Speaker Counting

Kanishk Yadav

Final year undergrad at BITS Pilani, India

31 May 2024

January 2024 - June 2024

Last week's task

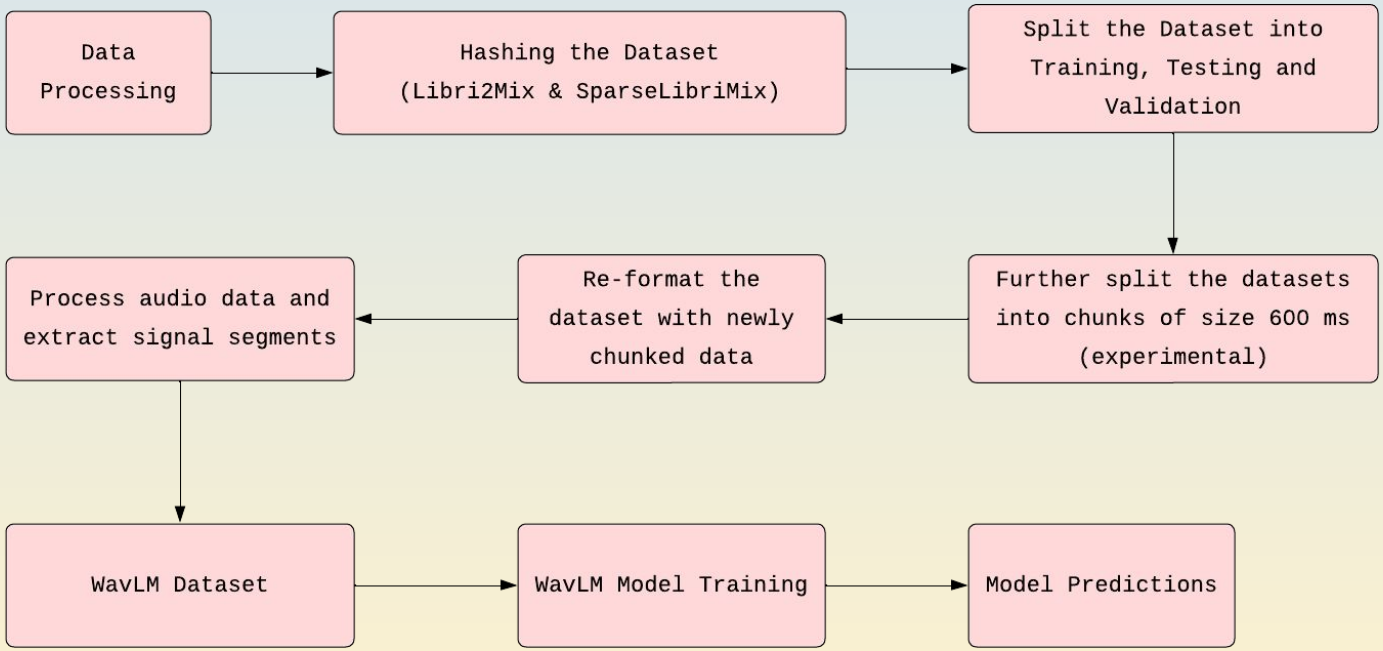
1. Implement Overlapped-Speech Detection (AISG) as baseline for upcoming week and report results

OSD (AISG): Datasets

1. **Libri2Mix**
   - mixtures generated for two random clean audio files from LibriSpeech Dataset
   - mono-channel, 16 kHz
2. **SparseLibriMix**
   - variable % overlap [0, 0.2, 0.4, 0.6, 0.8, 1]
   - added noise (optional)

