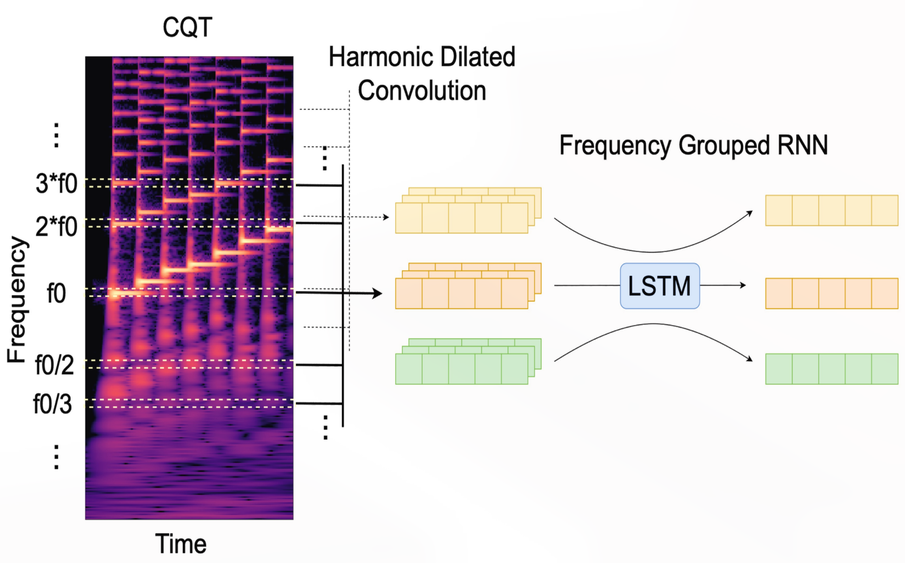
HPPNET :

MODELING THE HARMONIC STRUCTURE AND PITCH INVARIANCE IN PIANO TRANSCRIPTION

Weixin Wei
Peilin Li
Wei Li
Yi Yu

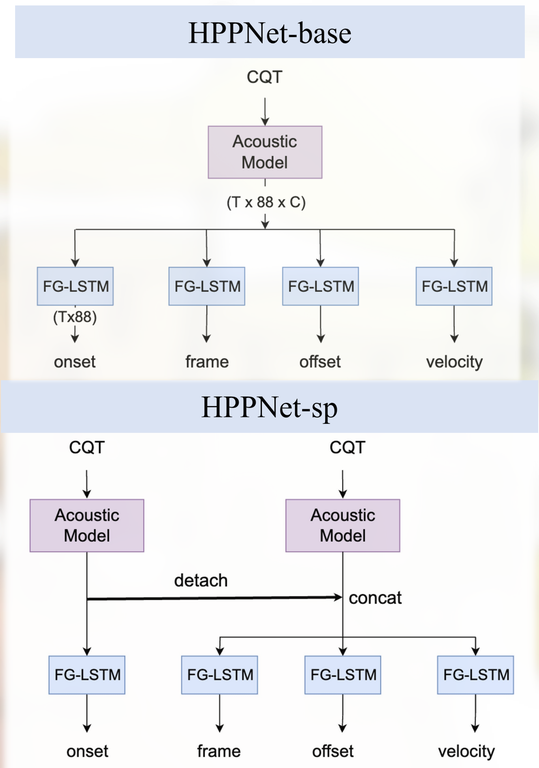
While neural network models are making significant progress in piano transcription, they are becoming more resource-consuming due to requiring larger model size and more computing power. In this paper, we attempt to apply more prior about piano to reduce model size and improve the transcription performance.
The sound of a piano note contains various overtones, and the pitch of a key does not change over time.
To make full use of such latent information, we propose HPPNet that using the harmonic dilated convolution to capture the harmonic structures and the frequency grouped recurrent neural network to model the pitch-invariance over time. Experimental results on the MAESTRO dataset show that our piano transcription system achieves state-of-the-art performance both in frame and note scores (frame F1 93.15\%, note F1 97.18\%). Moreover, the model size is much smaller than the previous state-of-the-art deep learning models.

Method



- Acoustic model with harmonic dilation convolution
  To capture harmonic information in a more efficient way, the Harmonic Dilated Convolution (HD-Conv) set customized dilation rates in the log-frequency dimension. We feed the CQT into multiple dilated convolution layers with different dilation rates and sum the outputs for the following layers.

