

Hello, ICASSP 2022!





About Ou Longshen

- PhD student in Sound and Music Computing Lab, National University of Singapore
- Research area: music information retrieval
 - Music / lyric transcription
 - Lyric generation
- Violin and guitar player
- Personal website: oulongshen.xyz

A close-up, shallow depth-of-field photograph of piano keys, showing the white and black keys receding into the distance. The image is partially obscured by a white curved shape on the right side.

Exploring Transformer's Potential on Automatic Piano Transcription

Longshen Ou 2022
oulongshen@u.nus.edu

Sound and Music Computing Lab
National University of Singapore

Motivation

Explore the applicability of Transformer to piano transcription problem

Other research fields, like speech processing, natural language processing, have moved towards Transformer architectures–

Can we get the performance increase on music transcription problem too?

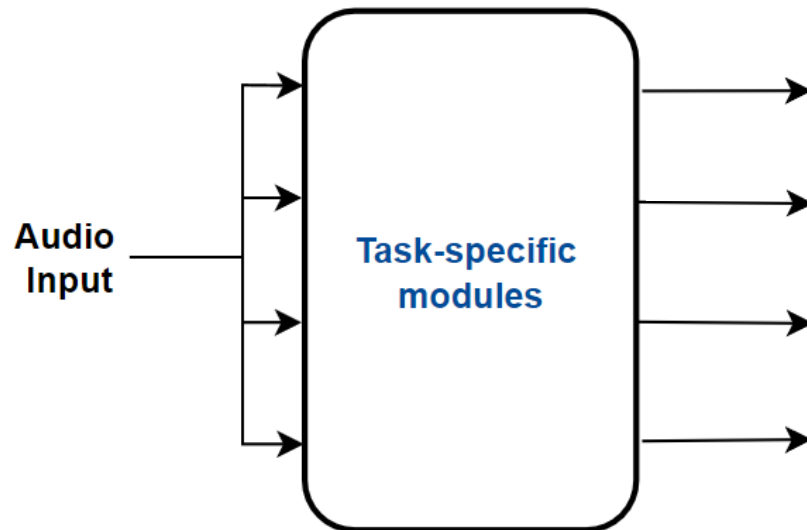
Baseline System^[1]: Overview

Audio
Input

Procedure:

[1] Kong, Qiuqiang, et al. "High-Resolution Piano Transcription With Pedals by Regressing Onset and Offset Times." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021, pp. 3707–17. *IEEE Xplore*, <https://doi.org/10.1109/TASLP.2021.3121991>.

Baseline System^[1]: Overview

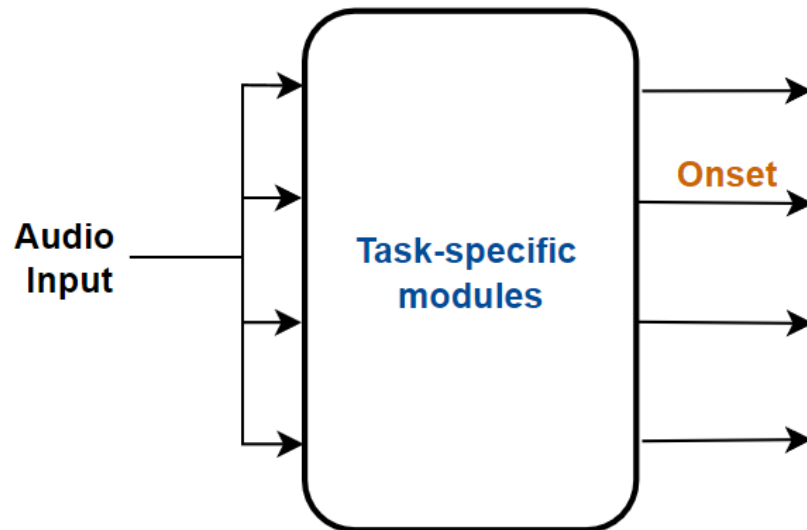


Procedure:

1. First, task-specific neural networks recognize note-element on frame-level;

[1] Kong, Qiuqiang, et al. "High-Resolution Piano Transcription With Pedals by Regressing Onset and Offset Times." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021, pp. 3707–17. *IEEE Xplore*, <https://doi.org/10.1109/TASLP.2021.3121991>.

Baseline System^[1]: Overview

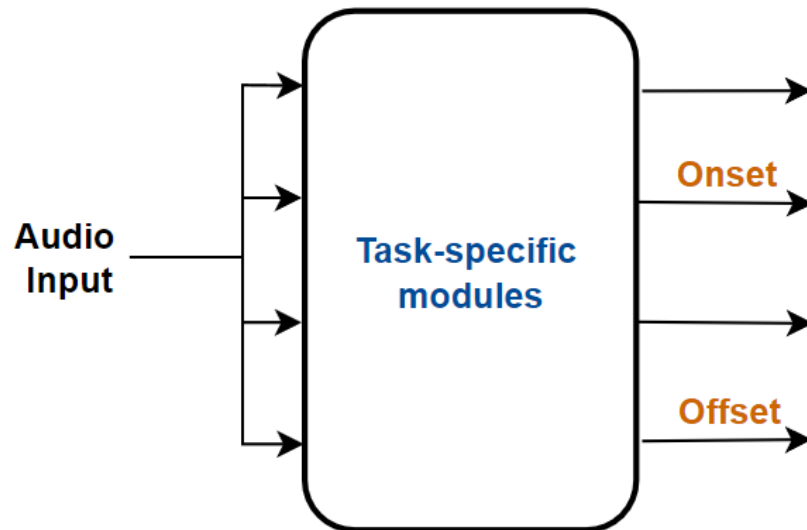


Procedure:

1. First, task-specific neural networks recognize note-element on frame-level;

[1] Kong, Qiuqiang, et al. "High-Resolution Piano Transcription With Pedals by Regressing Onset and Offset Times." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021, pp. 3707–17. *IEEE Xplore*, <https://doi.org/10.1109/TASLP.2021.3121991>.