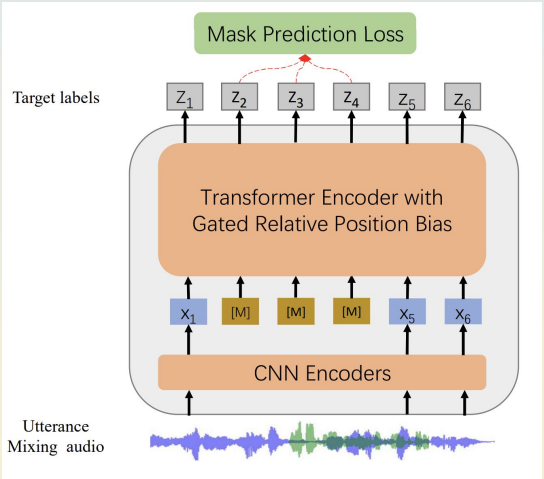
Total Model Params	94.70 million
Training Data	Libri2mix Training set (84 hours)
Evaluation Data	Libri2Mix Evaluation set (26 hours)
Test Data	Libri2Mix Testing set (25 hours) + SparseLibriMix (6 hours)

OSD (AISG) Pipeline Structure



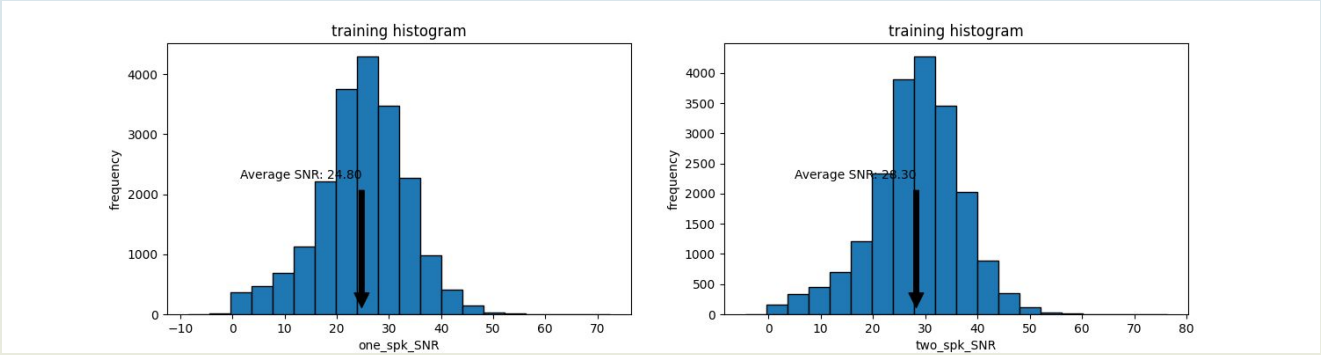
WavLM Model [1]

1. **Transformer-based** architecture: Utilizes a gated relative position bias to better capture sequential information
2. WavLM integrates **masked** speech prediction with denoising, handling both ASR and non-ASR tasks effectively
3. Trained on a large-scale dataset of **94k hours**, including data from Libri-Light, GigaSpeech, and VoxPopuli, enabling robust learning and adaptability to various speech applications.

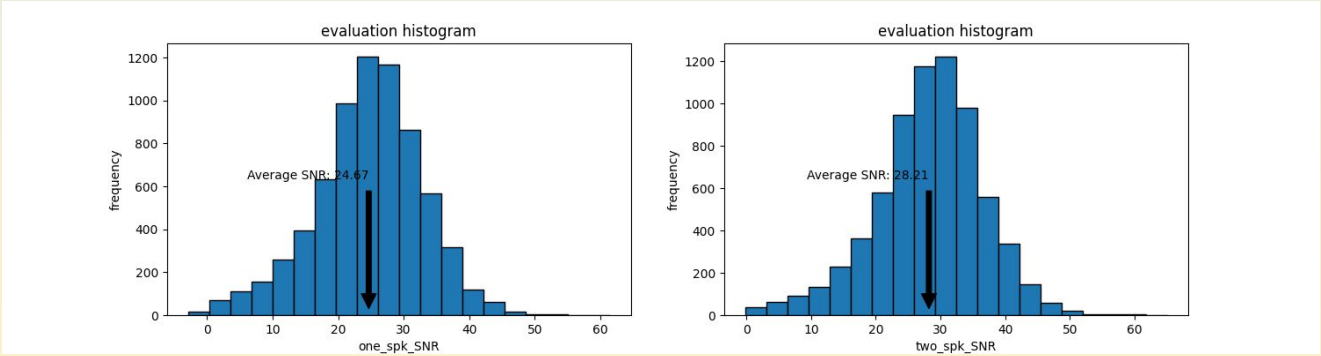


[1] Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Zeng, M., & Wei, F. (2022). WAVLM: Large-Scale Self-Supervised Pre-Training for Full stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1505–1518. https://doi.org/10.1109/jstsp.2022.3188113

