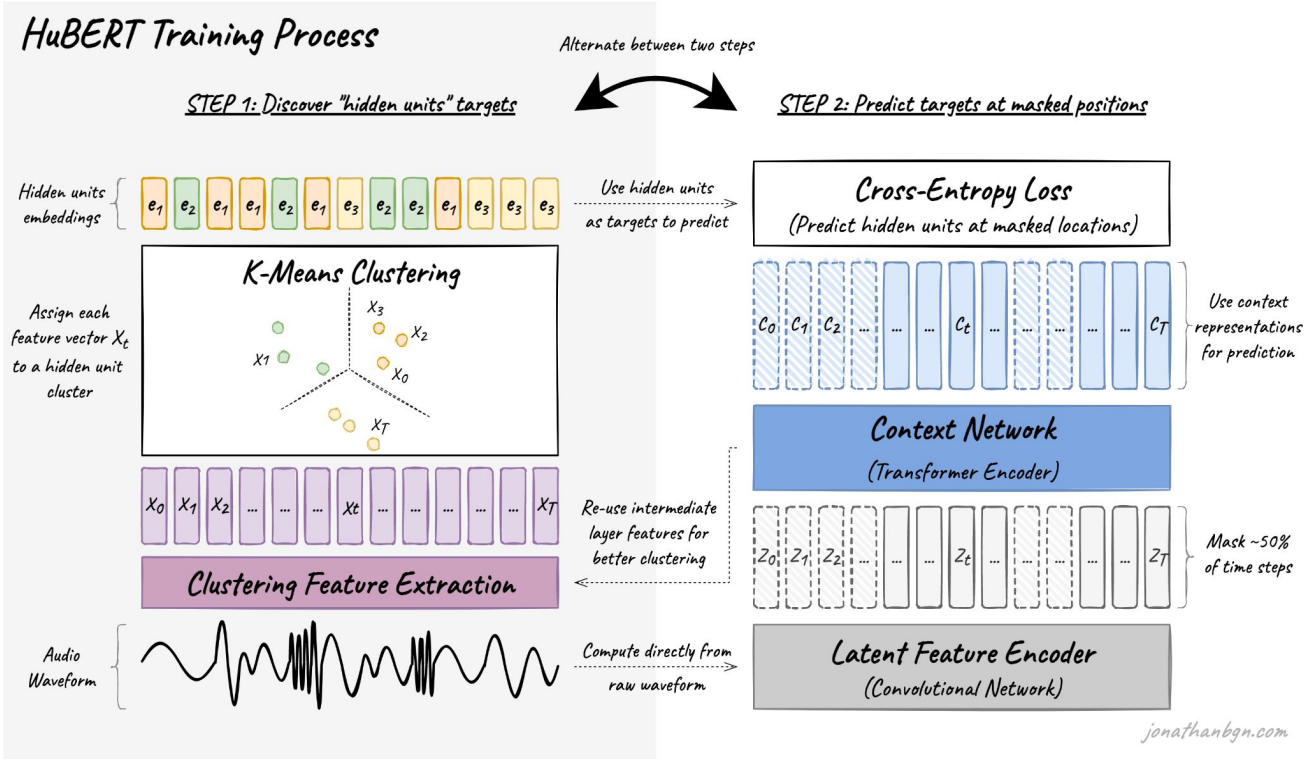
Rodmap of MFCC



HuBERT Framework

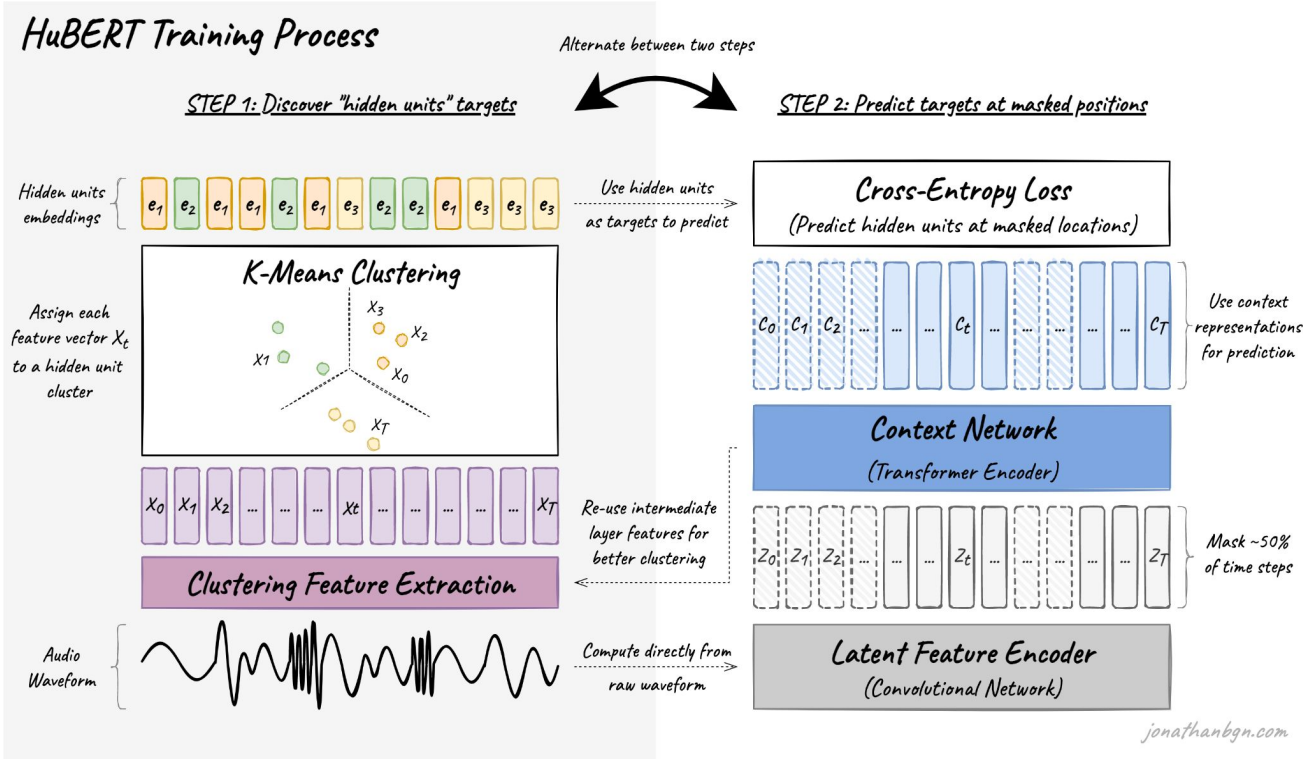
The training process alternates between two steps:

- clustering step to create pseudo-targets
- prediction step where the model tries to guess these targets at masked positions.



Discover “hidden units” targets through clustering

- Mel-Frequency Cepstral Coefficients (MFCCs) are used for the first clustering step.
- However, for subsequent clustering steps, representations from an intermediate layer of the HuBERT transformer encoder (from the previous iteration) are re-used.



Discover “hidden units” targets through clustering

The first step is to extract the hidden units (pseudo-targets) from the raw waveform of the audio. The K-means algorithm is used to assign each segment of audio (25 milliseconds) into one of K clusters.

