UC San Diego

2021 UC San Diego J Yang Scholars Symposium

iem
Institute of Engineering in Medicine

# Towards Automatic Instrumentation by Learning to Separate Parts in Symbolic Multitrack Music

**Hao-Wen Dong**
董皓文
Department of Computer Science and Engineering

Advisors: Prof. Julian McAuley and Prof. Taylor Berg-Kirkpatrick

September 25, 2021

# About me

Hi, I'm Herman.
I do Music x AI research.
I love music and movies!

**National Taiwan University**
*B.S. in Electrical Engineering*
2013 – 2017

2017 – 2019
**Academia Sinica**
*Research Assistant*
Music and AI Lab, Research Center for IT Innovation

Summer 2019
**Yamaha Corporation**
*Research Intern*
AI Group, Research and Development Division

**University of California San Diego**
*M.S. in Computer Science*
2019 – 2021

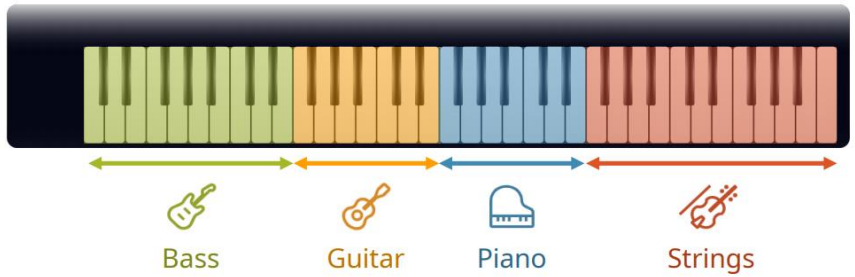
Summer 2021
**Dolby Laboratories**
*Deep Learning Audio Intern*
Applied AI Team, Advanced Technology Group

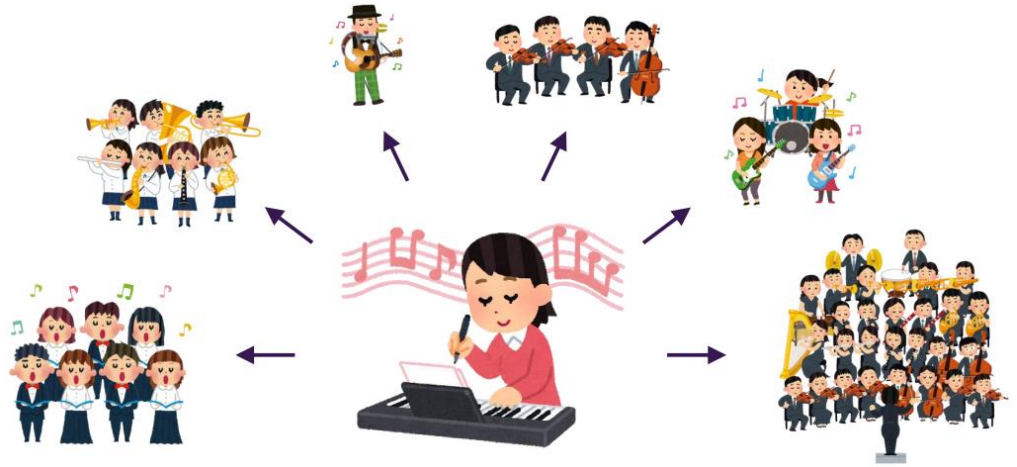
**University of California San Diego**
*Ph.D. in Computer Science (in progress)*
2019 – present