Baseline System^[1]: Overview

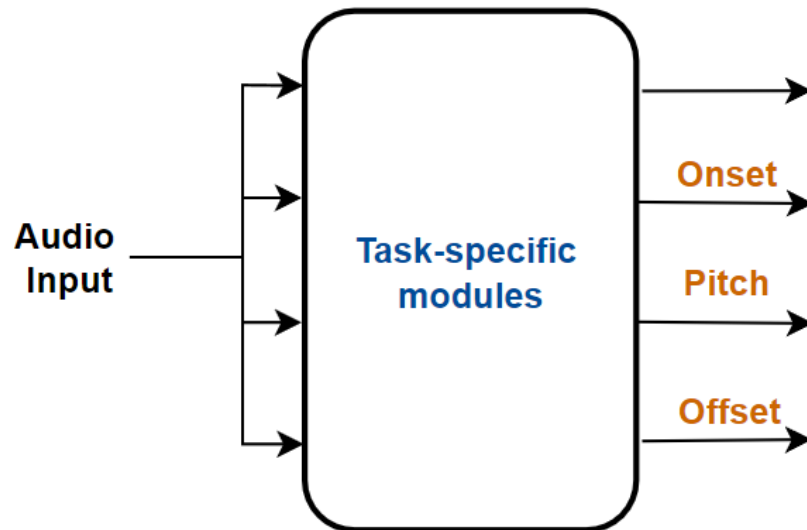


Procedure:

1. First, task-specific neural networks recognize note-element on frame-level;

[1] Kong, Qiuqiang, et al. "High-Resolution Piano Transcription With Pedals by Regressing Onset and Offset Times." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021, pp. 3707–17. *IEEE Xplore*, <https://doi.org/10.1109/TASLP.2021.3121991>.

Baseline System^[1]: Overview

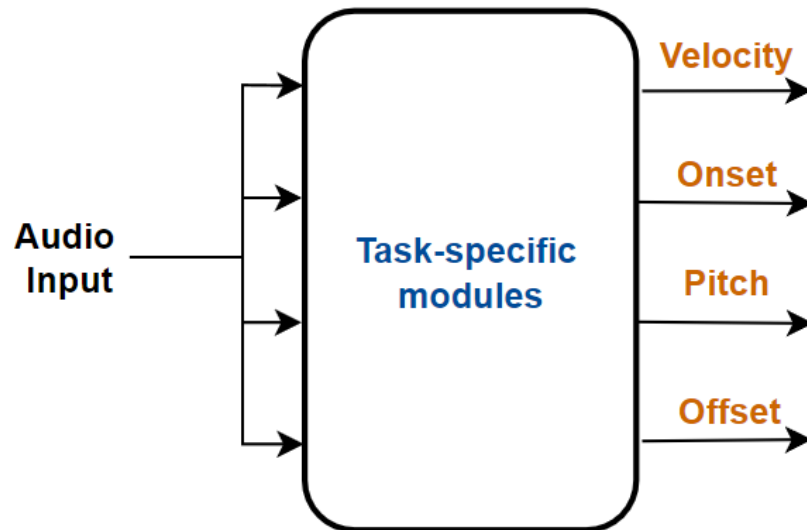


Procedure:

1. First, task-specific neural networks recognize note-element on frame-level;

[1] Kong, Qiuqiang, et al. "High-Resolution Piano Transcription With Pedals by Regressing Onset and Offset Times." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021, pp. 3707–17. *IEEE Xplore*, <https://doi.org/10.1109/TASLP.2021.3121991>.

Baseline System^[1]: Overview

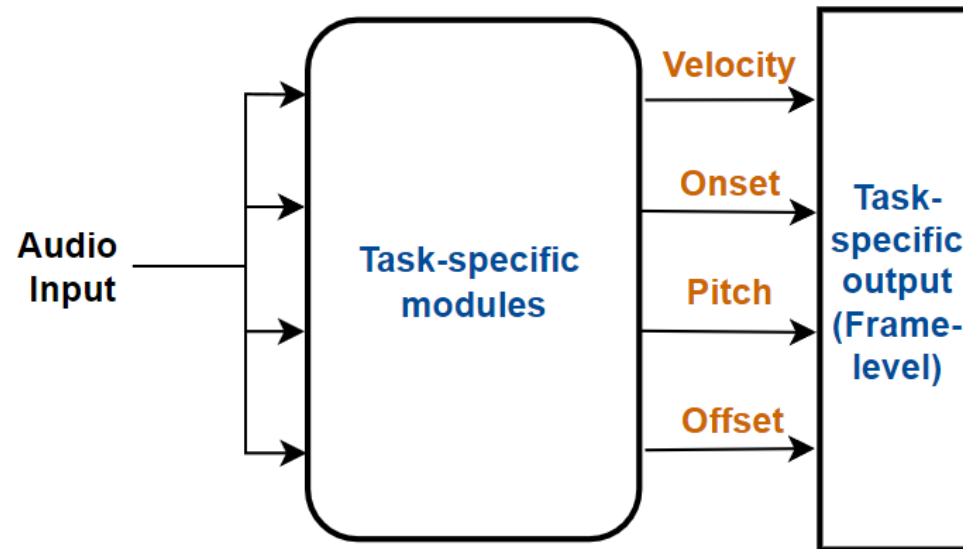


Procedure:

1. First, task-specific neural networks recognize note-element on frame-level;

[1] Kong, Qiuqiang, et al. "High-Resolution Piano Transcription With Pedals by Regressing Onset and Offset Times." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021, pp. 3707–17. *IEEE Xplore*, <https://doi.org/10.1109/TASLP.2021.3121991>.