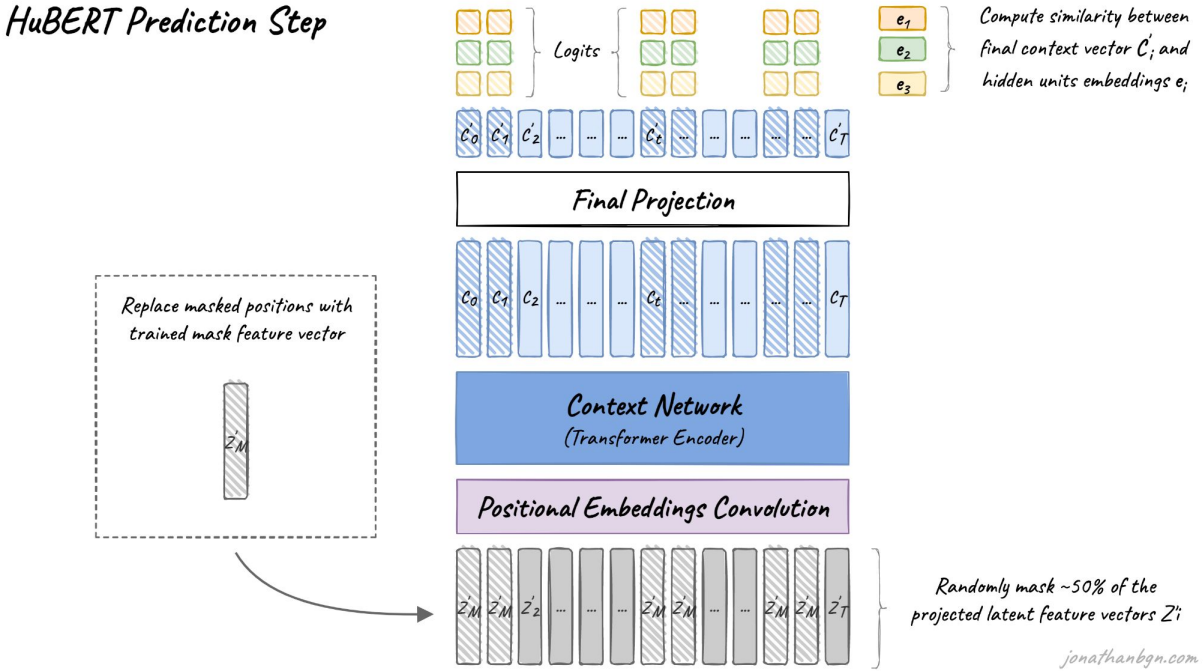
Each identified cluster will then become a hidden unit, and all audio frames assigned to this cluster will be assigned with this unit label.

Each hidden unit is then mapped to its corresponding embedding vector that can be used during the second step to make predictions

Predict “noisy” targets from context

50% of transformer encoder input features are masked, and the model is asked to predict the targets for these positions.

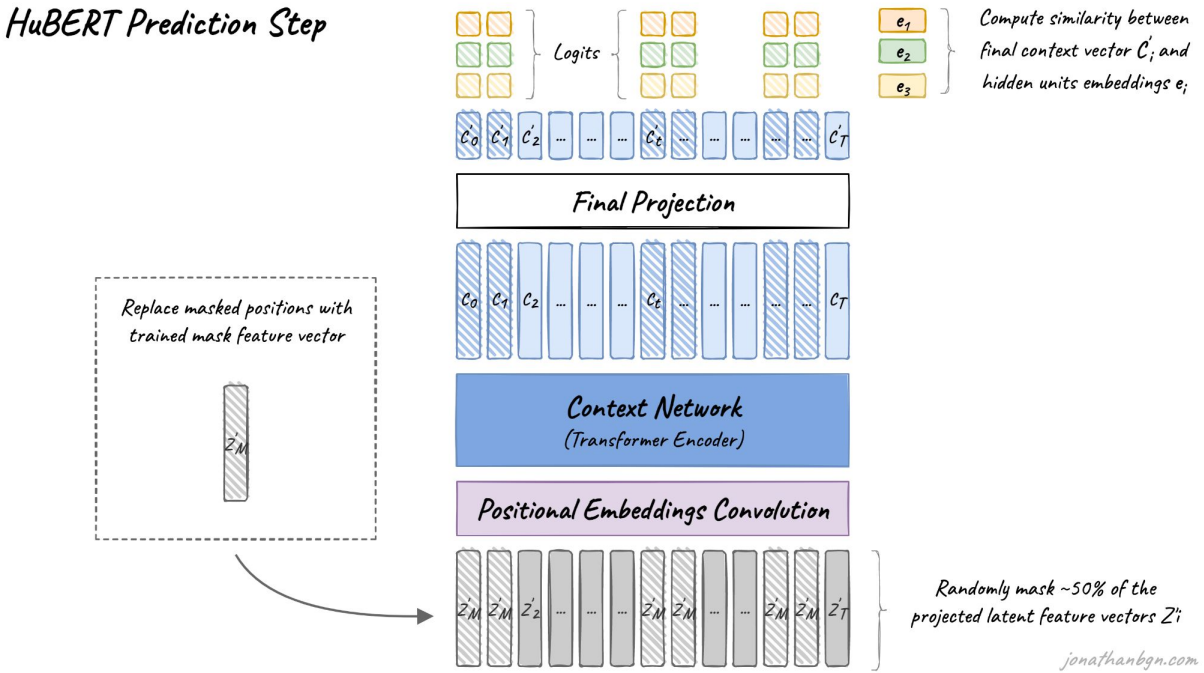
The cosine similarity is computed between the transformer outputs and each hidden unit embedding from all possible hidden units to give prediction logits.