UC San Diego

# Automatic instrumentation

- **Goal**—Dynamically assign instruments to notes in solo music

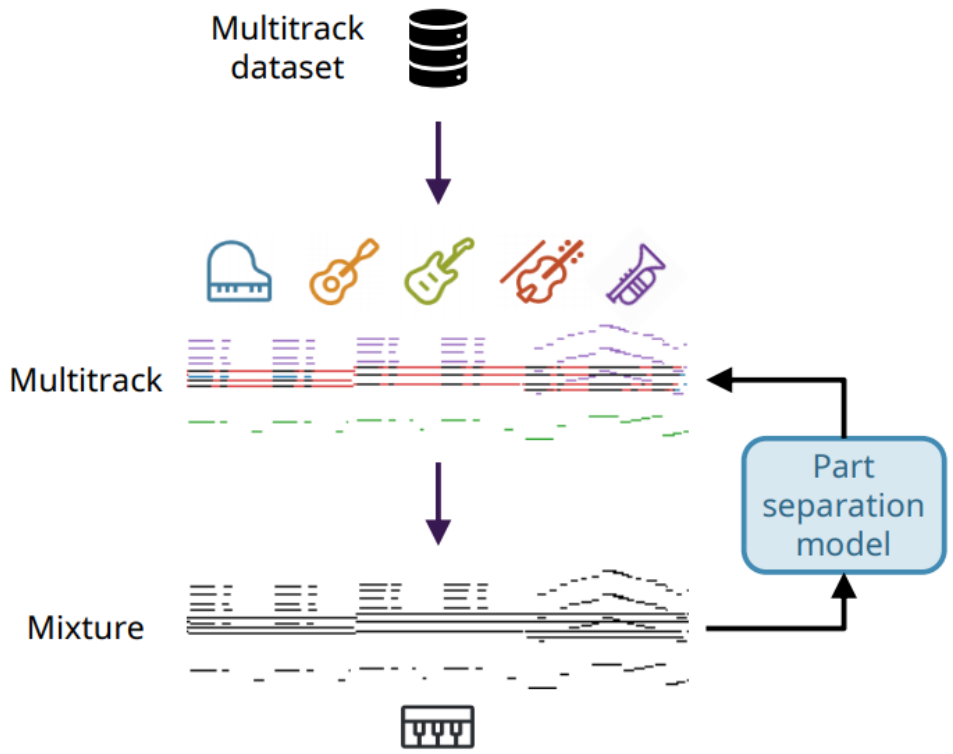


Hao-Wen Dong, Chris Donahue, Taylor Berg-Kirkpatrick and Julian McAuley, "Towards Automatic Instrumentation by Learning to Separate Parts in Symbolic Multitrack Music," Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR), in press, 2021.

# Overview

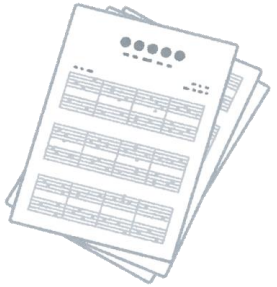
- Acquire paired data of solo music and its instrumentation
  - Downmix multitracks into single-track mixtures
- Train a part separation model
  - Learn to infer the part label for each note in a mixture
- Approach automatic instrumentation
  - Treat input from a keyboard player as a downmixed mixture
  - Separate out the relevant parts



# Data

- Four datasets of diverse genres and ensembles

| Dataset | Hours | Files | Notes | Parts | Ensemble | Most common label |
|---|---|---|---|---|---|---|
| Bach chorales [31] | 3.23 | 409 | 96.6K | 4 | soprano, alto, tenor, bass | bass (27.05%) |
| String quartets [32] | 6.31 | 57 | 226K | 4 | first violin, second violin, viola, cello | first violin (38.72%) |
| Game music [33] | 45.05 | 4.61K | 2.46M | 3 | pulse wave I, pulse wave II, triangle wave | pulse wave II (39.35%) |
| Pop music [34] | 1.02K | 16.2K | 63.6M | 5 | piano, guitar, bass, strings, brass | guitar (42.50%) |



# Models & input features

## Models

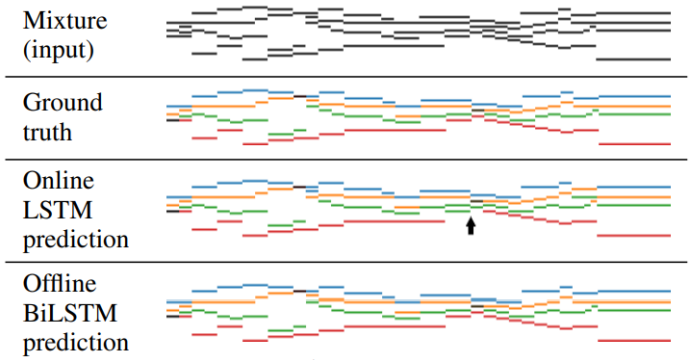
- Deep sequential models
  - Online LSTM
  - Offline BiLSTM
- Baseline models
  - Zone-based algorithm
  - Closest-pitch algorithm
  - Multilayer perceptron (MLP)

## Input features

- time—onset time (in time step)
- pitch—pitch as a MIDI note number
- duration—note length (in time step)
- frequency—frequency of the pitch (in Hz)
- beat—onset time (in beat)
- position—position within a beat (in time step)

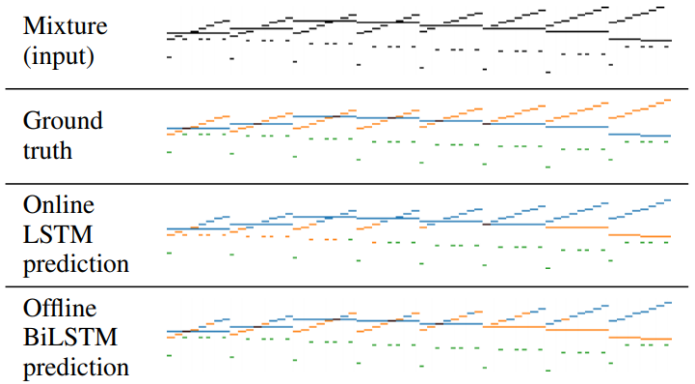
# Qualitative results

## Bach chorales




(Audio available. [1] Colors: piano, soprano, tenor, bass.)

