HW 1 - singer classification

Name: 譚至斌

ID: d12942015

Novelty

- Only the vocal is used for this task
  - The assumption is in pop songs, the musical elements are shared by different singer, but the sining voice is unique however.

Methodology

- The solution can be described by 4 steps:

    1. Separate vocals from song audio with ‘*demucs*’

    2. Split vocal audio on silence into 5-seconds segments (padding zeros if it’s not long enough)

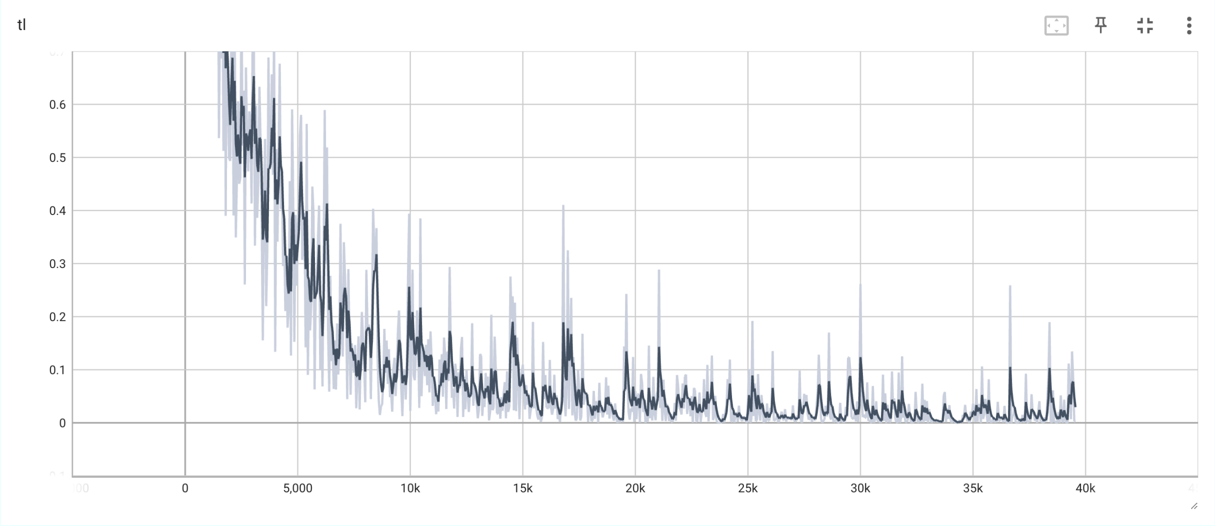
    3. In the training stage, training and validating on each spliced frame

    4. In the prediction stage, use all predicted results from each frame of one song (by voting) to decide which singer it belongs to

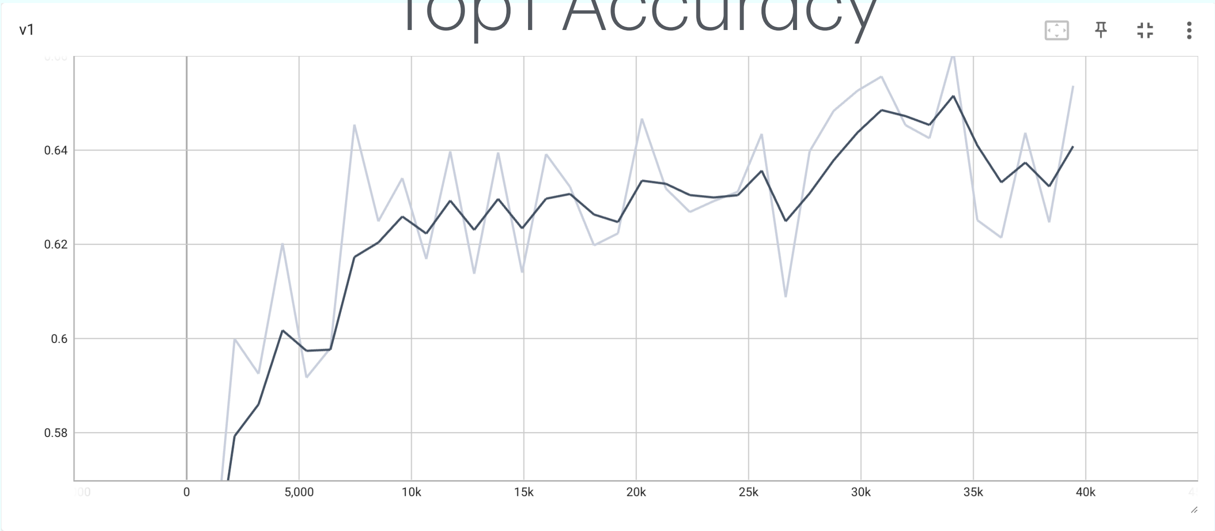
Result

- As the training loss decreased, the highest point of accuracies of top1 and top3 are ~**0.66** and ~**0.84**, respectively.