Baseline System^[1]: Overview

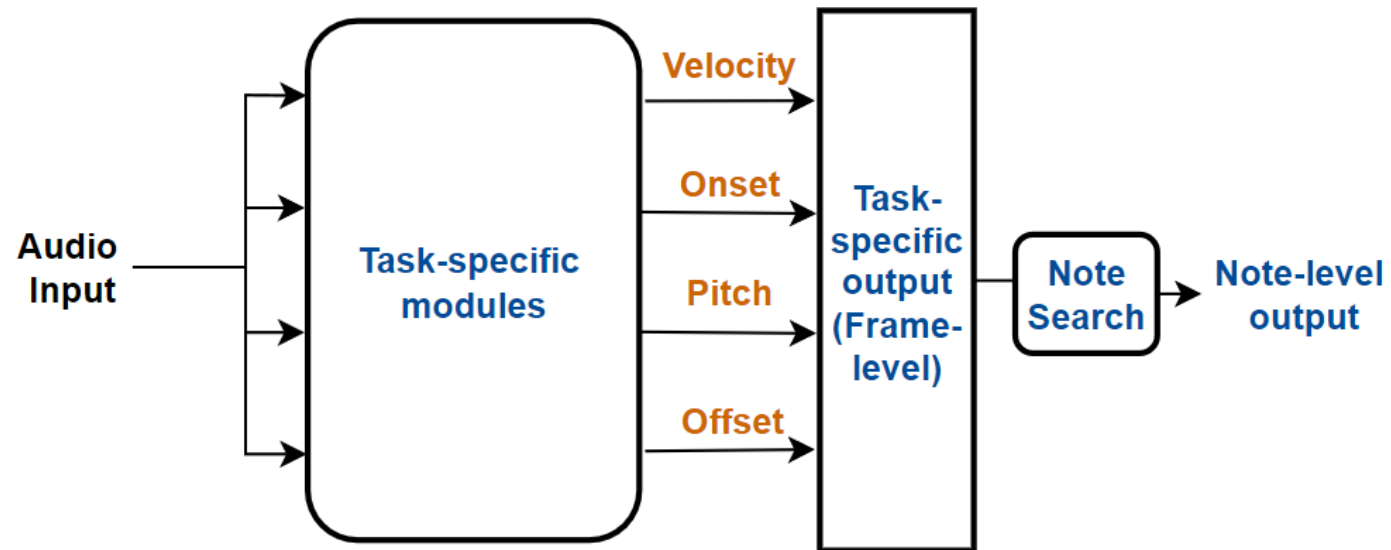


Procedure:

1. First, task-specific neural networks recognize note-element on frame-level;

[1] Kong, Qiuqiang, et al. "High-Resolution Piano Transcription With Pedals by Regressing Onset and Offset Times." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021, pp. 3707–17. *IEEE Xplore*, <https://doi.org/10.1109/TASLP.2021.3121991>.

Baseline System^[1]: Overview

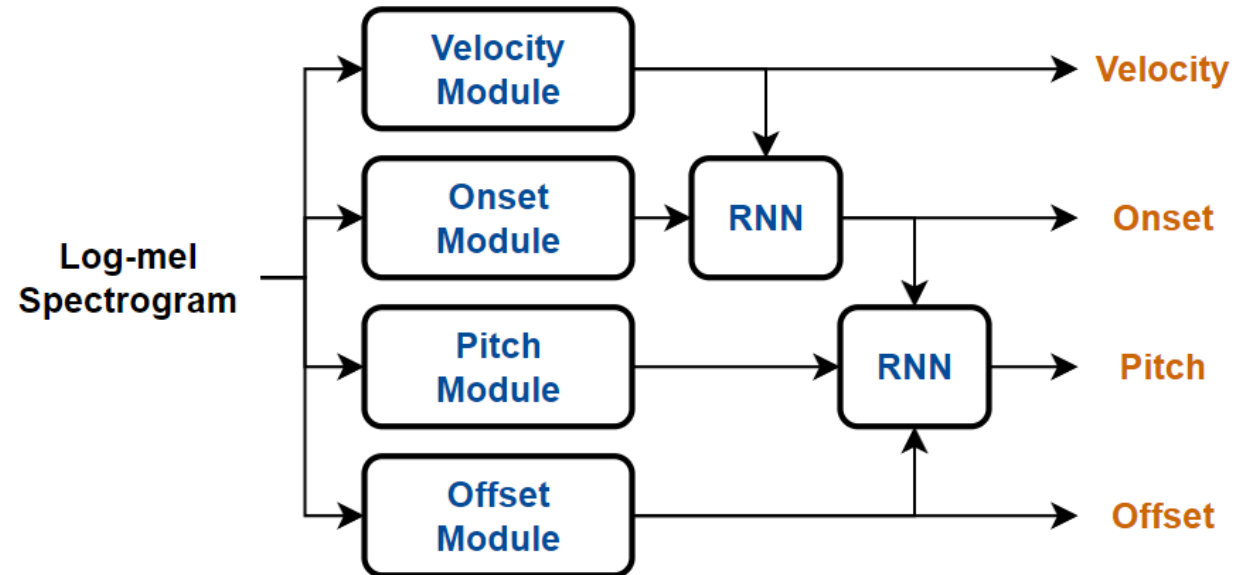


Procedure:

1. First, task-specific neural networks recognize note-element on frame-level;
2. then, note-level results are calculated through a search algorithm

[1] Kong, Qiuqiang, et al. "High-Resolution Piano Transcription With Pedals by Regressing Onset and Offset Times." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021, pp. 3707–17. *IEEE Xplore*, <https://doi.org/10.1109/TASLP.2021.3121991>.

