

# HW1: Singer Classification using SSL foundation models

Name: 汪宣甫  
Student ID: R11944074

Code and checkpoints' [link](#)

# Outline

- Motivation
- Method
  - Data preprocessing
  - SSL feature extractor
- Experiments
  - Top1 and Top3 acc
  - Confusion matrix
- Conclusion

# Motivation

- One of my research interests is self-supervised methods for speech.
- Recent research has suggested that SSL models are omnipotent in several downstream tasks (ASR, SID, ...) [1].
- Similarly, in music processing, there are exist also powerful SSL models [2].
- So in this homework, I want to examine those SSL models in the real work application.

# Method - data preprocessing

- Using hybrid method of demucs [3] to separate the singer voice from the song.
- Then I use [4] to remove silence from the singer voice, to make training more efficient.
- Lastly, for each song I segment it into 5 seconds of snippets.

# Method - SSL feature extractor

- Using pretrained MERT-330M [2] as feature extractor.
- MERT-330M has one 1D convolution layer, followed by 24 layers of transformer encoders. MERT-330M is trained on a large scale dataset, consisted of 160k hours of music.
- After extracting features from MERT-330M, those features are fed into a single MLP layer to predict singers' classes.
- There are two strategies to extract features from MERT. One is freeze the MERT but only train the weights that weighted sum all 25 hidden states.
- Another is taking the last hidden state of MERT and finetuning it.

# Experiments - top1 and top3 acc

- Suffix 'origin' means that we do not perform data preprocessing to the data.
- I also compare the SOTA audio SSL model, WavLM [5], in this table.

Model type	top 1	top 3
<i>Weighted sum</i>		
MERT	86.1	97.5
WavLM Large	78.2	89.9
<i>Finetune</i>		
MERT	<b>90.8</b>	<b>97.5</b>
MERT-origin	71.2	82.5

# Experiments - confusion matrix

- I use the best model, MERT with finetuning to plot the confusion matrix.
- Picture is at the next page.






# Conclusion

- I found that fine-tuning can lead to better performance than freezing the SSL model, the result is different from [1]. They suggest that whether it is frozen has no big difference.
- Although WavLM is not pre-training on music data, it still has acceptable performance. The result might be due to the fact that we remove the instrumental sounds of the inputs, so the inputs are mainly consisted of vocals.
- From the confusion matrix, we can see that the model failed to differentiate the band, Radiohead from U2. I conjecture that they have similar styles (they are both Britpop) of performing music.

# Reference

- [1] Yang, Shu-wen, et al. "Superb: Speech processing universal performance benchmark." *arXiv preprint arXiv:2105.01051* (2021).
- [2] Li, Yizhi, et al. "MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training." *arXiv preprint arXiv:2306.00107* (2023).
- [3] Défossez, Alexandre. "Hybrid spectrogram and waveform source separation." *arXiv preprint arXiv:2111.03600* (2021).
- [4] Unsilence git link: [lagmoellertim/unsilence: Console Interface and Library to remove silent parts of a media file](https://github.com/lagmoellertim/unsilence)  (github.com)
- [5] Chen, Sanyuan, et al. "Wavlm: Large-scale self-supervised pre-training for full stack speech processing." *IEEE Journal of Selected Topics in Signal Processing* 16.6 (2022): 1505-1518.