OSD (AISG): Data Analysis



Frequency Distribution of SNR for One Speaker and Two Speaker



OSD (AISG): Performance (Evaluation - Libri2Mix)

Class (#speakers)	Precision	Recall	F1-Score	Support (#occurrences)
0	0.97	0.47	0.63	7882
1	0.89	0.86	0.88	62321
2	0.91	0.97	0.94	95348
Overall (Accuracy)			0.91	165551
Macro Avg	0.92	0.77	0.82	165551
Weighted Avg	0.91	0.91	0.90	165551

Confusion Matrix

Actual/Predicted	Class 0	Class 1	Class 2
Class 0	3719	3741	422
Class 1	115	53902	8304
Class 2	1	2957	92390

OSD (AISG): Performance (Test - Libri2Mix)

Class (#speakers)	Precision	Recall	F1-Score	Support (#occurrences)
0	0.97	0.35	0.51	6329
1	0.88	0.93	0.91	38237
2	0.93	0.97	0.95	40141
Overall (Accuracy)			0.91	84707
Macro Avg	0.93	0.75	0.79	84707
Weighted Avg	0.91	0.91	0.90	84707

Confusion Matrix

Actual/Predicted	Class 0	Class 1	Class 2
Class 0	2185	3625	519
Class 1	75	35736	2426
Class 2	1	1062	39078

OSD (AISG): Performance (Test - SparseLibriMix)

Class (#speakers)	Precision	Recall	F1-Score	Support (#occurrences)
0	0.97	0.28	0.44	2348
1	0.92	0.96	0.94	21139
2	0.91	0.98	0.94	10138
Overall (Accuracy)			0.91	33625
Macro Avg	0.93	0.74	0.77	33625
Weighted Avg	0.92	0.91	0.90	33625

Confusion Matrix

Actual/Predicted	Class 0	Class 1	Class 2
Class 0	661	1580	107
Class 1	21	20191	927
Class 2	1	238	9899

Next week's task

1. Experiment with chunk size and other hyperparameters during WavLM training
2. Implement OSD (AISG) for 3 speakers (Libri3Mix) and report results
3. Try integrating different loss function during WavLM training

Speaker Counting

Kanishk Yadav

Final year undergrad at BITS Pilani, India

06 June 2024

January 2024 - June 2024

Last week's task

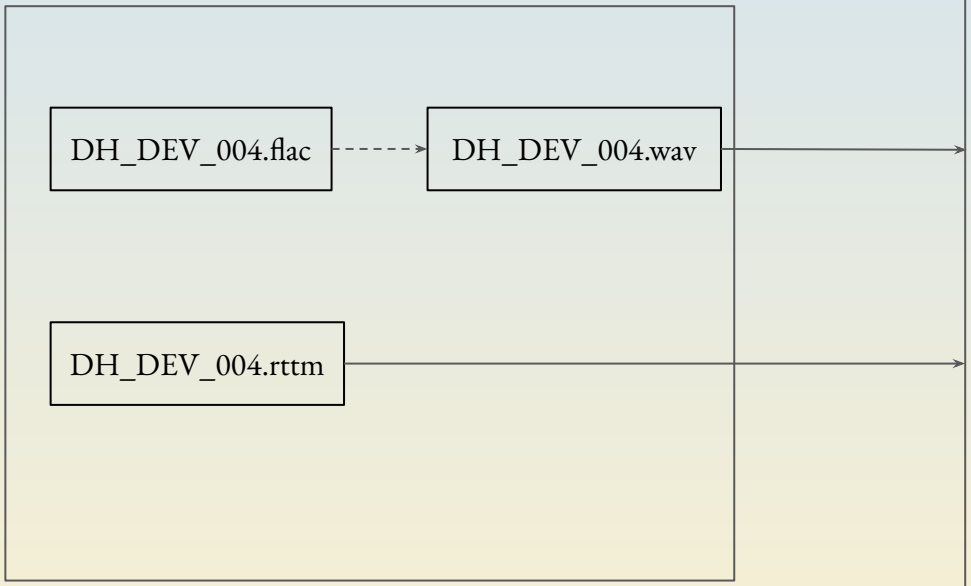
1. Inference Pre-trained Overlapped-Speech Detection (AISG) for DIHARD III dataset (0.6 s and 0.4 s chunk)

