Paper ID: 250

Singing Beat Tracking With Self-supervised Front-end and Linear Transformers

Mojtaba Heydari and Zhiyao Duan

Audio Information Research Lab, University of Rochester, Rochester NY, USA

23nd International Society for Music Information Retrieval Conference, ISMIR (Winter 2022)

Motivations / Contributions

➢ **Motivations:**

- Singing voice beat tracking is useful for many applications in music production, processing, analysis and interaction
- Due to the lake of percussive components and strong harmonic content, current generic music beat tracking systems don't work well for singing voice beat tracking

➢ **Contributions:**

✓ Introducing singing voice beat tracking as a novel MIR task

✓ Proposing two strategies to create annotated datasets for this task

✓ New evaluation scheme that can account for phase ambiguities

✓ Proposing two neural models for the task leveraging pre-trained speech self-supervised models and linear transformers

Proposed Approach to Create Annotated Data

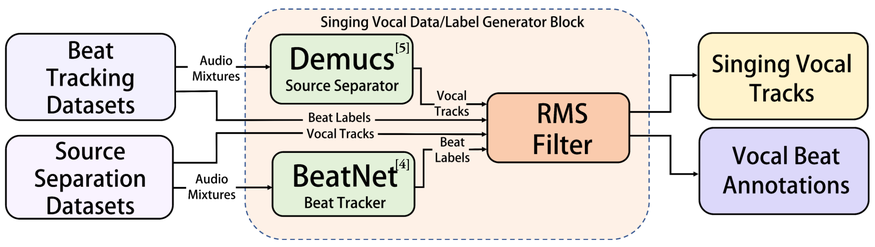

Figure 1. Singing data and label generation pipeline

Dataset	*# Number of vocal excerpts*	*Total Length*
Ballroom*	452	2 h 38 m
GTZAN *	741	5 h 44 m
Hainsworth*	154	1 h 47 m
MUSDB18 †	263	6 h 21 m
Rock Corpus*	315	9 h 23 m
RWC pop *	188	5 h 06 m
RWC Royalty free *	29	19 m
URSing †	106	3 h 17 m

Table 1. Datasets collected and adapted for singing beat tracking. ∗ denotes beat tracking datasets and † denotes music separation datasets.

Proposed Neural Network Structures

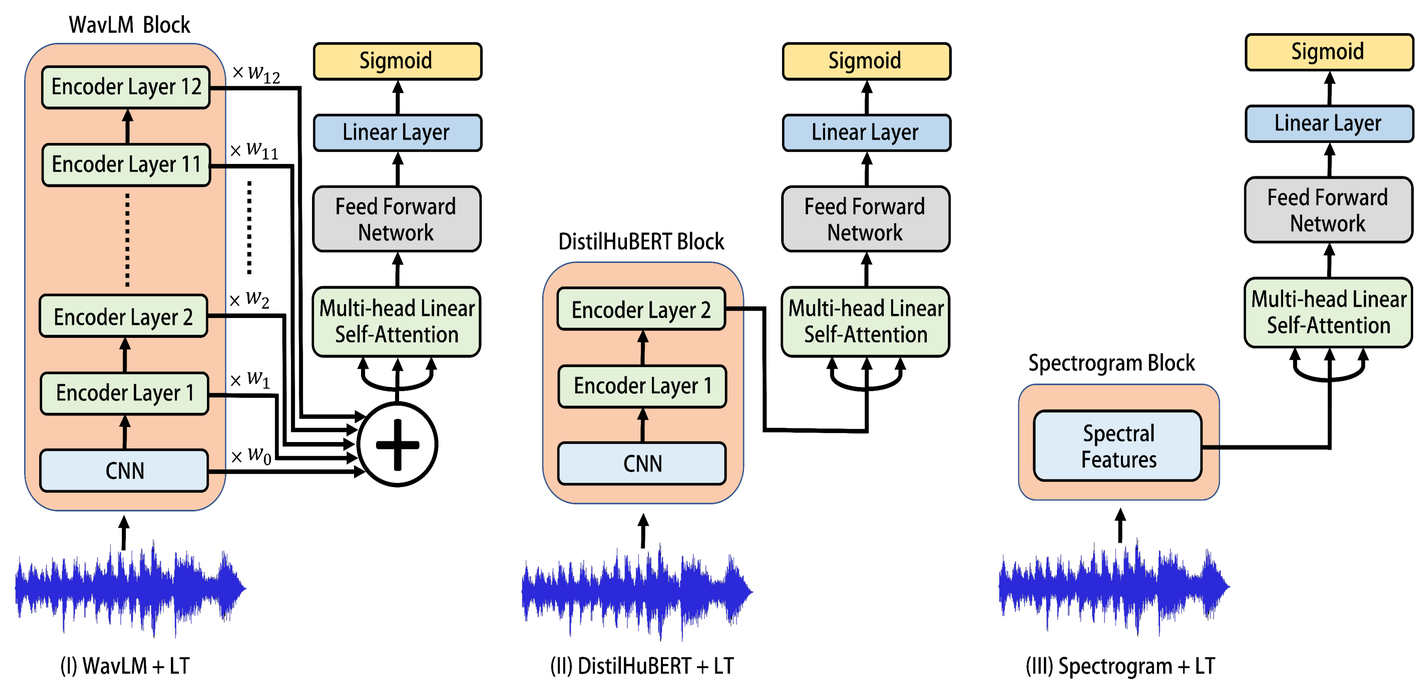

Figure 2. Neural network structures of the proposed models. (I) utilizes a weighted sum of all layers of pre-trained WavLM self-supervised speech model to calculate the feature embeddings. (II) uses the outputs of the DistilHuBERT teacher student model and (III) employs the spectrogram representations that are used in conventional beat tracking methods as input feature representations. All models take advantage of the same self-attention network to fine-tune them for the task.