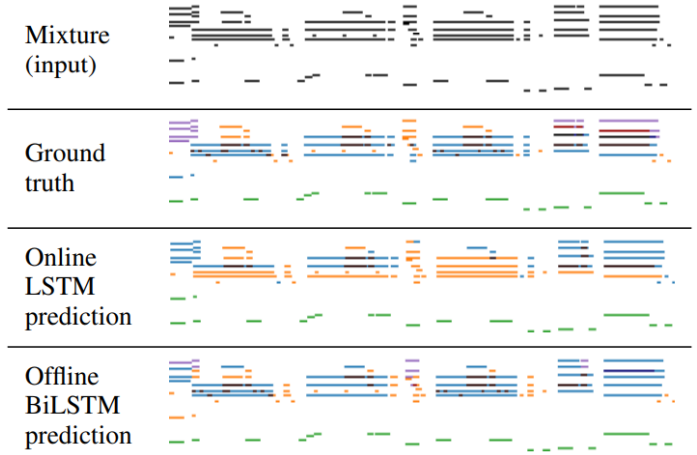
## Game music




(Audio available. [1] Colors: pulse wave I, pulse wave II, triangle wave.)

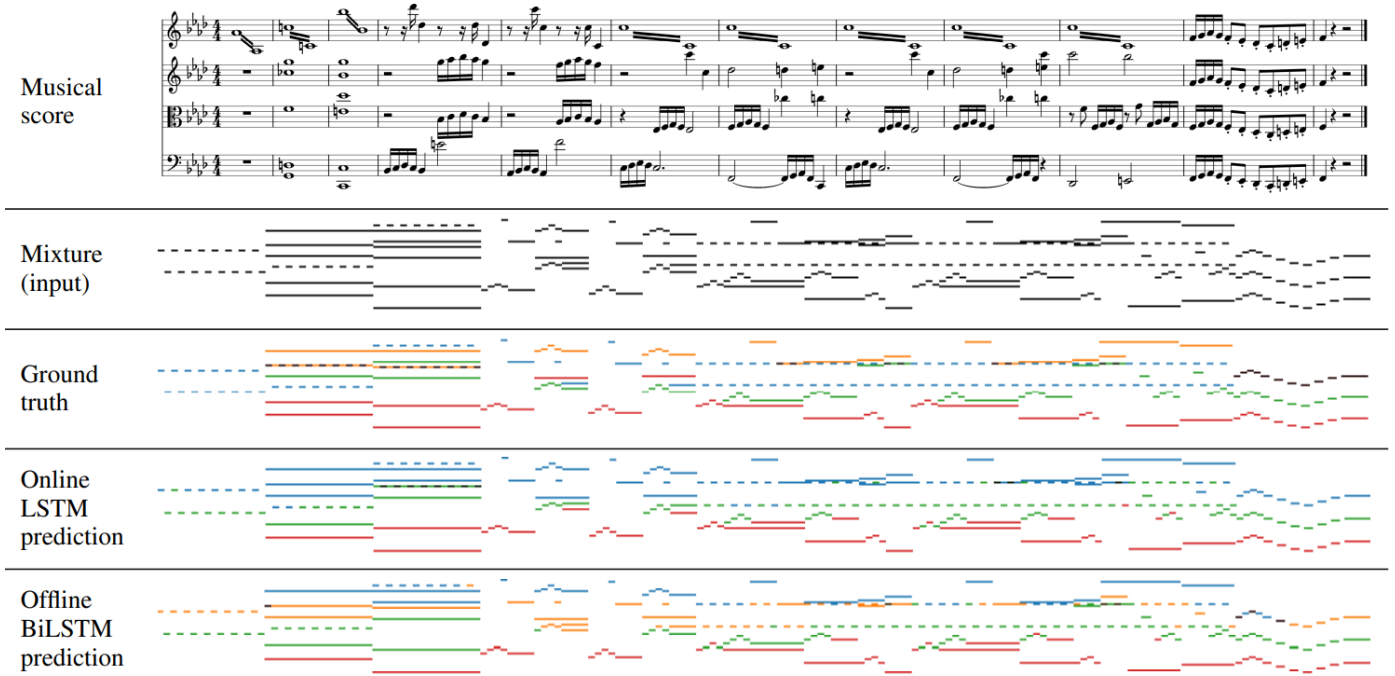
**These examples are all hard cases!**

## Pop music




(Audio available. [1] Colors: piano, guitar, bass, strings, brass.)