OSD (AISG): Datasets

1. **Libri2Mix**
   - mixtures generated for two random clean audio files from LibriSpeech Dataset
   - mono-channel, 16 kHz
2. **Modified DIHARD III**
   - modified into a hugging face object for use with WavLM model

Total Model Params	94.70 million
Training Data	Libri2mix Training set (84 hours)
Inference Data	**Modified** DIHARD III set (67 hours)

Modified DIHARD III



hugging face object

[
{
audio id: DH_DEV_004,

input values: [0.000213623046875, -3.0517578125e-05, -0.000213623046875 . . . . . .]

label: 'speaker count for the chunk' (0/1/2)

chunk_filename: 'DH_DEV_004_0.0_0.6'
}

{
audio id: DH_DEV_004,

input values: [6.103515625e-05, 0.0001220703125, 0.000213623046875 . . . . . .]

label: 'speaker count for the chunk' (0/1/2)

chunk_filename: 'DH_DEV_004_0.6_1.2'
}
]

OSD (AISG): Performance (dev - DIHARD III)

Inference performed with a chunk of 0.6 s using the WavLM model, which also trained on a chunk of 0.6 s

(0.6, 0.6)

Class (#speakers)	Precision	Recall	F1-Score	Support (#occurrences)
0	0.28	0.86	0.42	13987
1	0.81	0.82	0.81	147567
2	0.69	0.19	0.30	43450
Overall (Accuracy)			0.69	205004
Macro Avg	0.59	0.62	0.51	205004
Weighted Avg	0.74	0.69	0.68	205004

Confusion Matrix

Actual/Predicted	Class 0	Class 1	Class 2
Class 0	**12084**	1832	71
Class 1	23354	**120546**	3667
Class 2	7896	**27311**	8243

OSD (AISG): Performance (eval - DIHARD III)

Inference performed with a chunk of 0.4 s using the WavLM model, which also trained on a chunk of 0.6 s

(0.6, 0.4)

Class (#speakers)	Precision	Recall	F1-Score	Support (#occurrences)
0	0.26	0.71	0.38	29843
1	0.83	0.77	0.80	224224
2	0.71	0.13	0.23	43094
Overall (Accuracy)			0.67	297161
Macro Avg	0.60	0.54	0.47	297161
Weighted Avg	0.76	0.67	0.67	297161

Confusion Matrix

Actual/Predicted	Class 0	Class 1	Class 2
Class 0	**21175**	8618	50
Class 1	49527	**172419**	2278
Class 2	11433	**25877**	5784

Next week's task

1. Improve class imbalance for 0, 1 and 2 speakers in the dataset
2. Train the WavLM model on the DIHARD III dataset and report results

Speaker Counting

Kanishk Yadav

Final year undergrad at BITS Pilani, India

21 June 2024

January 2024 - June 2024

Last week's task

1. Train the WavLM model on the DIHARD III dataset, report results and compare with the WavLM model trained on LibriSpeech

OSD (AISG): Performance on DIHARD III

Inference performed with a chunk of 0.6 s using the WavLM model which is fine-tuned on **LibriSpeech** dataset on a chunk of 0.6 s

(0.6, 0.6)

