Multi-modal Music Transcription System

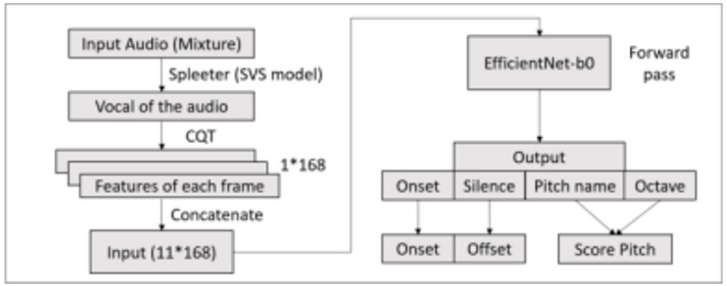
© Copyright National University of Singapore. All Rights Reserved.

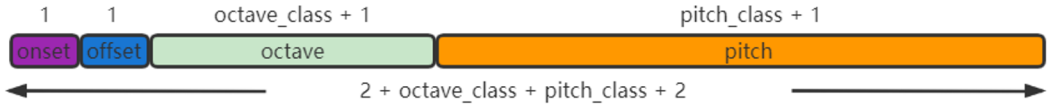
Automatic Melody Transcrption

© Copyright National University of Singapore. All Rights Reserved.

AMT Assignment: Original Model Output

- **ONE** Nerual Network, **ONE** OutTensor to predict 4 targets





$$loss = onset\_loss + offset\_loss + octave\_loss + pitch\_loss \quad (weight?)$$

$$loss = k_1 * onset\_loss + k_2 * offset\_loss + k_3 * octave\_loss + k_4 * pitch\_loss$$

— Hard to **judge the best factor** (ki) of 4 different target

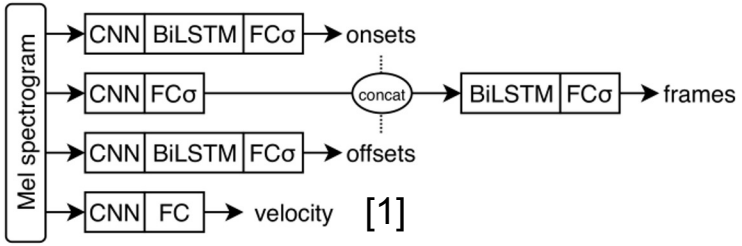
— Hard for model to predict 4 target **best at same time**

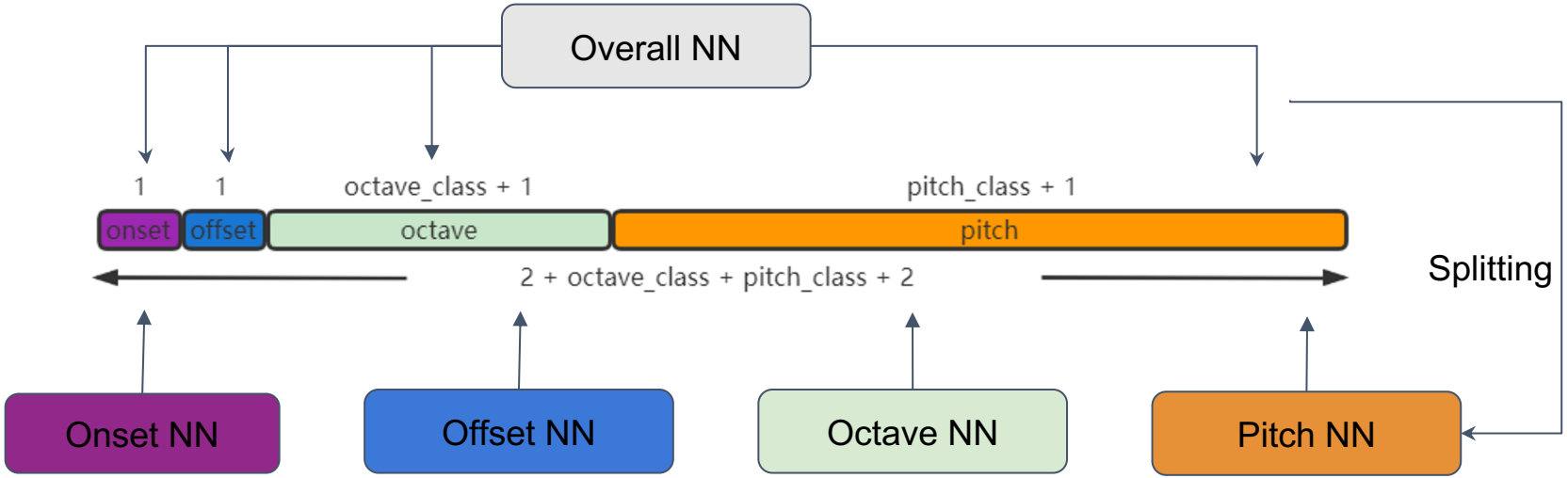
[1] J. -Y. Wang and J. -S. R. Jang, "On the Preparation and Validation of a Large-Scale Dataset of Singing Transcription," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

© Copyright National University of Singapore. All Rights Reserved.

Improvement: Model Splitting

IDEA: split different model to predict different metrics





[1] Github: BShakhovsky/PolyphonicPianoTranscription: Recurrent Neural Network for generating piano MIDI-files from audio (MP3, WAV, etc.) (github.com)

© Copyright National University of Singapore. All Rights Reserved.

Some improvement tricks

- *MIR-ST500 Dataset: only take annotation for vocal part ?*

  Use extract tool (E.G. **demucs**) to extract vocal part !