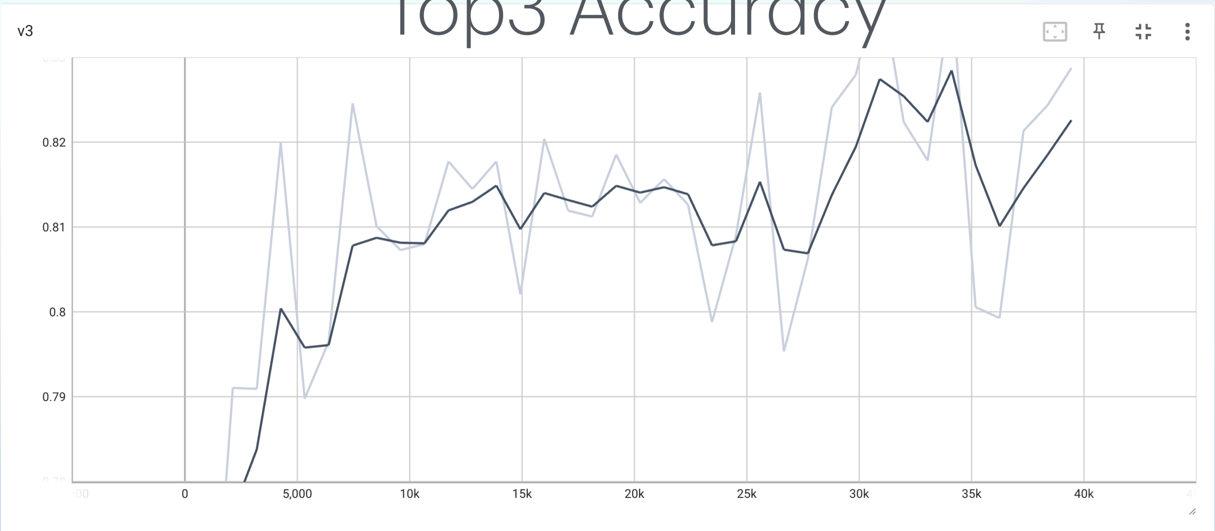
Training Loss



Top1 Accuracy



Top3 Accuracy

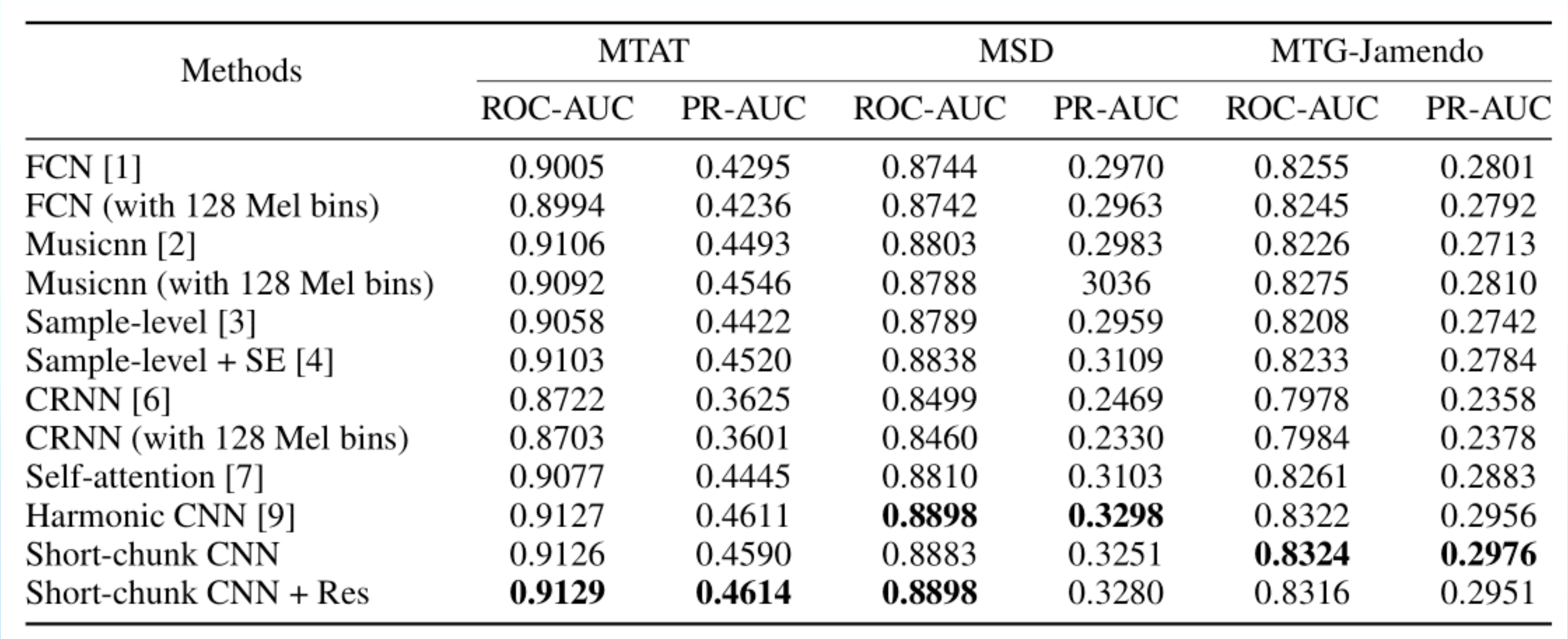


Findings

- Source separation tools is useful
- MelSpectrogram is better than normal spectrogram while processing musical data

Details of my approach

- My approach is based on the survey, I find all the NN architectures in this survey have nearly the same performance (<5 points)



Methods	MTAT		MSD		MTG-Jamendo	
	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
FCN [1]	0.9005	0.4295	0.8744	0.2970	0.8255	0.2801
FCN (with 128 Mel bins)	0.8994	0.4236	0.8742	0.2963	0.8245	0.2792
Musicnn [2]	0.9106	0.4493	0.8803	0.2983	0.8226	0.2713
Musicnn (with 128 Mel bins)	0.9092	0.4546	0.8788	3036	0.8275	0.2810
Sample-level [3]	0.9058	0.4422	0.8789	0.2959	0.8208	0.2742
Sample-level + SE [4]	0.9103	0.4520	0.8838	0.3109	0.8233	0.2784
CRNN [6]	0.8722	0.3625	0.8499	0.2469	0.7978	0.2358
CRNN (with 128 Mel bins)	0.8703	0.3601	0.8460	0.2330	0.7984	0.2378
Self-attention [7]	0.9077	0.4445	0.8810	0.3103	0.8261	0.2883
Harmonic CNN [9]	0.9127	0.4611	**0.8898**	**0.3298**	0.8322	0.2956
Short-chunk CNN	0.9126	0.4590	0.8883	0.3251	**0.8324**	**0.2976**
Short-chunk CNN + Res	**0.9129**	**0.4614**	**0.8898**	0.3280	0.8316	0.2951

Details of my approach

- There 2 SOTA models, Harmonic CNN and Short-chunk CNN. Each of them outperform on different datasets.

- As the survey describe, Harmonic CNN is best for generalization ability. However, I think singer detection is a **tagging task focus on voice**. (which is also one of my assumption)

- I choose Short-chunk CNN as the backend model.

Discussion

- I notice increase batch size from 32 to 64 slightly decrease the validation accuracy (but still in statistical error)

- MelSpectrogram also increase the performance slightly compared to ordinary spectrogram (still in statistical error)

- I think currently the development of machine learning techniques have achieved a good stage. It can help people to **eliminate the gap of some feature-extracting techniques** (like MelSpec vs Spec).

- From my observation, the most important thing is how to **apply the domain knowledge** to the problem we want to solve.

Appendix

- source code: https://github.com/tanchihpin0517/NTU/tree/master/deepmir/hw1

- model: https://www.dropbox.com/scl/fi/gkgzrvntww5rc5jtcp4h2/epoch-028-tl-0.00-v1-0.66-v3-0.84-step-30914.ckpt?rlkey=vdwerwwlgok4t94q8v501iw41&dl=0