Class (#speakers)	Precision	Recall	F1-Score	Support (#occurrences)
0	0.28	0.89	0.43	14363
1	0.82	0.81	0.81	144056
2	0.66	0.18	0.28	39728
Overall (Accuracy)			**0.69**	198147
Macro Avg	0.59	0.62	0.61	198147
Weighted Avg	0.75	0.69	0.68	198147

Confusion Matrix

Actual/Predicted	Class 0	Class 1	Class 2
Class 0	**12764**	1499	100
Class 1	24295	**116146**	3615
Class 2	8254	**24322**	7152

WavLM model trained on DIHARD III

Inference performed with a chunk of 0.6 s using the WavLM model which is fine-tuned on **DH3** dataset on a chunk of 0.6 s

(0.6, 0.6)

Class (#speakers)	Precision	Recall	F1-Score	Support (#occurrences)
0	0.83	0.78	0.80	14363
1	0.89	0.94	0.91	144056
2	0.77	0.65	0.70	39728
Overall (Accuracy)			**0.87**	198147
Macro Avg	0.83	0.79	0.81	198147
Weighted Avg	0.86	0.87	0.86	198147

Confusion Matrix

Actual/Predicted	Class 0	Class 1	Class 2
Class 0	**11189**	2940	234
Class 1	2022	**134769**	7265
Class 2	350	13751	**25627**