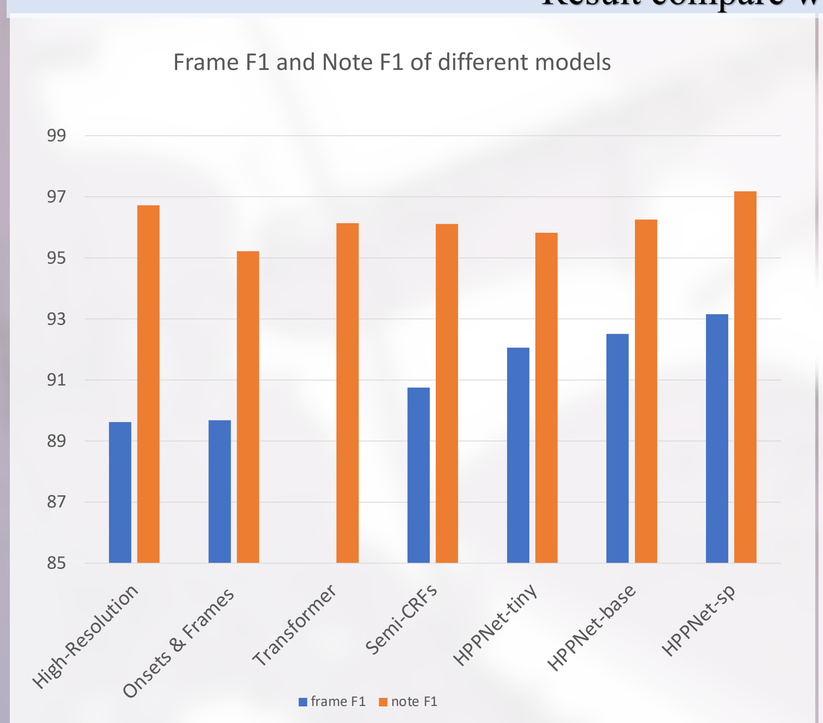
- Frequency Grouped LSTM
  To lightweight our model, we first segment the output of harmonic dilated convolution to 88 frequency according piano keyboard. Then we feed each frequency group to the same LSTM layer individually to model the temporal relationship. We term this the Frequency Grouped LSTM (FG-LSTM)