Baseline System: NNs



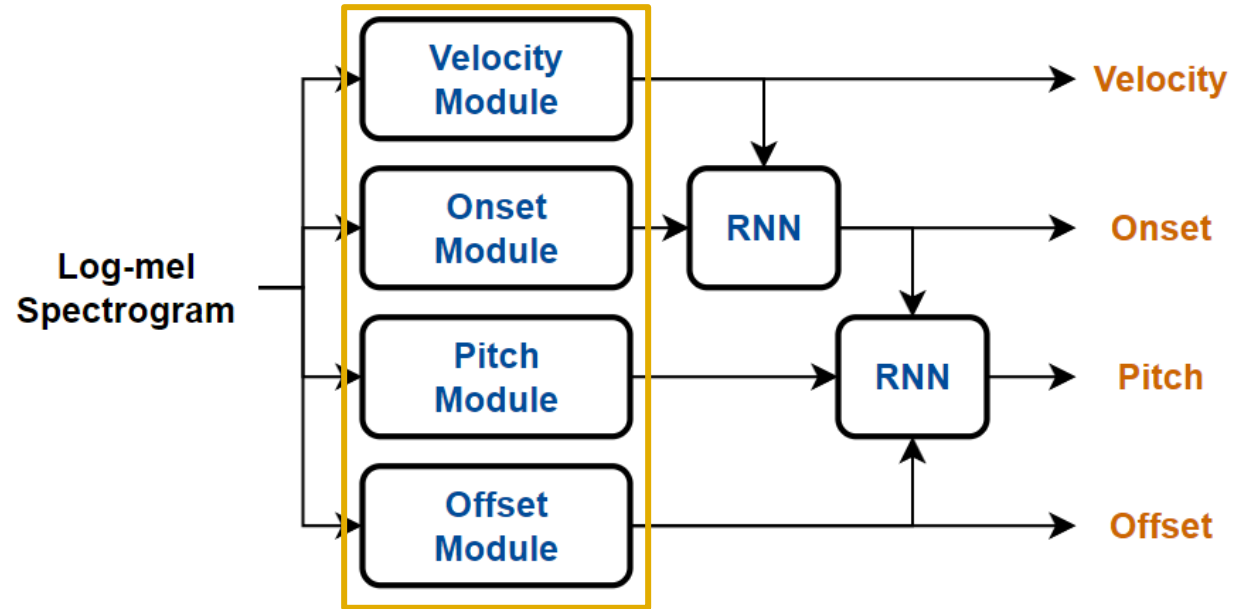
Separate network for each subtask

Network for each subtask: CNN-GRU, Same structure for every subtask

Audio input: Log-mel spectrogram

Output of each branch: Does this element exist, on a specific timestep, of a specific pitch

Baseline System: NNs



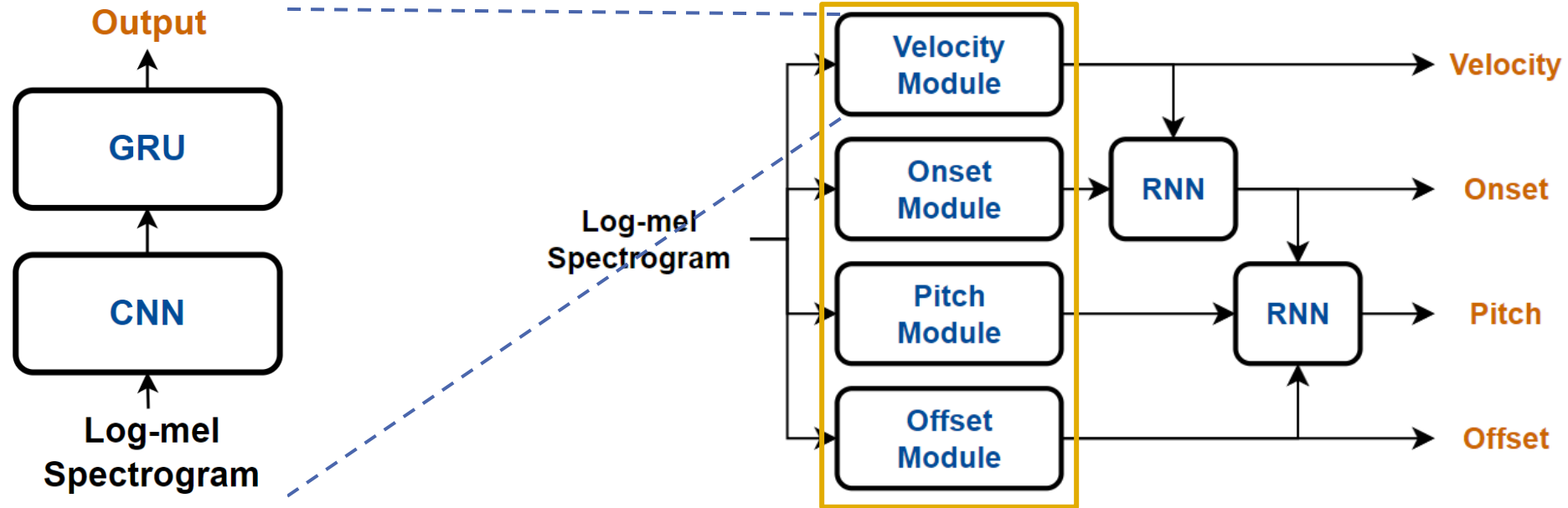
Separate network for each subtask

Network for each subtask: CNN-GRU, Same structure for every subtask

Audio input: Log-mel spectrogram

Output of each branch: Does this element exist, on a specific timestep, of a specific pitch

Baseline System: NNs



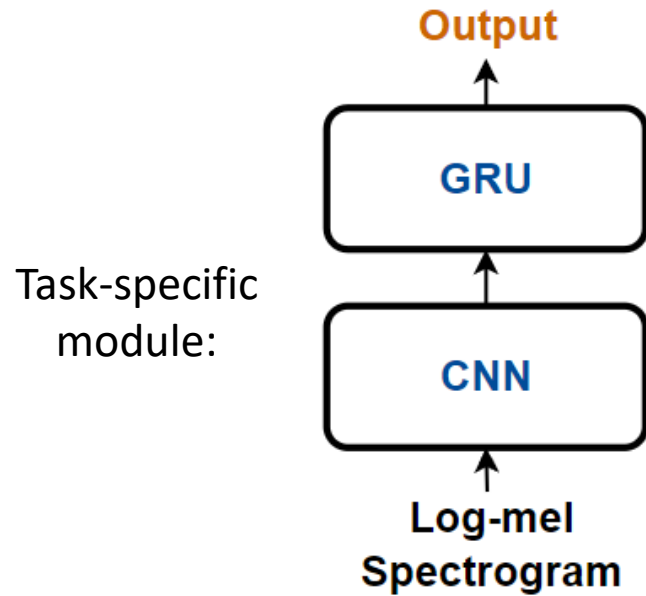
Separate network for each subtask

Network for each subtask: **CNN-GRU**, Same structure for every subtask

Audio input: Log-mel spectrogram

Output of each branch: Does this element exist, on a specific timestep, of a specific pitch