Significant reduction in misclassification →

Speaker Counting

Kanishk Yadav

Final year undergrad at BITS Pilani, India

19 July 2024

January 2024 - July 2024

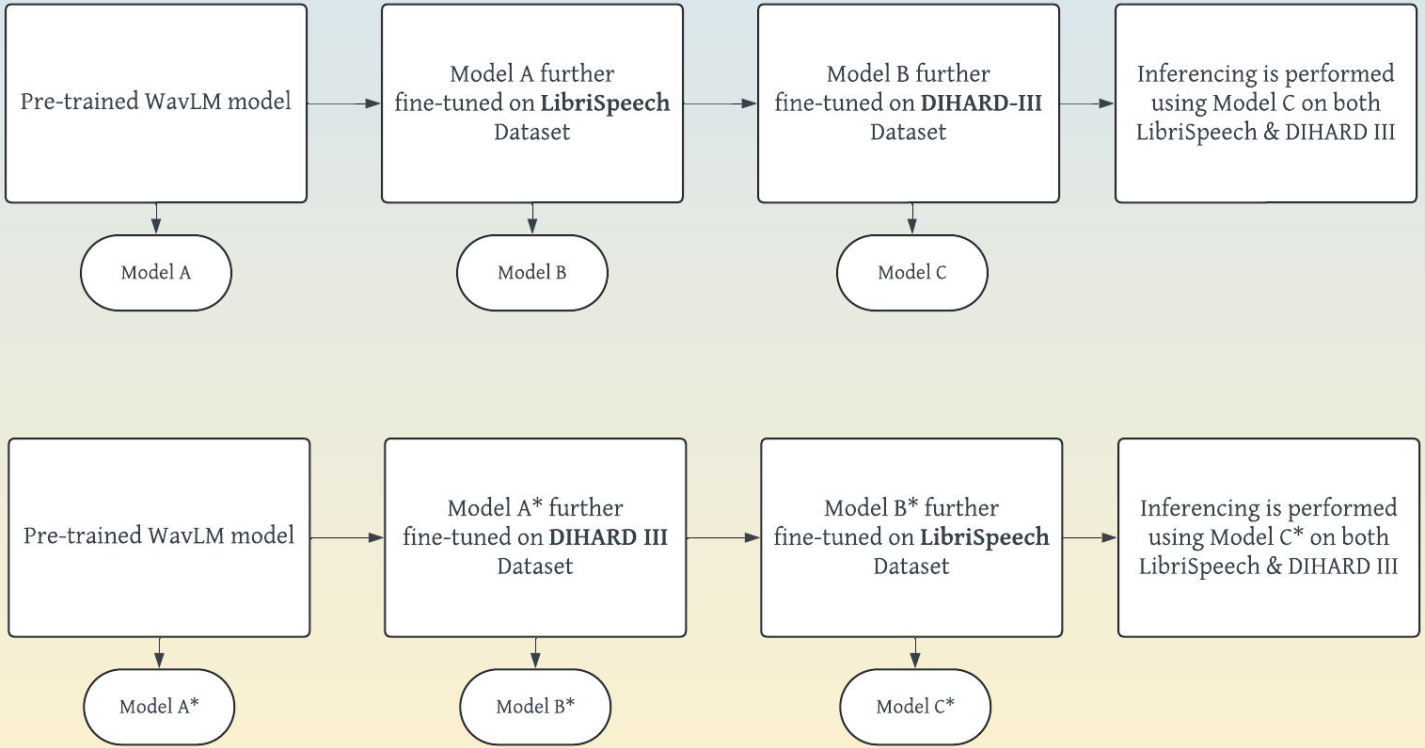
Last week's task

1. Sequentially fine-tune the WavLM model on the combined datasets and compare results:

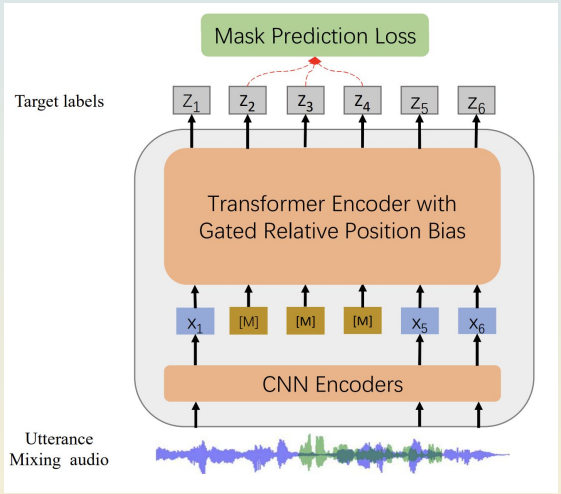
[1] LibriSpeech *followed by* DIHARD-III
[2] DIHARD-III *followed by* LibriSpeech