- *Large-scale dataset, quick optimize and get worse soon ?*

  May skip some well-performed training model during epoch

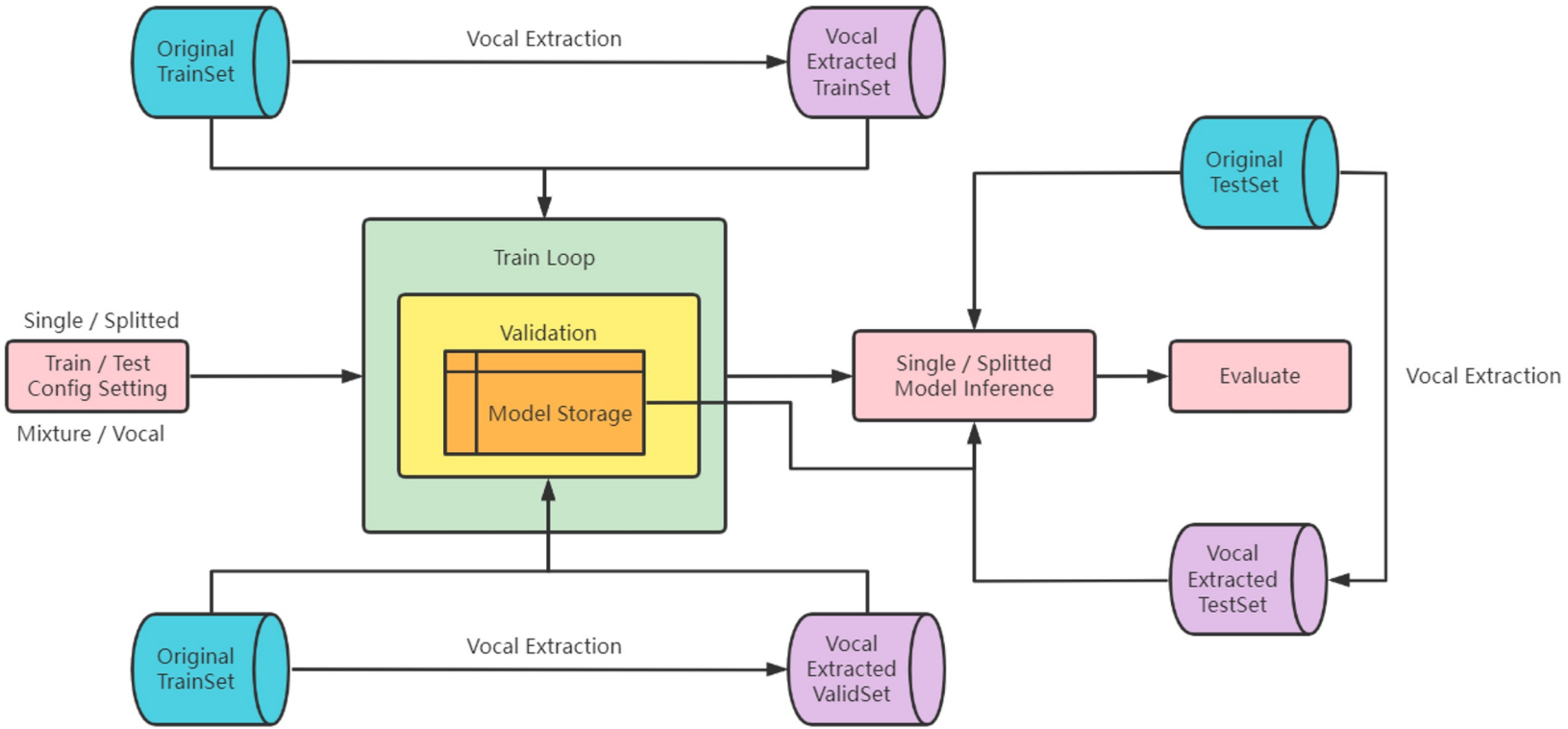
  Valid and save model after training **certain amount of data** !

  For splitted model, **save best individually** and **combine** !

```
python inference.py --save_model_dir ./models7_vocal
       --best_model_id_onset 59 --best_model_id_offset 70
       --best_model_id_octave 19 --best_model_id_pitch 34
```

© Copyright National University of Singapore. All Rights Reserved.

Melody System Implementation



© Copyright National University of Singapore. All Rights Reserved.

Metrics Performance (TestSet)

EffNetb0 F1-Score MIR-ST500 [1]		
	ConPOff	0.4578
	COnP	0.6663 ↓
	COn	0.7544 ↓

Single Model + **Mixture** Audio **MIR-ST500**	Metric	Precision	Recall	F1-Score
	ConPOff	0.334695	0.409107	0.366952
	COnP	0.569715	0.703664	0.627377
	COn	0.635402	0.784728	0.699628

Single Model + **Vocal** Audio **MIR-ST500**	Metric	Precision	Recall	F1-Score
	ConPOff	0.402688	0.403332	0.402210
	COnP	0.693781	0.703302	**0.697048** ↑
	COn	0.762988	0.772290	**0.765943** ↑

Splitted Model + **Mixture** Audio **MIR-ST500**	Metric	Precision	Recall	F1-Score
	ConPOff	**0.357746** ↑	0.405023	0.378657
	COnP	**0.605917** ↑	0.692694	**0.644244** ↑
	COn	**0.679555** ↑	0.775688	**0.721938** ↑

Splitted Model + **Vocal** Audio **MIR-ST500**	Metric	Precision	Recall	F1-Score
	ConPOff	**0.402671**	**0.423197**	**0.411709** ↑
	COnP	**0.674778**	**0.718016**	**0.694057** ↑
	COn	**0.742230**	**0.788608**	**0.762811** ↑

· **Comparison Between Model Combination:** Better **Precision (6%) (Mixture),** obvious improvement when **using mixture**

· **Comparison Between Audio Input:** Significant improvement on **COnP, COnPOff** with vocal, better than **EffNetb0**

· **Splitted Model+Vocal Audio Structure:** Perform **best F1-Score** of **COnPOff (0.4117, 12.2% higher than baseline)** !

[1] J. -Y. Wang and J. -S. R. Jang, "On the Preparation and Validation of a Large-Scale Dataset of Singing Transcription," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

© Copyright National University of Singapore. All Rights Reserved.