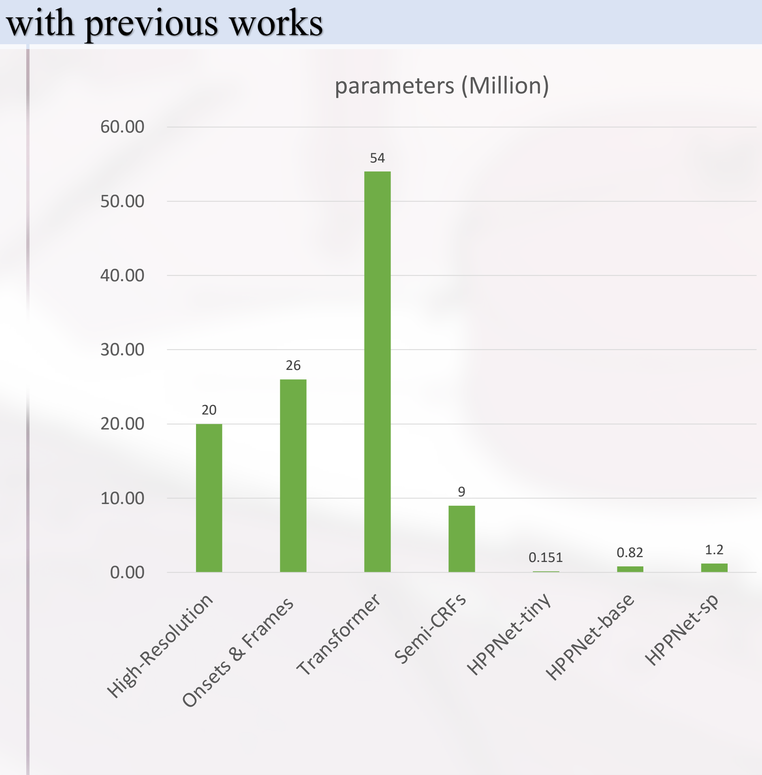


Experimental Result

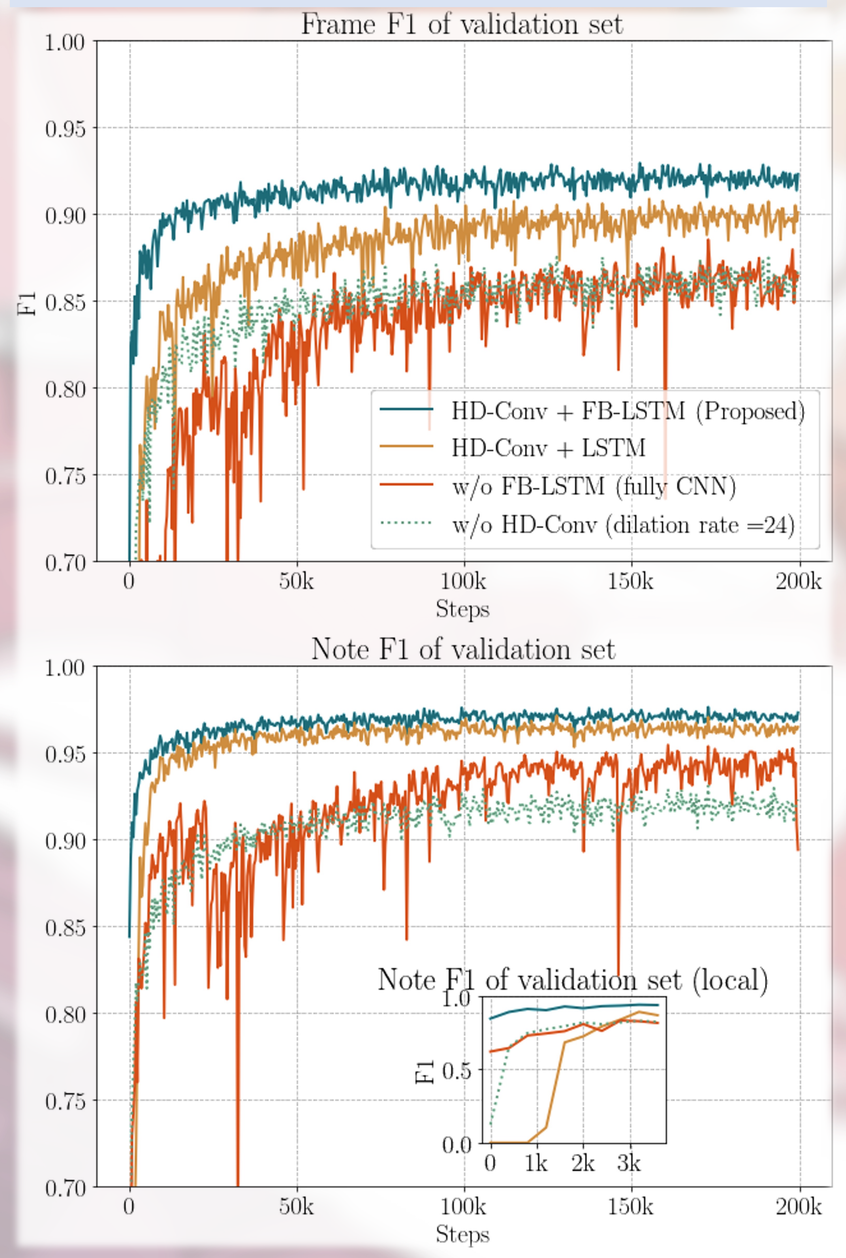
We compare HPPNet with some existing state-of-the-art methods using the MAESTRO dataset and the MAPS dataset. Then, ablation studies are done to demonstrate the affects of FG-LSTM, and HD-Conv. We also examine the model performance on small datasets. The success stems from two aspects: (i) the harmonic dilated convolution exploited to capture harmonics structure; and (ii) the frequency grouped LSTM designed based on the pitch-invariance of piano key over time.

Result compare with previous works





F1 of MAESTROv3 validation set.



Result training was done on the MAESTRO v3 train split and evaluation result on MAPS test split

Model	Params	FRAME			NOTE			NOTE W/OFFSET			NOTE W/OFFSET & VEL.		
		P (%)	R (%)	F1(%)	P (%)	R (%)	F1(%)	P (%)	R (%)	F1(%)	P (%)	R (%)	F1(%)
Onsets & Frames	26M	-	-	82.02	-	-	83.04	-	-	61.84	-	-	48.07
Onsets & Frames*	26M	92.86	78.46	84.91	87.46	85.58	86.44	68.22	66.75	67.43	52.41	51.22	51.77
HPPNet-sp	1.2M	88.42	86.81	87.56	91.61	82.38	86.63	65.01	63.84	64.39	60.35	59.26	59.77

Reference

[1] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.- Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in In- ternational Conference on Learning Representations, ICLR, 2019, pp. 1–6.
[2] J. W. Kim and J. P. Bello, "Adversarial learning for im- proved onsets and frames music transcription," in Pro- ceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR, 2019, pp. 670–677.
[3] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, "High- resolution piano transcription with pedals by regress- ing onset and offset times," IEEE ACM Transactions on Audio Speech and Language Processing. TASLP, vol. 29, pp. 3707–3717, 2021.
[4] Y. Yan, F. Cwitkowitz, and Z. Duan, "Skipping the frame-level: Event-based piano transcription with neu- ral semi-crfs," in Advances in Neural Information Pro- cessing Systems, NeurIPS, vol. 34, 2021, pp. 20 583– 20 595.
[5] C. Hawthorne, I. Simon, R. Swavely, E. Manilow, and J. H. Engel, "Sequence-to-sequence piano transcription with transformers," in Proceedings of the 22nd Interna- tional Society for Music Information Retrieval Confer- ence, ISMIR, 2021, pp. 246–253.
[6] W. Wei, P. Li, Y. Yu, and W. Li, "Harmof0: Loga- rithmic scale dilated convolution for pitch estimation," in 2022 IEEE International Conference on Multimedia and Expo, ICME, 2022, pp. 1–6.
[7] Y. Luo, C. Han, and N. Mesgarani, "Ultra-lightweight speech separation via group communication," in IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2021, pp. 16–20.