2. Implemented batch-training for the inference pipeline (reduction in inference time by 0.5x)

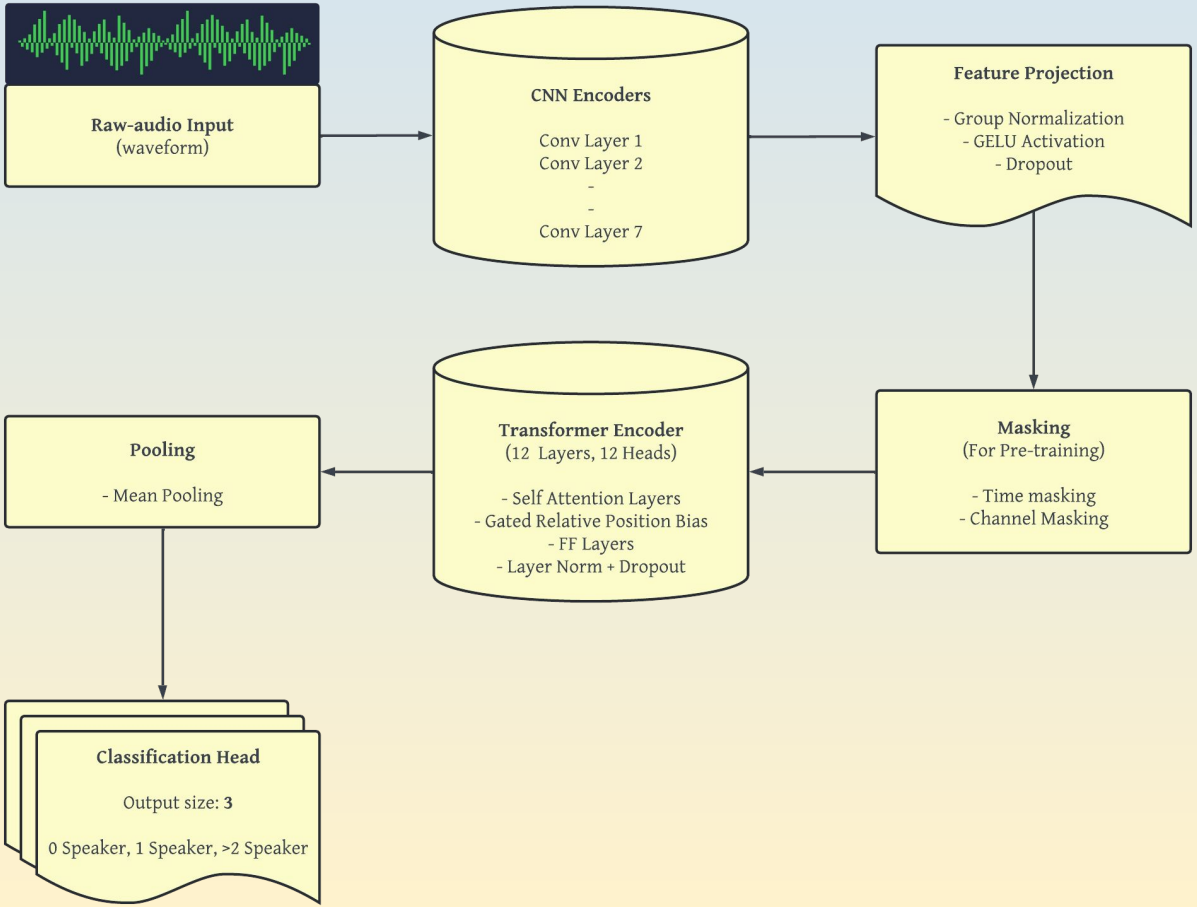
Procedure


Pre-trained WavLM model
Model A further fine-tuned on **LibriSpeech** Dataset
Model B further fine-tuned on **DIHARD-III** Dataset
Inferencing is performed using Model C on both LibriSpeech & DIHARD III
Model A
Model B
Model C
Pre-trained WavLM model
Model A* further fine-tuned on **DIHARD III** Dataset
Model B* further fine-tuned on **LibriSpeech** Dataset
Inferencing is performed using Model C* on both LibriSpeech & DIHARD III
Model A*
Model B*
Model C*

WavLM Architecture



WavLM Architecture (OSD)



WavLM model fine-tuned on DIHARD III

Inference performed with a chunk of 0.6 s using the WavLM model which is fine-tuned on **only DH3** dataset on a chunk of 0.6 s

(0.6, 0.6)

Class (#speakers)	Precision	Recall	F1-Score	Support (#occurrences)
0	0.83	0.78	0.80	14363
1	0.89	0.94	0.91	144056
2	0.77	0.65	0.70	39728
Overall (Accuracy)			**0.87**	198147
Macro Avg	0.83	0.79	0.81	198147
Weighted Avg	0.86	0.87	0.86	198147

Confusion Matrix

Actual/Predicted	Class 0	Class 1	Class 2
Class 0	**11189**	2940	234
Class 1	2022	**134769**	7265
Class 2	350	13751	**25627**

WavLM model fine-tuned on LibriSpeech

Inference performed with a chunk of 0.6 s using the WavLM model which is fine-tuned on **only LibriSpeech** dataset on a chunk of 0.6 s

(0.6, 0.6)

Class (#speakers)	Precision	Recall	F1-Score	Support (#occurrences)
0	0.97	0.35	0.51	6329
1	0.88	0.93	0.91	38237
2	0.93	0.97	0.95	40141
Overall (Accuracy)			**0.91**	84707
Macro Avg	0.93	0.75	0.79	84707
Weighted Avg	0.91	0.91	0.90	84707

Confusion Matrix

Actual/Predicted	Class 0	Class 1	Class 2
Class 0	2185	**3625**	519
Class 1	75	**35736**	2426
Class 2	1	1062	**39078**

WavLM model fine-tuned on (LibriSpeech ⟶ DIHARD III)

Inference performed with a chunk of 0.6 s using the WavLM model on **DH3** dataset on a chunk of 0.6 s

(0.6, 0.6)