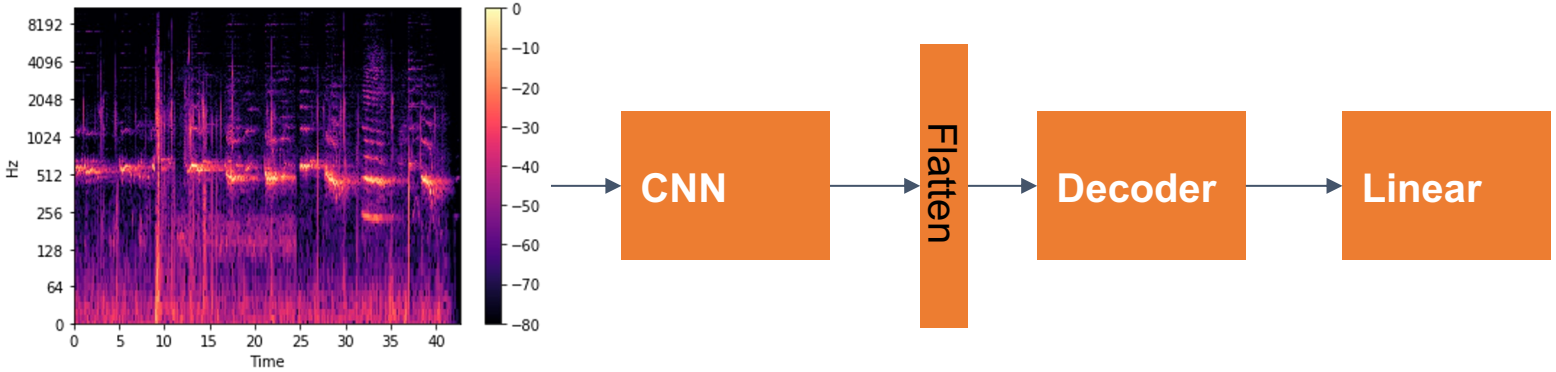
Automatic Lyric Transcrption

© Copyright National University of Singapore. All Rights Reserved.

ASR: Mel-Spectrogram Based

1. **Mel-Spectrogram + CRDNN(CNN, RNN, DNN)**
   Spectrogram (raw feature) + CNN (feature extractor) + RNN (decoder) + DNN (linear)

1. **Mel-Spectrogram + CNN + Transformer**
   Spectrogram (raw feature) + CNN (feature extractor) + Transformer (decoder) + DNN (linear)



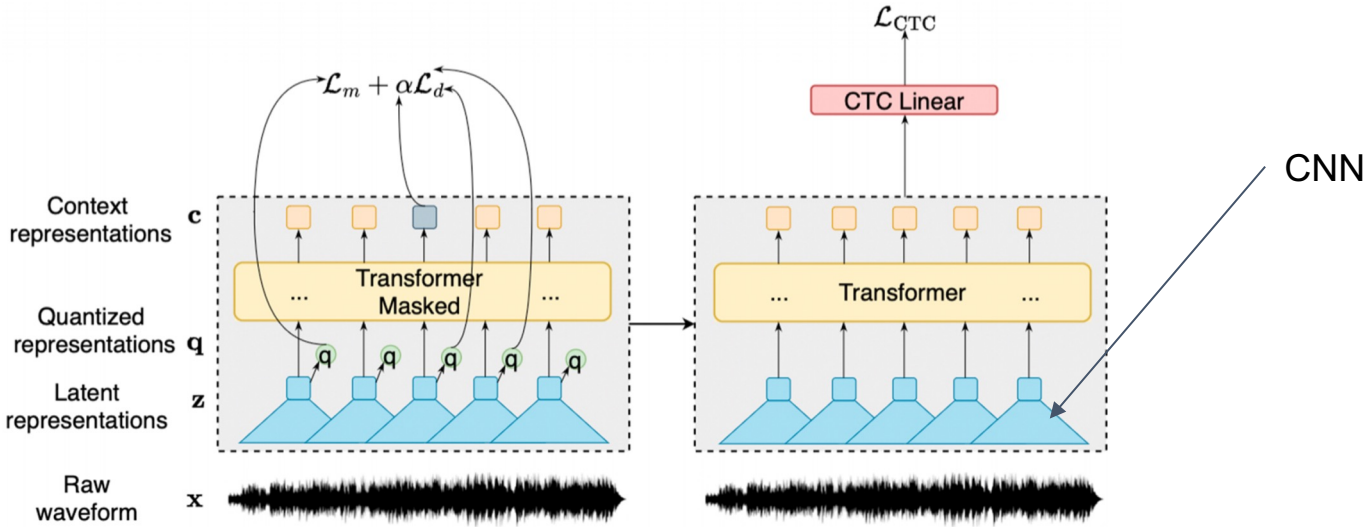
© Copyright National University of Singapore. All Rights Reserved.

ASR: End-to-End Wac2Vec 2.0

- Training wav2vec 2.0 has two stages:

Stage I: Self-supervised Contrastive Learning

Stage II: Supervised Fine-tuning



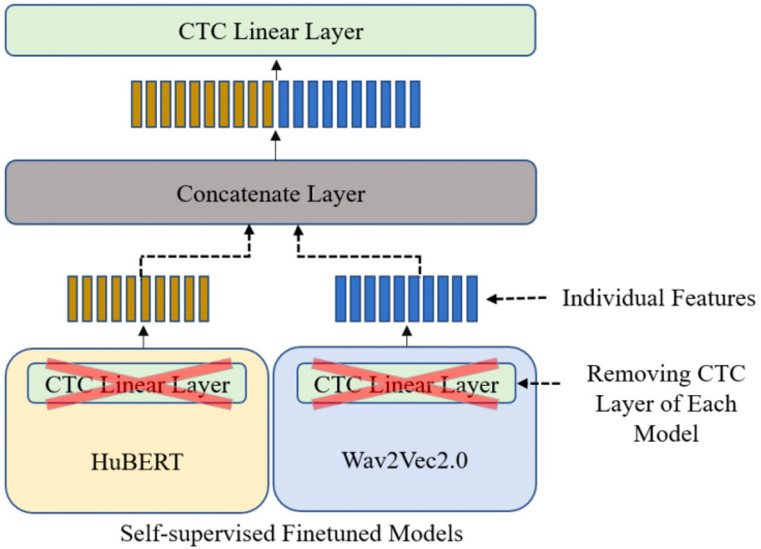
[1] Gu X, Ou L, Ong D, Wang Y. TRANSFER LEARNING OF WAV2VEC 2.0 FOR AUTOMATIC LYRIC TRANSCRIPTION[C]//Proceedings of the 30th ACM International Conference on Multimedia. 2022.

© Copyright National University of Singapore. All Rights Reserved.

Ensemble Methods on ASR

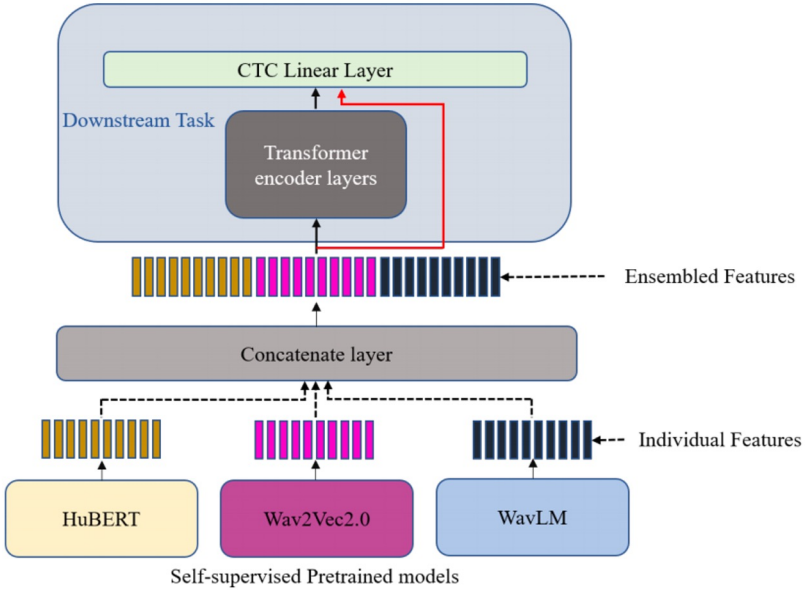
1. Model Ensemble

Concat before CTC Linear



2. Feature Ensemble

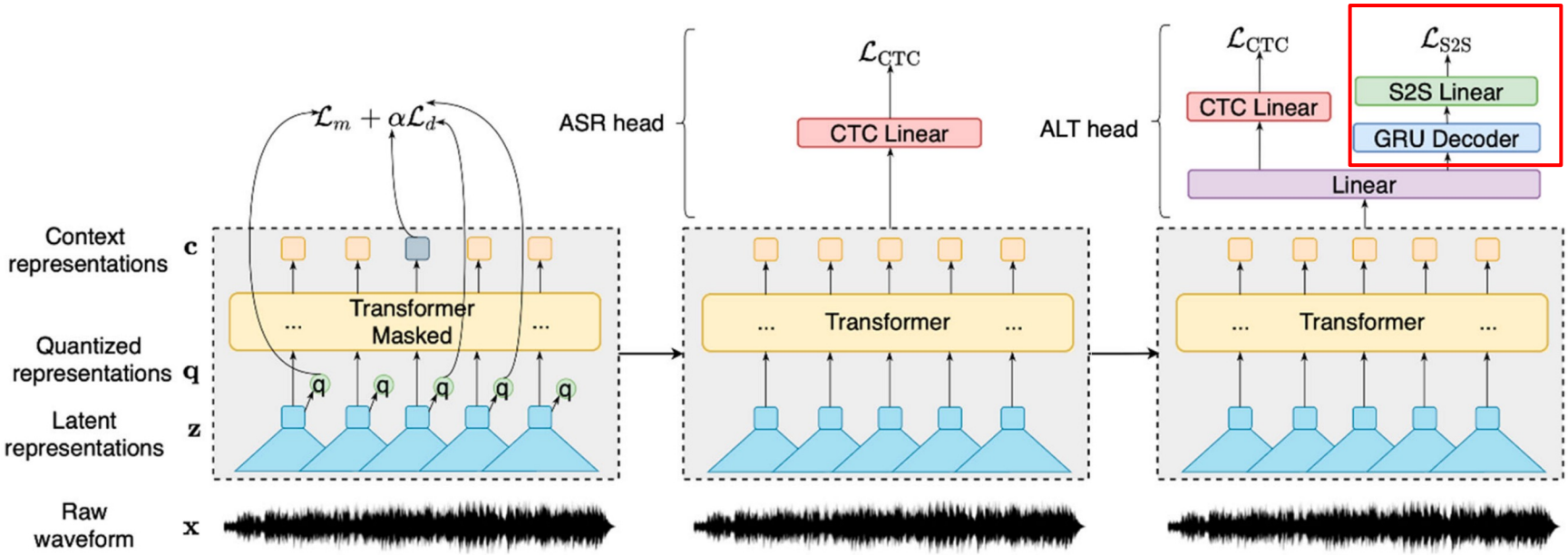
Concat before Decoder



[1] *A Arunkumar , Vrunda N Sukhadia, S. Umesh* Investigation of Ensemble features of Self-Supervised Pretrained Models for Automatic Speech Recognition[C] International Speech Communication Association

© Copyright National University of Singapore. All Rights Reserved.