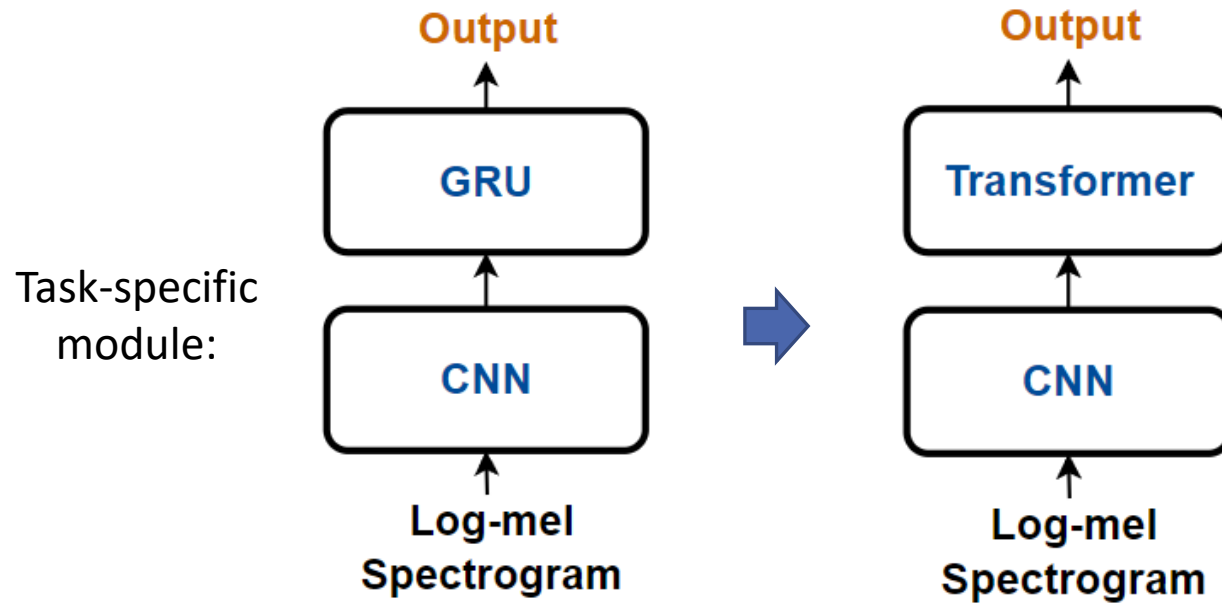
Model: CNN-Transformer



CNN-Transformer

Transformer: 4 layers of Transformer blocks

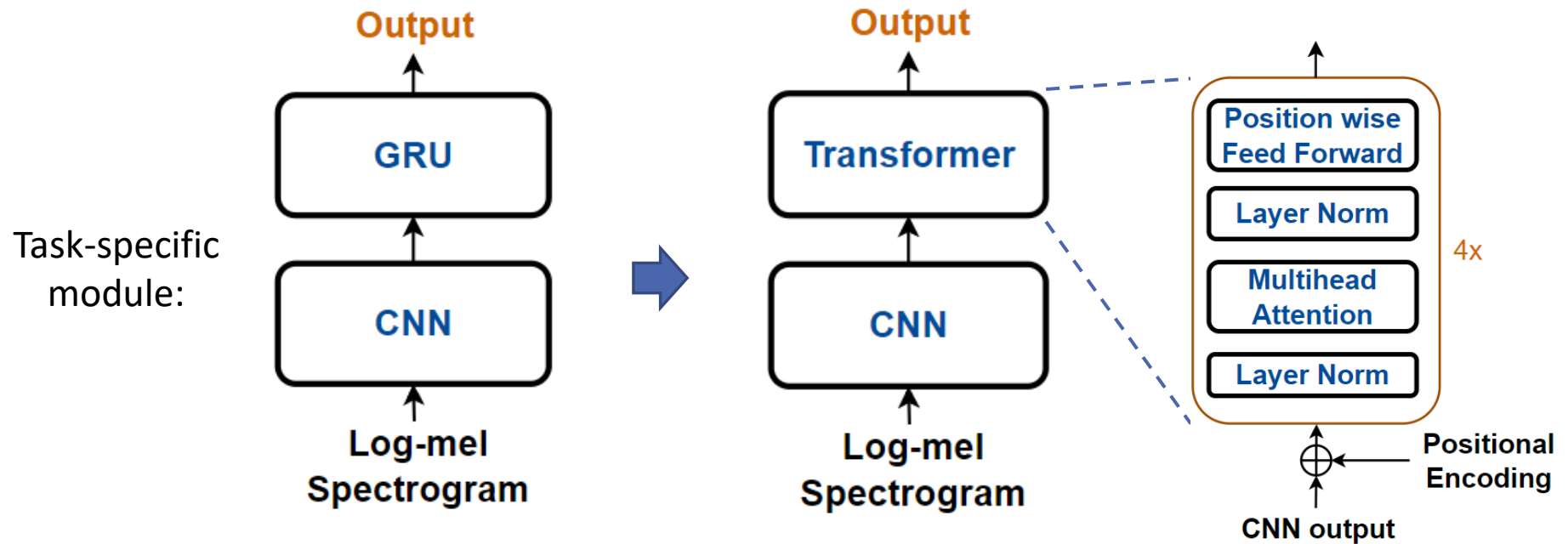
Model: CNN-Transformer



CNN-Transformer

Transformer: 4 layers of Transformer blocks

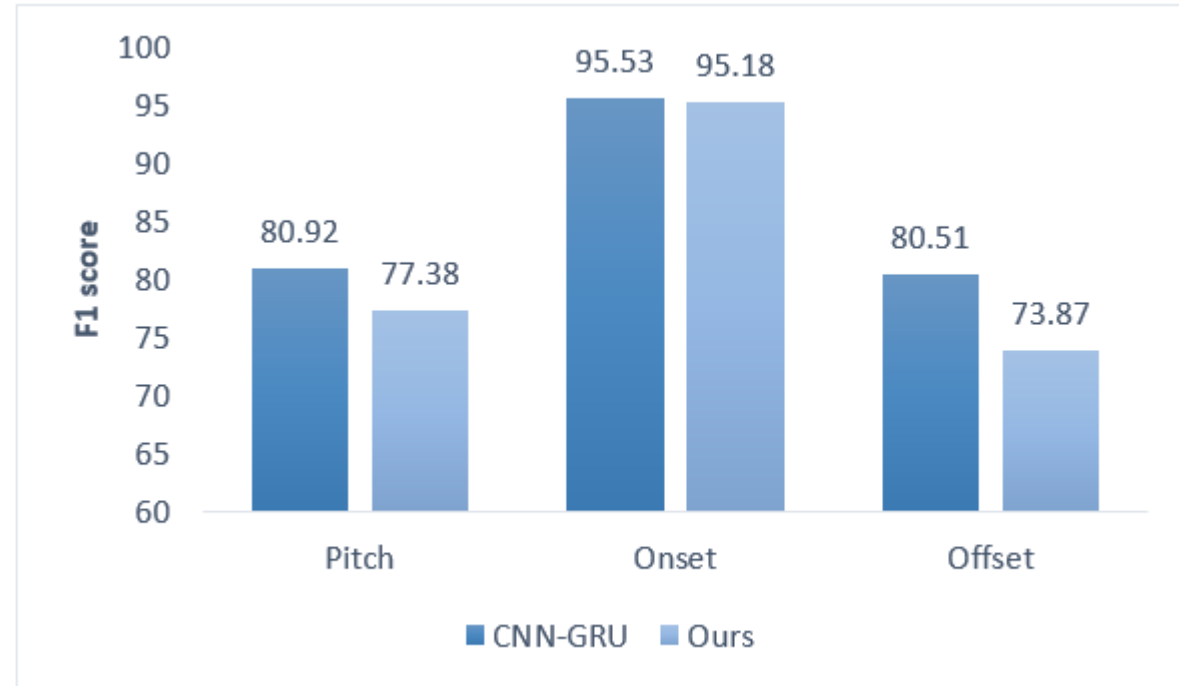