## String quartets




(Audio available. [1] Colors: first violin, second violin, viola, cello.)

Audio samples are available at salu133445.github.io/arranger/ .

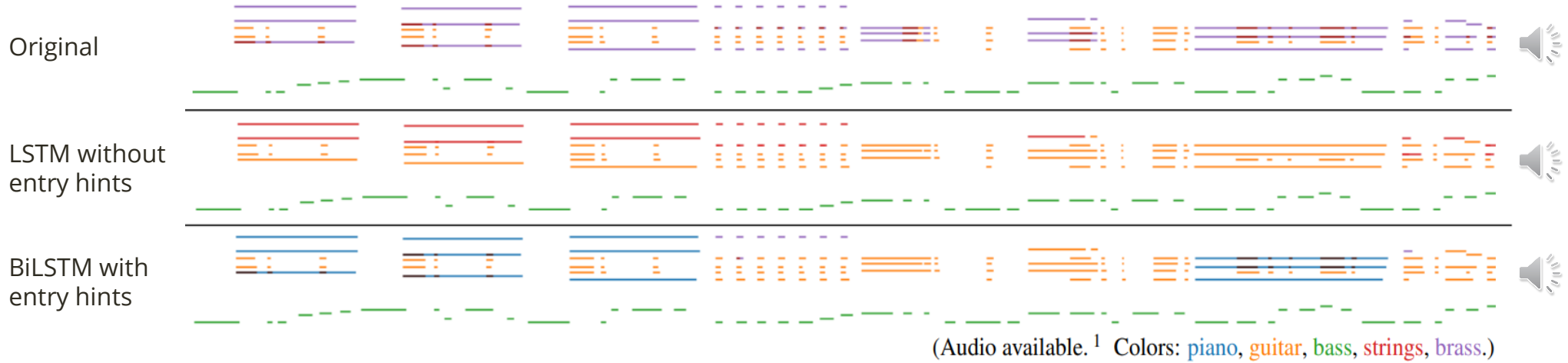
# Quantitative results

- Proposed models outperform baseline models
- BiLSTM outperforms LSTM
- LSTM models outperform their Transformer counterparts

| Model | Bach | String | Game | Pop |
|---|---|---|---|---|
| **Online models** | | | | |
| Zone-based | 73.14 | 58.85 | 43.67 | 57.07 |
| MLP [9] | 81.63 | 29.85 | 43.08* | 33.50* |
| LSTM | **93.02** | **67.43** | **50.22** | **74.14** |
| Transformer-Dec | 91.51 | 57.03 | 45.82 | 62.14 |
| Zone-based (oracle) | 78.33 | 66.89 | 79.54* | † |
| MLP [9] (oracle) | 97.59 | 58.16 | 65.30 | 44.62 |
| **Offline models** | | | | |
| BiLSTM | **97.13** | **74.38** | **52.93** | **77.23** |
| Transformer-Enc | 96.81 | 58.86 | 49.14 | 66.57 |
| **Online models (+entry hints)** | | | | |
| Closest-pitch | 68.87 | 50.69 | 57.14 | 47.45 |
| Closest-pitch (mono) | 89.76 | 42.82 | 49.91 | 32.28 |
| LSTM | **92.70** | **62.64** | **62.11** | **74.19** |
| Transformer-Dec | 91.17 | 62.12 | 56.73 | 67.19 |
| **Offline models (+entry hints)** | | | | |
| BiLSTM | **97.39** | **71.51** | **64.79** | **75.59** |
| Transformer-Enc | 93.81 | 56.72 | 54.67 | 67.23 |

# Demo

- The proposed models can produce alternative convincing instrumentations for an existing arrangement




(Audio available. [1] Colors: piano, guitar, bass, strings, brass.)

More examples are available at salu133445.github.io/arranger/ .

# Summary

- Proposed a new task of part separation
- Showed that our proposed models outperform various baselines
- Presented promising results for applying a part separation model to automatic instrumentation

# Future directions

- Generative modeling of automatic instrumentation
- Unpaired automatic instrumentation
- Large-scale pretraining for symbolic music models

# Acknowledgement


- This is a joint work with Chris Donahue, Taylor Berg-Kirkpatrick and Julian McAuley.
- I would like to thank the J. Yang and Family Foundation for supporting my PhD study with the J. Yang Scholarship.

Thank you!