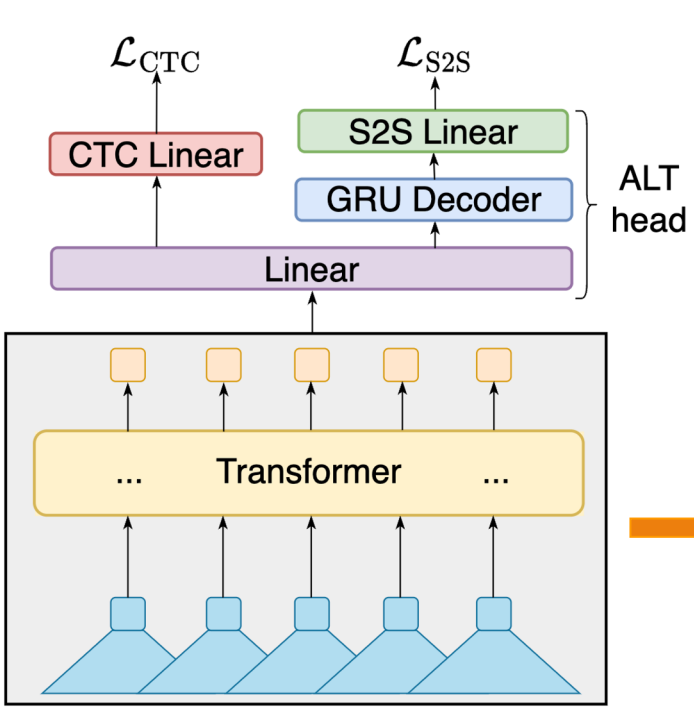
Transfer Learning to ALT

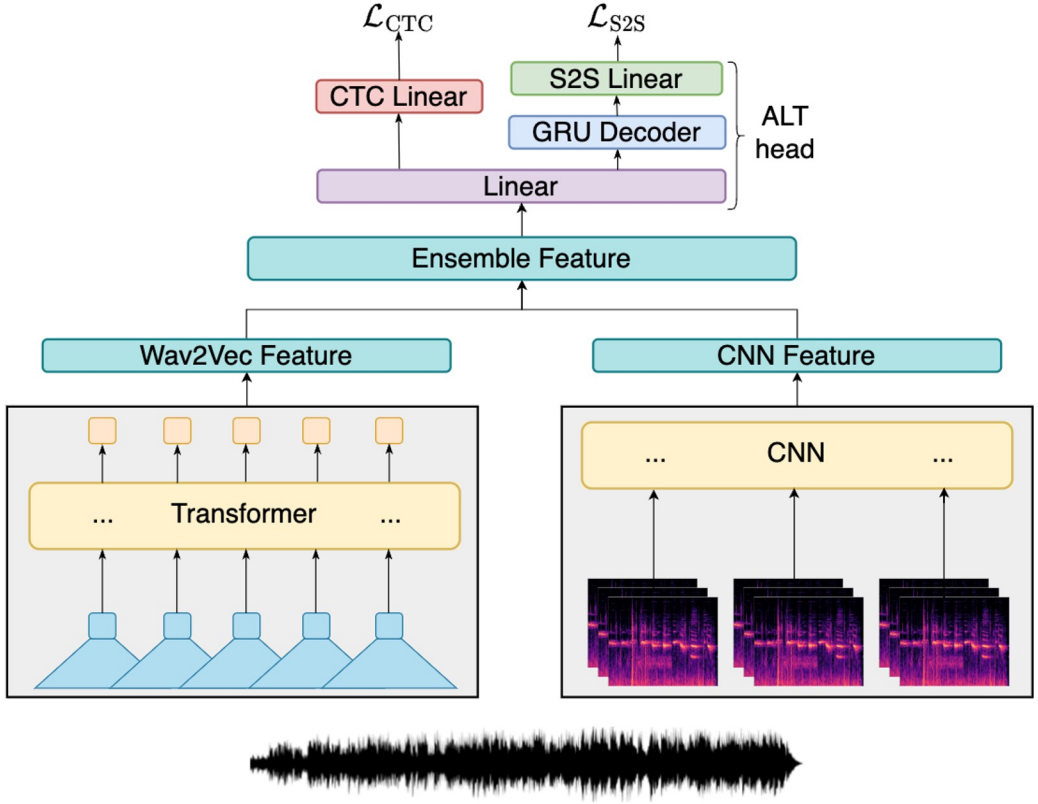


[1] Gu X, Ou L, Ong D, Wang Y. TRANSFER LEARNING OF WAV2VEC 2.0 FOR AUTOMATIC LYRIC TRANSCRIPTION[C]//Proceedings of the 30th ACM International Conference on Multimedia. 2022.

© Copyright National University of Singapore. All Rights Reserved.

Ensemble Features on ALT(Ours)





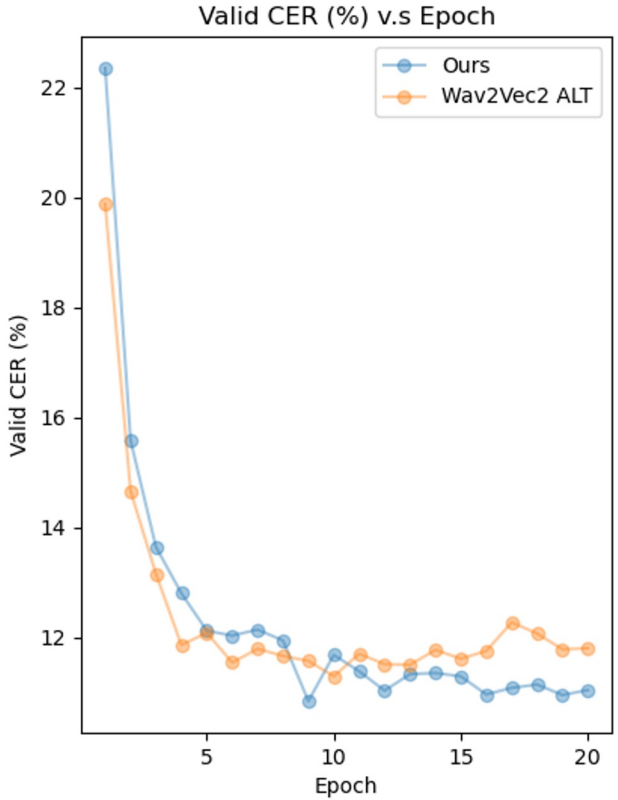
© Copyright National University of Singapore. All Rights Reserved.