Class (#speakers)	Precision	Recall	F1-Score	Support (#occurrences)
0	0.82	0.80	0.81	14363
1	0.89	0.93	0.91	144056
2	0.76	0.66	0.71	39728
Overall (Accuracy)			**0.87**	198147
Macro Avg	0.83	0.79	0.81	198147
Weighted Avg	0.86	0.87	0.86	198147

overall accuracy remains same

Confusion Matrix

Actual/Predicted	Class 0	Class 1	Class 2
Class 0	**11424**	2663	276
Class 1	2082	**134149**	7825
Class 2	382	13285	**26061**

Reduction in misclassification

WavLM model fine-tuned on (LibriSpeech ⟶ DIHARD III)

Inference performed with a chunk of 0.6 s using the WavLM model on **LibriSpeech** dataset on a chunk of 0.6 s

(0.6, 0.6)

Class (#speakers)	Precision	Recall	F1-Score	Support (#occurrences)
0	1.00	0.04	0.08	6329
1	0.87	0.85	0.86	38237
2	0.84	0.99	0.91	40141
Overall (Accuracy)			**0.86**	84707
Macro Avg	0.90	0.63	0.62	84707
Weighted Avg	0.87	0.86	0.82	84707

overall accuracy falls

Confusion Matrix

Actual/Predicted	Class 0	Class 1	Class 2
Class 0	248	**4336**	1745
Class 1	1	**32443**	5793
Class 2	0	398	**39743**

WavLM model fine-tuned on (DIHARD III ⟶ LibriSpeech)

Inference performed with a chunk of 0.6 s using the WavLM model on **DH3** dataset on a chunk of 0.6 s

(0.6, 0.6)

Class (#speakers)	Precision	Recall	F1-Score	Support (#occurrences)
0	0.42	0.96	0.58	14363
1	0.84	0.88	0.86	144056
2	0.81	0.27	0.41	39728
Overall (Accuracy)			**0.76**	198147
Macro Avg	0.69	0.71	0.62	198147
Weighted Avg	0.80	0.76	0.75	198147

overall accuracy falls

Confusion Matrix

Actual/Predicted	Class 0	Class 1	Class 2
Class 0	**13857**	449	57
Class 1	14773	**126809**	2474
Class 2	4565	24421	**10742**

WavLM model fine-tuned on (DIHARD III ⟶ LibriSpeech)

Inference performed with a chunk of 0.6 s using the WavLM model on **LibriSpeech** dataset on a chunk of 0.6 s

(0.6, 0.6)

Class (#speakers)	Precision	Recall	F1-Score	Support (#occurrences)
0	0.97	0.22	0.36	6329
1	0.88	0.92	0.90	38237
2	0.91	0.98	0.95	40141
Overall (Accuracy)			**0.90**	84707
Macro Avg	0.92	0.71	0.73	84707
Weighted Avg	0.90	0.90	0.88	84707

overall accuracy remains same

Confusion Matrix

Actual/Predicted	Class 0	Class 1	Class 2
Class 0	**1383**	4251	695
Class 1	42	**35088**	3107
Class 2	0	620	**39521**

Summary

Fine-tuned on	Tested on	Overall average accuracy	Average Precision	Average Recall	Average F1-score
LibriSpeech (AISG)	LibriSpeech	**0.91**	**0.93**	**0.75**	**0.79**
	DIHARD III	**0.69**	**0.59**	**0.62**	**0.61**
DIHARD III	LibriSpeech	*Yet to perform inference (expecting poor performance)*			
	DIHARD III	0.87	0.83	0.79	0.81
LibriSpeech → DIHARD III	LibriSpeech	**0.86**	**0.90**	**0.63**	**0.62**
	DIHARD III	**0.87**	**0.83**	**0.79**	**0.81**
DIHARD III → LibriSpeech	LibriSpeech	0.90	0.92	0.71	0.73
	DIHARD III	0.76	0.69	0.71	0.62

Average metrics denote the specific average metric for all classes: 0 speaker, 1 speaker and >2 speakers

Next week's task

1. Combined fine-tuning on DIHARD III + LibriSpeech dataset rather than sequential (facing memory issues)