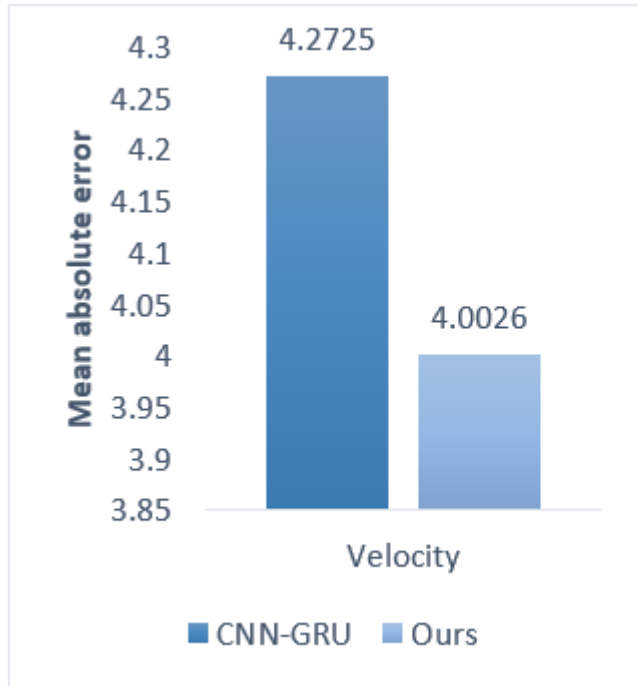
Model: CNN-Transformer



CNN-Transformer

Transformer: 4 layers of Transformer blocks

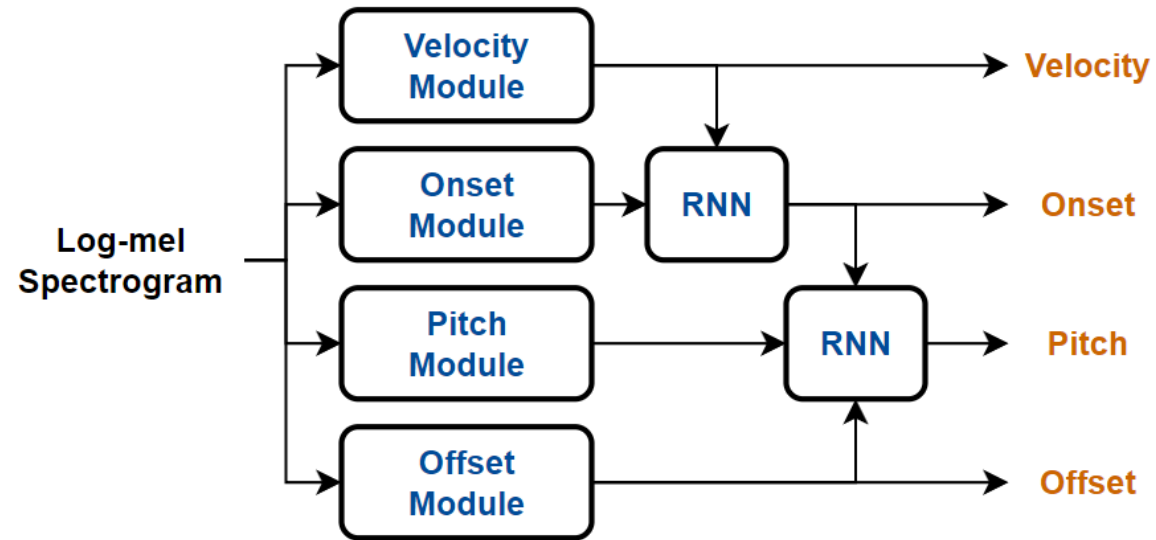
Evaluation 1/2: performance on subtasks



Result:

- Our model is better at velocity task, with relative 6.3% lower MAE;
- But not for the other three subtasks