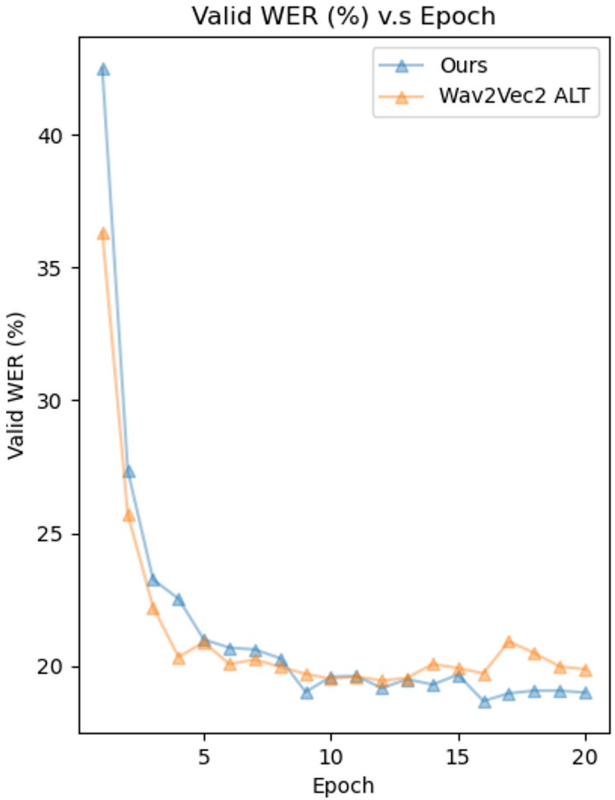
Performance

Model \ Metric	Language Model	CER (%)	WER (%)	SER (%)
Wav2vec2+ALT head	RNNLM trained on DSing1	12.28	21.66	64.38
Wav2Vec 2.0 + CNN +ALT Head(Ours)		10.66	19.29	61.06
Error Rate Decrease By		**13.2%**	**10.9%**	**5.2%**
Wav2vec2+ALT head	RNNLM trained on DSing30 (10 times larger than Dsing1)	11.35	17.75	50.83
Wav2Vec 2.0 + CNN +ALT Head(Ours)		10.83	17.30	49.58
Error Rate Decrease By		**4.5%**	**2.5%**	**2.5%**

© Copyright National University of Singapore. All Rights Reserved.

Performance





© Copyright National University of Singapore. All Rights Reserved.

Model Parameters

ALT Model	Model Parameters
Wav2Vec 2.0	316.5M
ALT Head	89.4M
Wav2Vec 2.0 + ALT Head(Current SOTA)	405.9 M
Wav2Vec 2.0 + CNN + ALT Head(Ours)	408.7 M
Model Increase By	**0.5% (2.8M)**

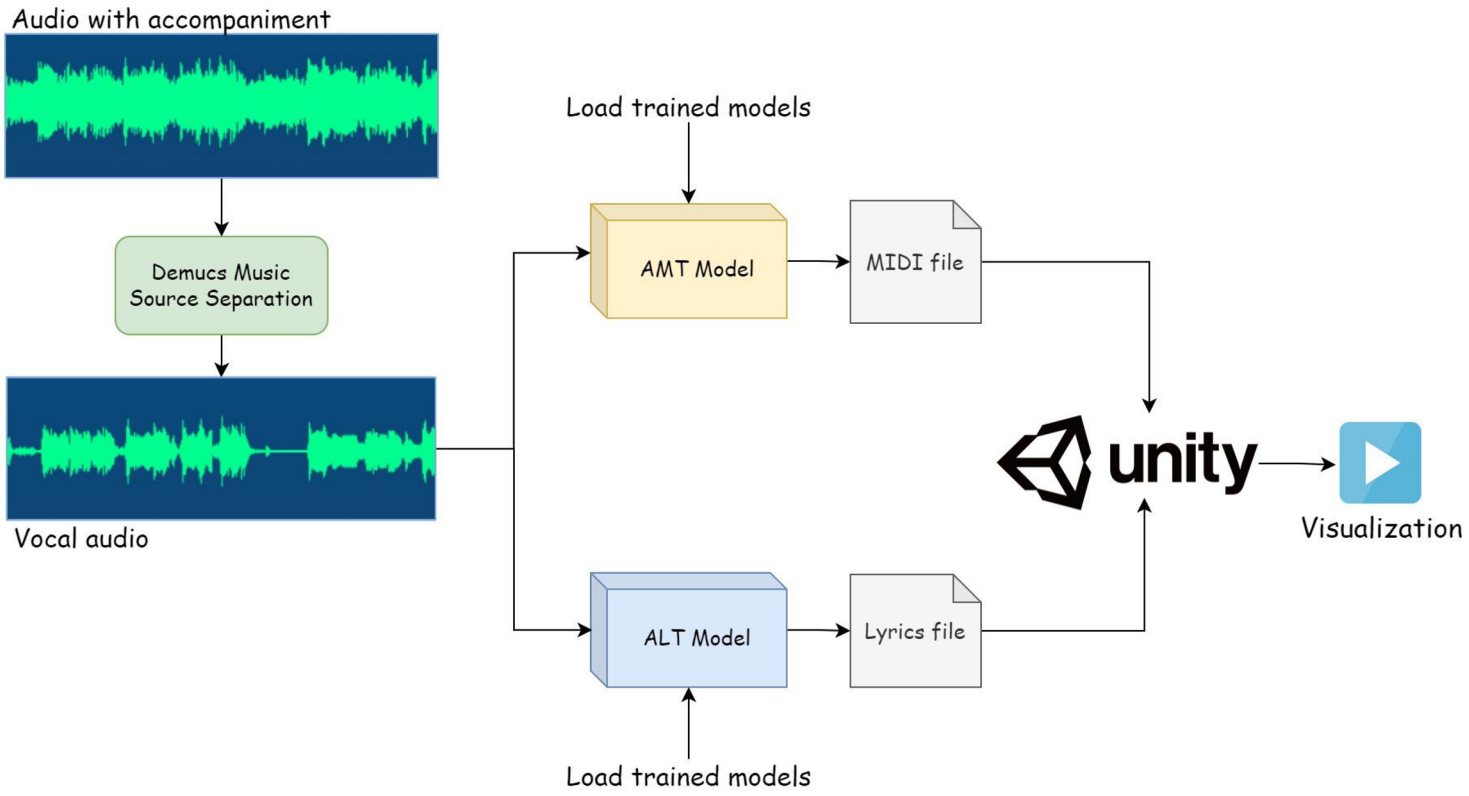
Our model improves the performance with only a slight increase of parameters!