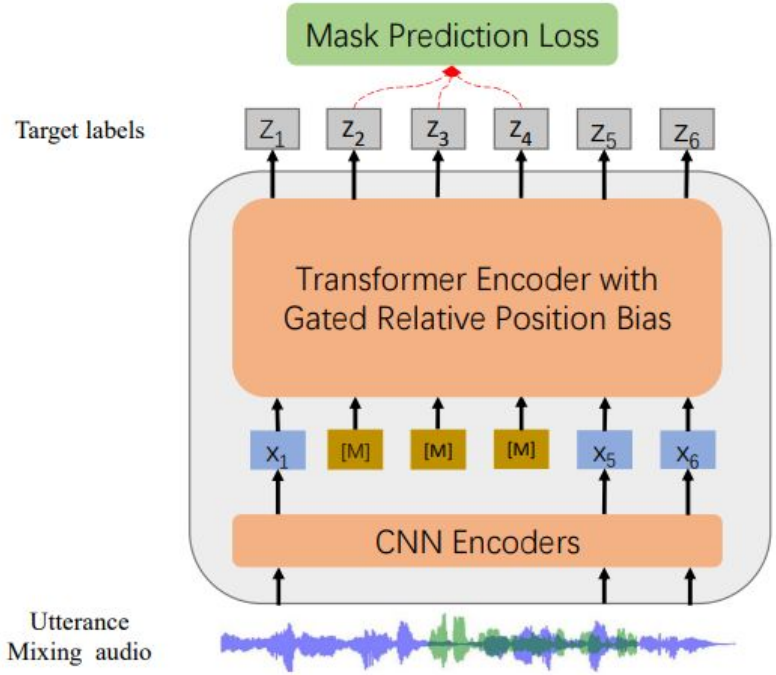
Predict “noisy” targets from context

The cross-entropy loss is then used to penalize wrong predictions.

The loss is only applied to the masked positions as it has been shown to perform better when using noisy labels.



WavLM Architecture



Results

- Total 7 types of noises used on SNR ranging from 0-25
- 1000 Audio files from Libreespeech test dataset used to obtain the codebook pairs corresponding to clean and noisy versions
- Types of comparison
  - Offset + noise
  - Reverb + offset
  - Reverb + offset + noise

- Average accuracy with various reverb

RIR	0.1	0.2	0.4	0.5	0.7	0.8
Percent change	26.27	23.54	23.11	13.48	19.59	8.26

- Average accuracy for different type of Noise + offset

Noise v/s SNR	Babble	Cafeteria	Car	Livingroom	Shopping	Traffic	Train station	Avg
0	3.65	8.22	30.34	16.26	13.73	19.78	3.65	13.66
5	13.34	16.33	31.22	23.37	21.89	25.41	13.34	20.7
10	22.76	24.55	31.64	27.91	26.62	26.11	22.76	26.05
15	26.7	28.62	32.02	30.04	27.16	27.43	26.7	28.38
20	28.97	30.28	32.24	31.06	28.2	28.02	28.97	29.68
25	30.02	31.12	32.45	31.42	28.72	29.18	30.12	30.45
Avg	28.93	23.19	31.66	26.68	24.39	25.99	20.92	25.97

Percentage Change for the SNRs and given noise file with reverb = 0.5

SNR	Babble	Car	Cafeteria	Livingroom	Shopping	Traffic	Train station	Avg
0	1.39	8.77	1.74	3.39	3.31	4.99	3.45	3.86
5	2.48	10.73	2.77	5.14	5.4	7.47	6.48	5.78
10	5.35	11.99	4.96	7.44	8.2	9.82	9.11	8.12
15	8.69	13.75	7.7	9.83	10.07	11.38	9.79	10.17
20	10.76	15.79	9.79	11.37	11.58	12.33	10.57	11.74
25	11.79	16.41	10.12	12.28	12.49	12.85	13.49	12.78
Avg	6.74	12.91	6.18	8.24	8.51	9.81	8.81	8.74