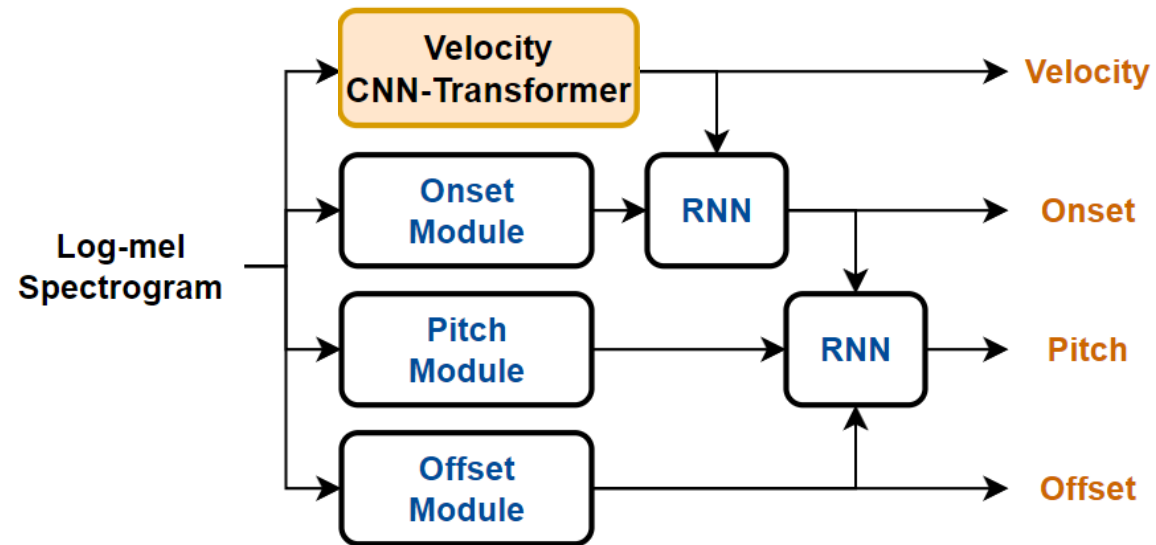
Model: Baseline + velocity Transformer



Afterward

Switch CNN-GRU to CNN-Transformer on velocity branch, to observe the impact on overall transcription performance.

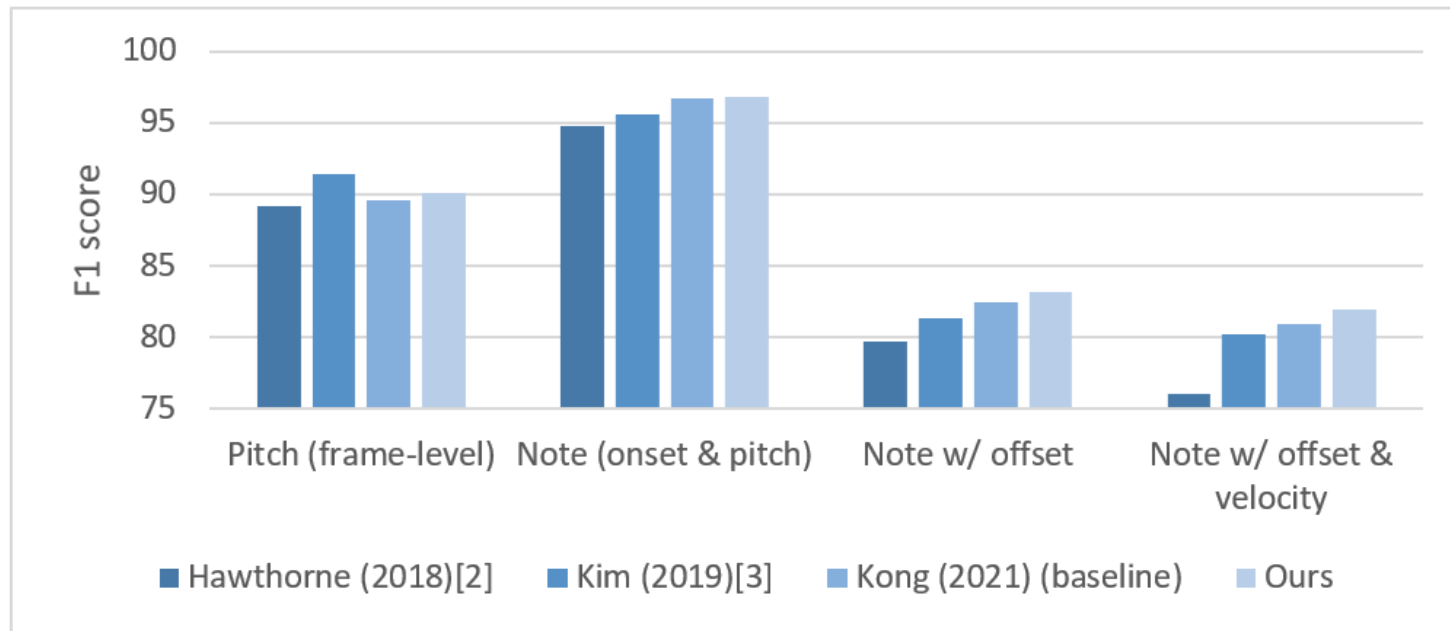
Model: Baseline + velocity Transformer



Afterward

Switch CNN-GRU to CNN-Transformer on velocity branch, to observe the impact on overall transcription performance.

Evaluation 2/2: transcription performance



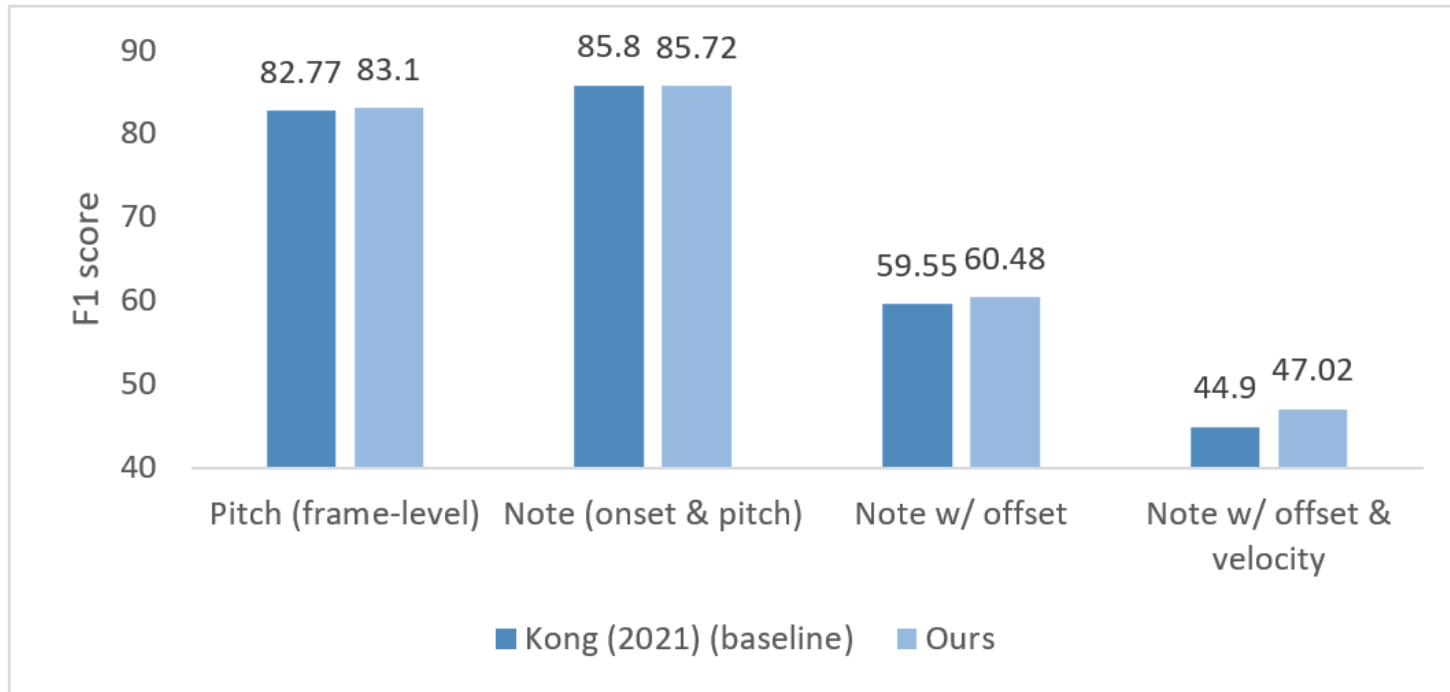
Performance:

- Better than baseline model on both frame-level and note-level metrics
- 2nd best frame-level multipitch estimation result among SOTA models
- Better when testing with out-of-domain data

[2] Hawthorne, Curtis, et al. "Onsets and Frames: Dual-Objective Piano Transcription." *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, edited by Emilia Gómez et al., 2018, pp. 50–57, http://ismir2018.ircam.fr/doc/pdfs/19_Paper.pdf.

[3] Kim, Jong Wook, and Juan Pablo Bello. "Adversarial Learning for Improved Onsets and Frames Music Transcription." *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, the Netherlands, November 4-8, 2019*, edited by Arthur Flexer et al., 2019, pp. 670–77, <http://archives.ismir.net/ismir2019/paper/000081.pdf>.

Evaluation 2/2: transcription performance



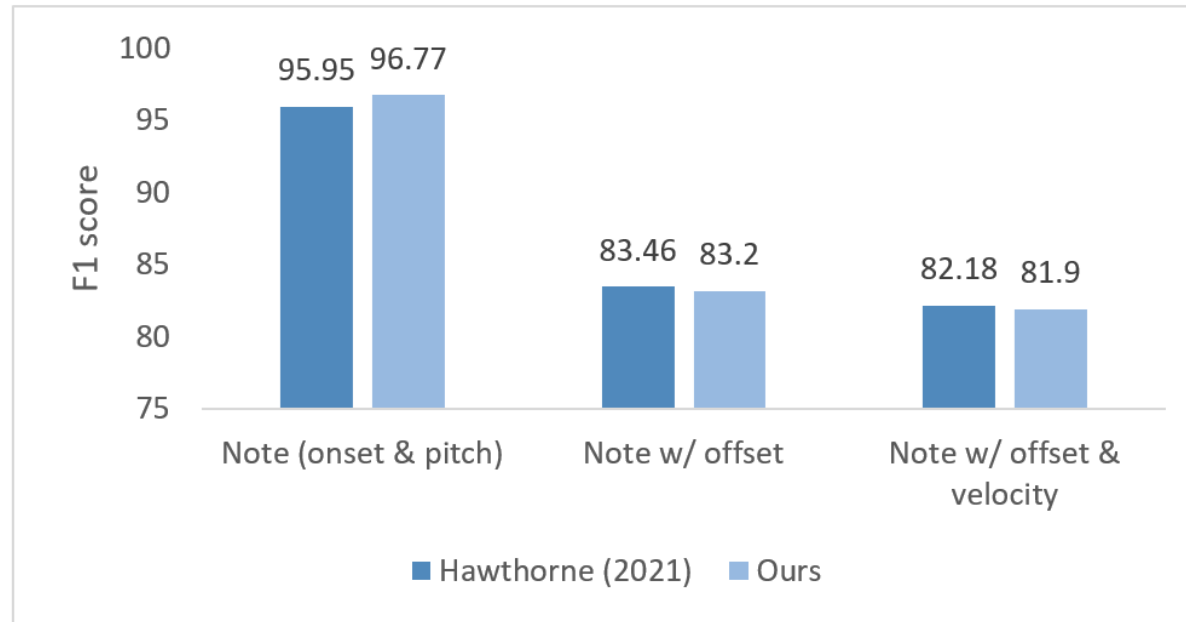
Performance:

- Better than baseline model on both frame-level and note-level metrics
- 2nd best frame-level multipitch estimation result among SOTA models
- Better when testing with out-of-domain data

[2] Hawthorne, Curtis, et al. "Onsets and Frames: Dual-Objective Piano Transcription." *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, edited by Emilia Gómez et al., 2018, pp. 50–57, http://ismir2018.ircam.fr/doc/pdfs/19_Paper.pdf.

[3] Kim, Jong Wook, and Juan Pablo Bello. "Adversarial Learning for Improved Onsets and Frames Music Transcription." *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, the Netherlands, November 4-8, 2019*, edited by Arthur Flexer et al., 2019, pp. 670–77, <http://archives.ismir.net/ismir2019/paper/000081.pdf>.

Compare with generic Transformer^[4]



Performance:

- Slightly lower result on two note-level metrics
- Ours has **higher onset & pitch score**
- **Provide frame-level output**, may facilitate other tasks


[4] Hawthorne, Curtis, et al. "Sequence-to-Sequence Piano Transcription with Transformers." *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, edited by Jin Ha Lee et al., 2021, pp. 246–53, <https://archives.ismir.net/ismir2021/paper/000030.pdf>.

Main Takeaways

- CNN-Transformer does better on velocity task—capturing long-range dependencies benefits the velocity estimation task
- A piano transcription system with competitive note-level results, while also provide decent frame-level result to detect note elements

Possible future direction:

- How well this model perform on transcription of other instruments?
- Multitask setup?
- Can we build semi-/self-supervised learning transcription models?

A close-up, shallow depth-of-field photograph of piano keys, showing the white and black keys receding into the distance. The image is partially obscured by a white curved shape on the right side.

Exploring Transformer's Potential on Automatic Piano Transcription

Longshen Ou 2022
oulongshen@u.nus.edu

Sound and Music Computing Lab
National University of Singapore