© Copyright National University of Singapore. All Rights Reserved.

Visuialization

© Copyright National University of Singapore. All Rights Reserved.

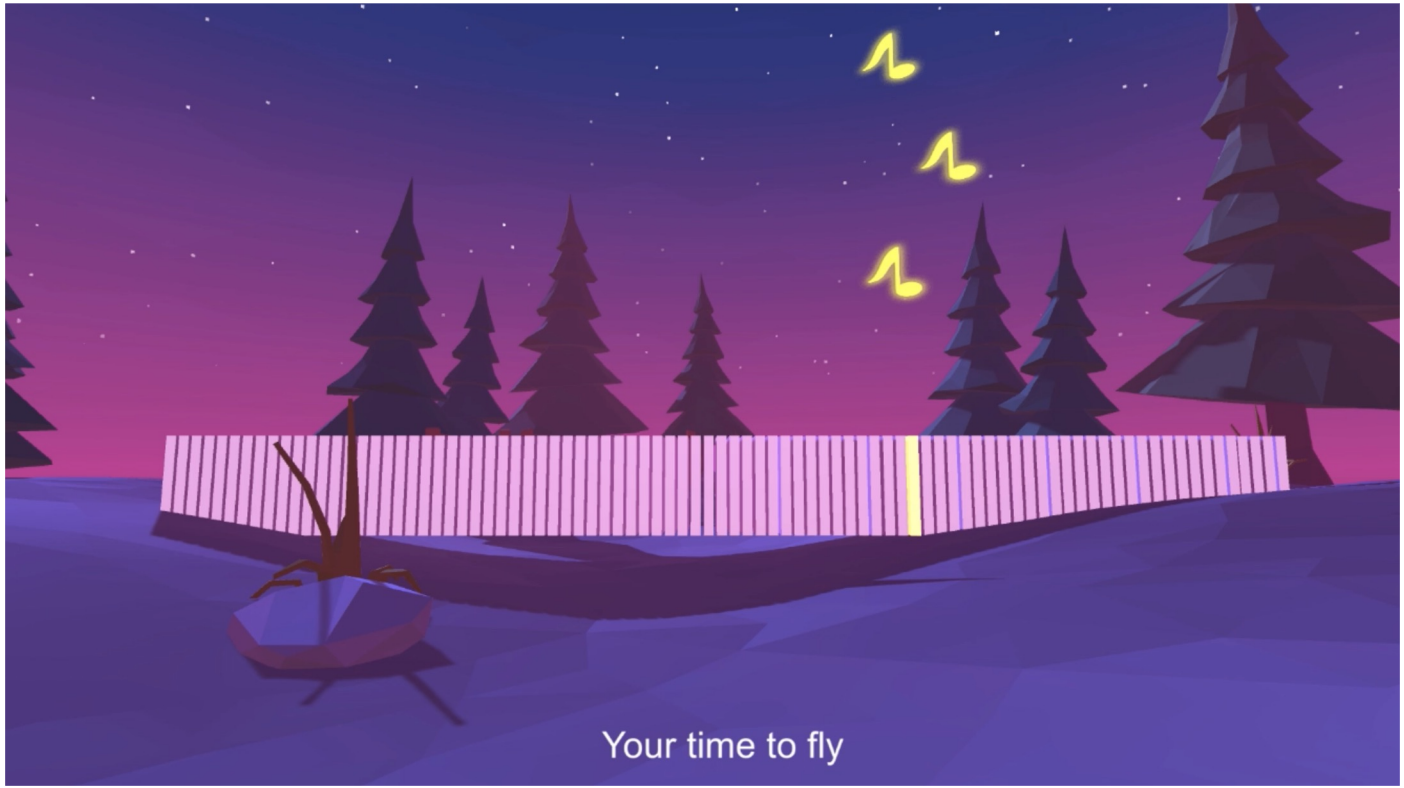
System Workflow



© Copyright National University of Singapore. All Rights Reserved.

refer: https://github.com/facebookresearch/demucs

Visualize pitch and lyrics



© Copyright National University of Singapore. All Rights Reserved.