Evaluation

	Method	F-Measure	P-Score	Cemgil	Goto	Comp. Time
Baseline	BeatNet [4]	0.243	0.327	0.173	0.003	0.13 (s)
	BeatRoot [6]	0.301	0.394	0.22	0.066	**0.03 (s)**
	Böck [7]	0.171	0.195	0.122	0.009	1.56 (s)
Proposed	WavLM + LT	**0.733**	**0.704**	**0.618**	**0.560**	4.09 (s)
	DistilHuBERT + LT	0.703	0.668	0.593	0.516	1.83 (s)
	Spectrogram + LT	0.454	0.438	0.367	0.223	0.32 (s)
Proposed (PI Results)	WavLM + LT	0.745	0.715	0.627	0.574	4.09 (s)
	DistilHuBERT + LT	0.721	0.684	0.608	0.537	1.83 (s)
	Spectrogram + LT	0.489	0.477	0.391	0.265	0.32 (s)

Table 2. Average performance and speed across segments of several methods on the GTZAN separated vocal tracks, including baselines and the proposed models and the Phase Inclusive (PI) evaluation for the proposed models.

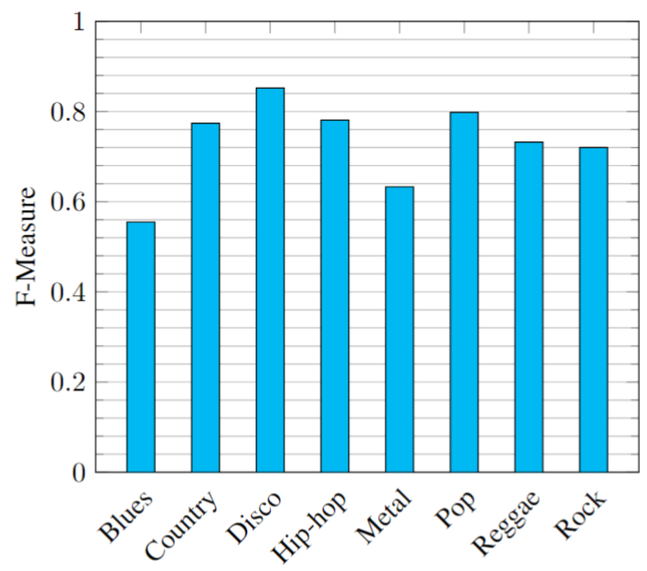

Figure 3. F-measure performance of WavLM + LT model on the GTZAN separated vocal tracks for different genres.

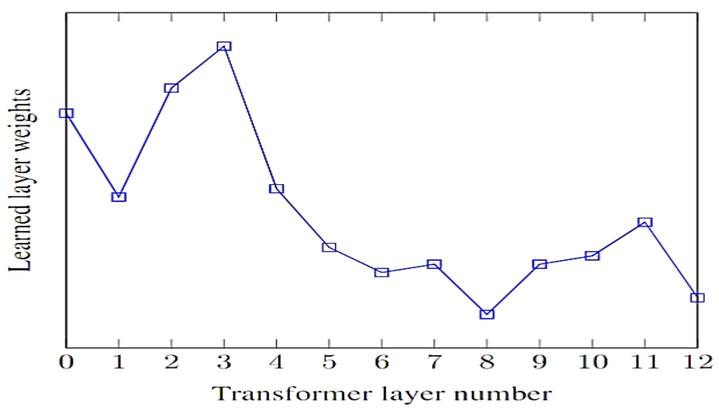

Figure 4. Learned weights for different encoder layers of WavLM in the WavLM+LT model

Acknowledgement

This work was supported by National Science Foundation grant No. 1846184.

References

1. S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," arXiv preprint arXiv:2110.13900, 2021.
2. H.-J. Chang, S.-w. Yang, and H.-y. Lee, "Distilhu- bert: Speech representation learning by layer-wise dis- tillation of hidden-unit bert," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 7087–7091.
3. A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are rnns: Fast autoregressive transform- ers with linear attention," in International Conference on Machine Learning. PMLR, 2020, pp. 5156–5165.
4. M. Heydari, F. Cwitkowitz, and Z. Duan, "BeatNet: CRNN and particle filtering for online joint beat downbeat and meter tracking," in Proc. of the 22th Intl. Conf. on Music Information Retrieval (ISMIR), 2021.
5. A. Défossez, "Hybrid spectrogram and waveform source separation," in Proceedings of the ISMIR 2021 Workshop on Music Source Separation, 2021.P. M. Brossier, "Automatic annotation of musical audio for interactive applications," pp. 58–102. Queen Marry University, London, UK, August 2006.
6. S. Dixon, "Evaluation of the audio beat tracking system beatroot," Journal of New Music Research, vol. 36, no. 1, pp. 39–50, 2007.
7. S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer, "madmom: a new Python Audio and Music Signal Processing Library," in Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 10 2016, pp. 1174– 1178.