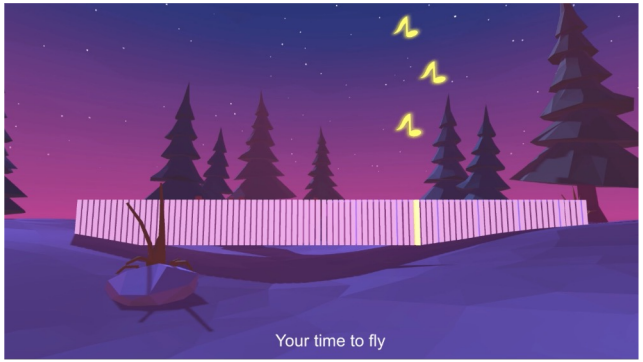
Visualize pitch and lyrics

Melody

- 88 keys
- Input: 2d array
- [[start, end, pitch],...[start, end, pitch]]
- generate a note when time to start

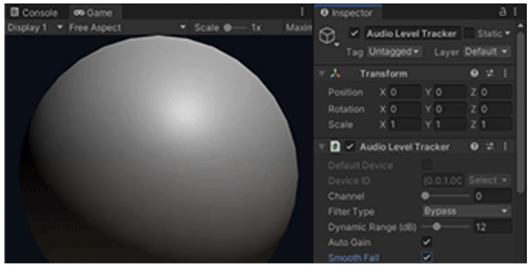
Lyrics

- Input: 2d array
- [[start, end,lyrics],...[start, end,lyrics]]

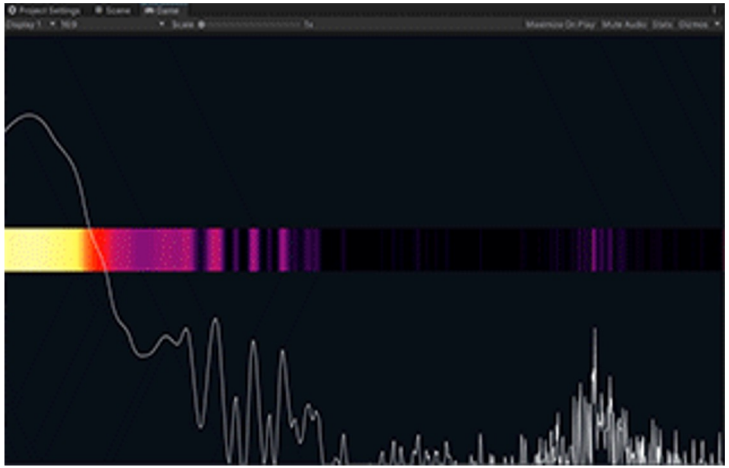


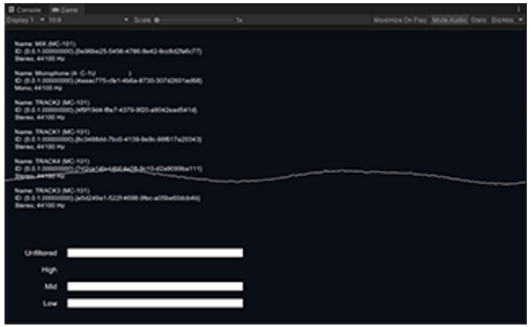
© Copyright National University of Singapore. All Rights Reserved.

Audio Active visual



sphere scaled by the **audio level**



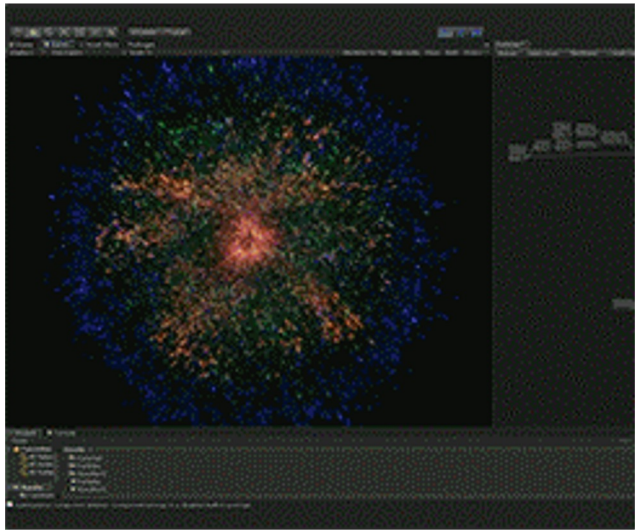


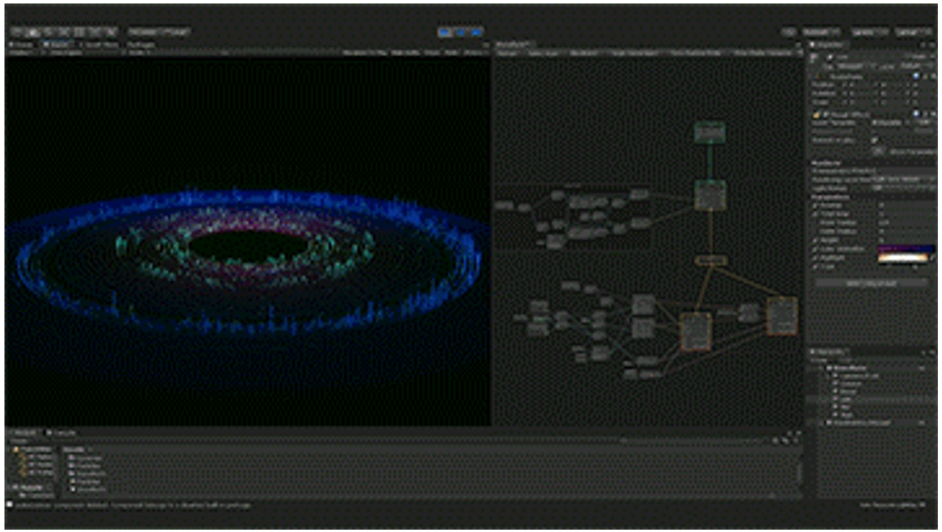
draw **frequency** band levels with raw wave function

get the audio **frequency spectrum** of an input stream

© Copyright National University of Singapore. All Rights Reserved.

Audio Active visual





Customize particle system provided by Visual Effact Graphics in Unity

© Copyright National University of Singapore. All Rights Reserved.

Demonstration

© Copyright National University of Singapore. All Rights Reserved.

Import Audio

So if you ever feel like giving up

THANK YOU

© Copyright National University of Singapore. All Rights Reserved.