INCREASING DRUM TRANSCRIPTION VOCABULARY USING DATA SYNTHESIS

Mark Cartwright and Juan Pablo Bello
Music and Audio Research Laboratory
New York University

http://www.markcartwright.com

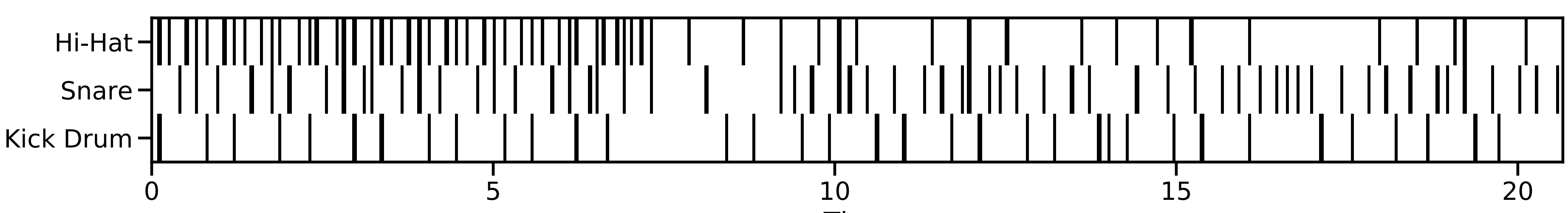
Problem

Most Automatic Drum Transcription (ADT) algorithms are limited to simply onset times of 3 classes:

1. Bass Drum (BD)
2. Snare Drum (SD)
3. Hi-Hat (HH)

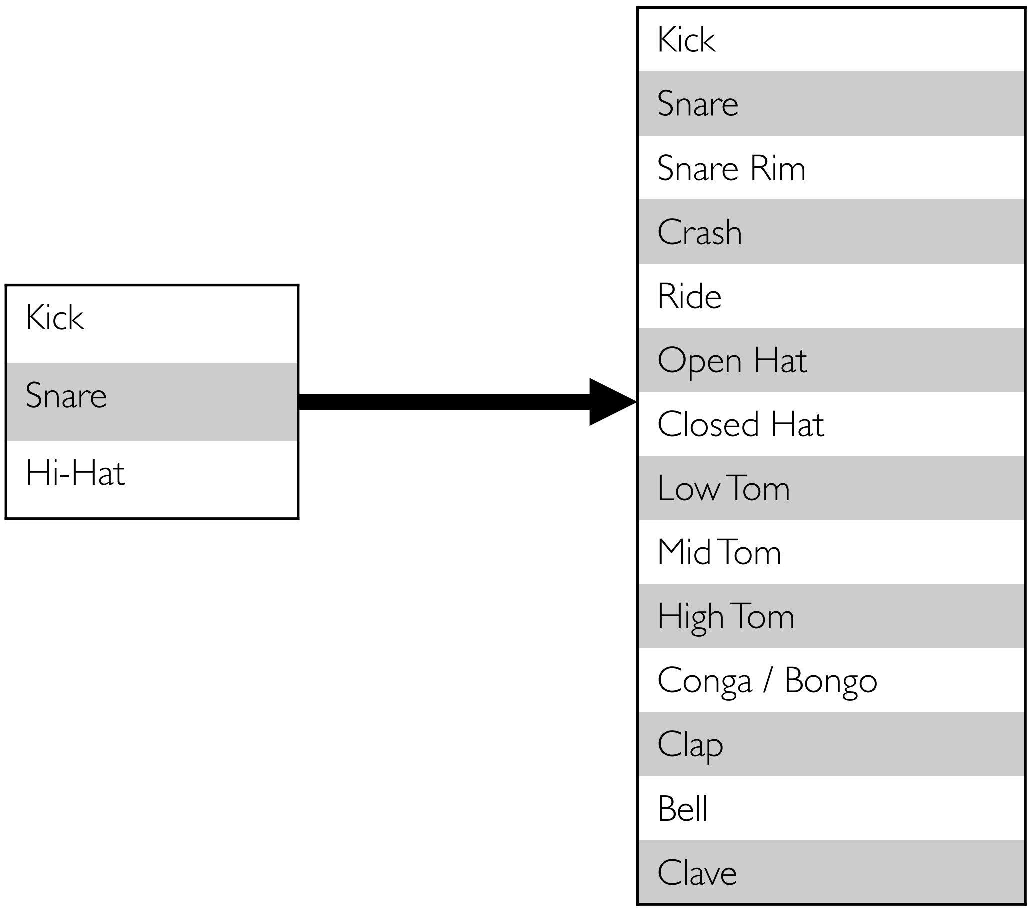
Example:

Original   3-class re-synthesis:

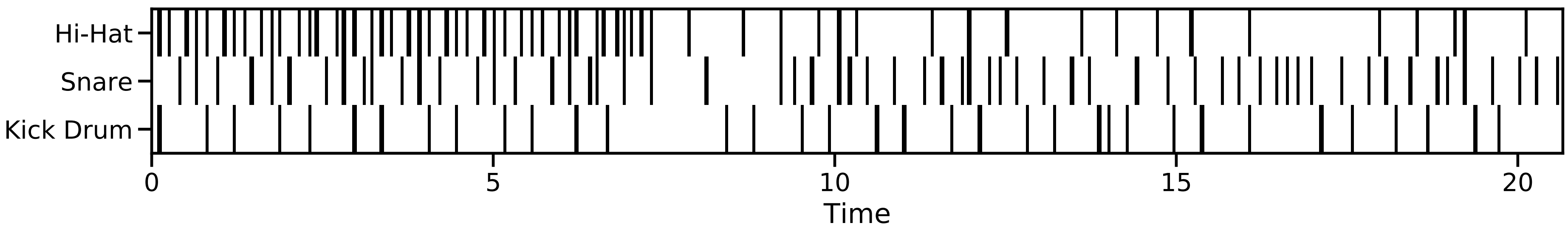


Goal

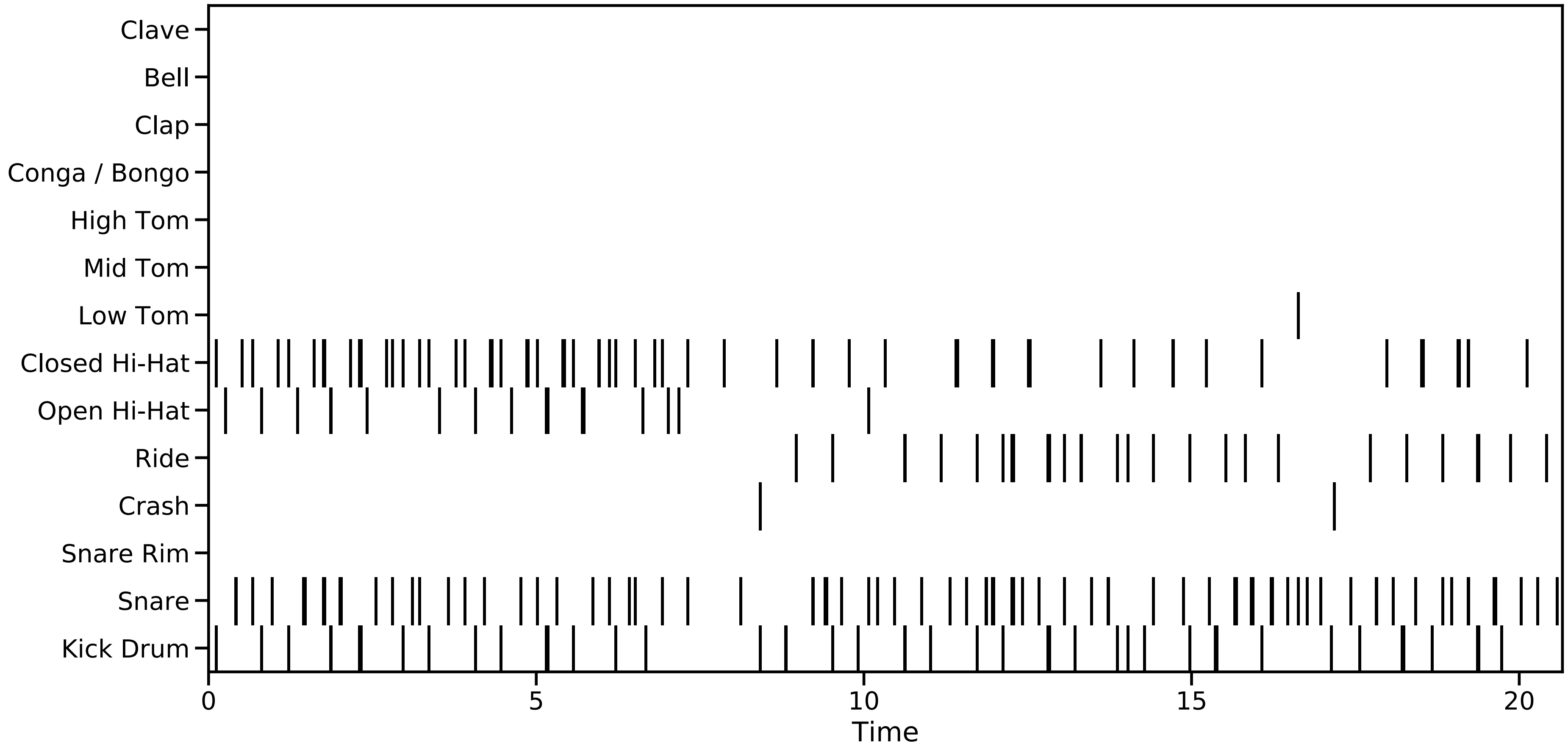
- Increase vocabulary to onset times of 14 drum classes:



Goal



Goal

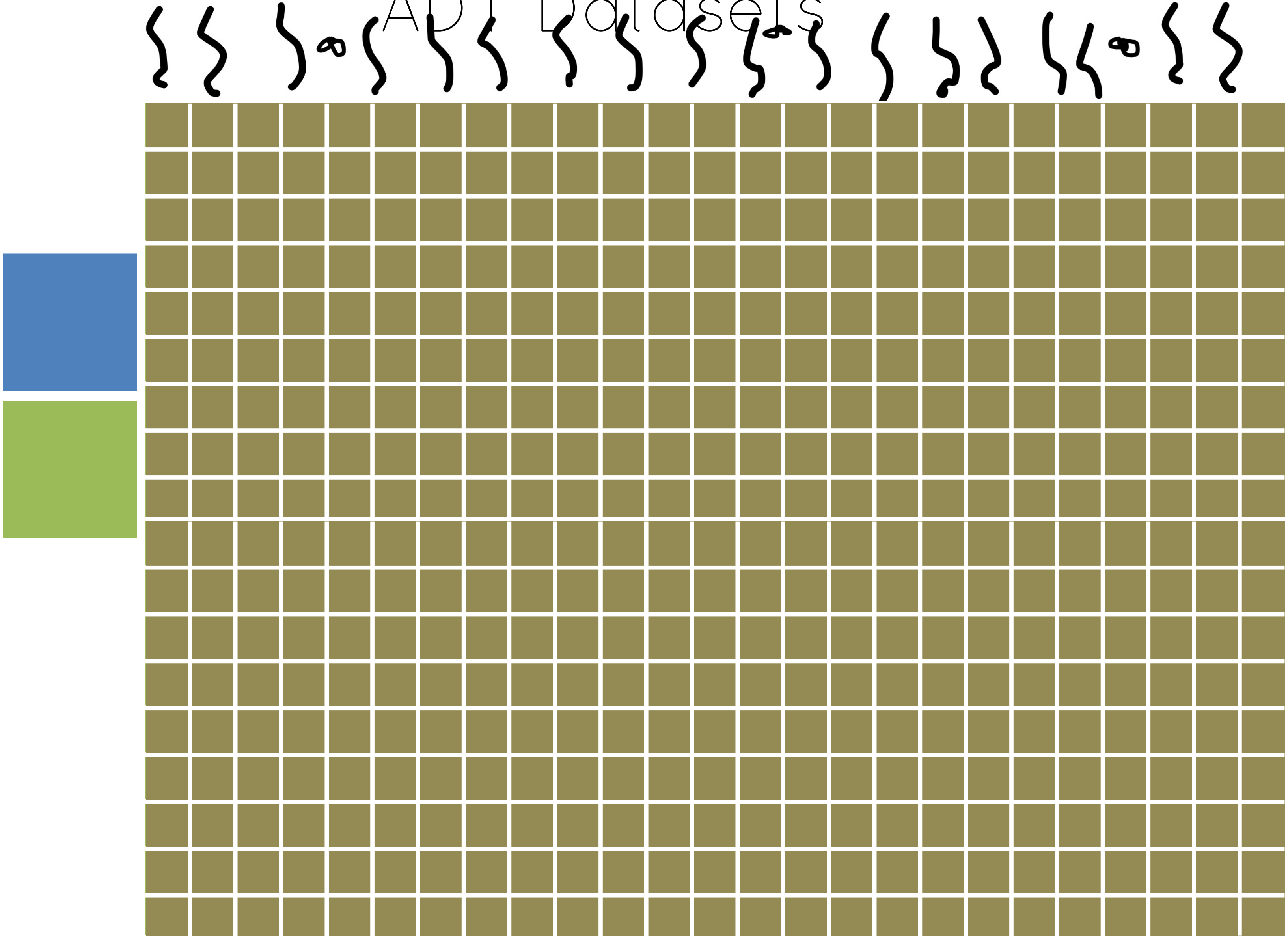

Clave
Bell
Clap
Conga / Bongo
High Tom
Mid Tom
Low Tom
Closed Hi-Hat
Open Hi-Hat
Ride
Crash
Snare Rim
Snare
Kick Drum
0
5
10
15
20
Time

ADT Datasets

~33,000 onsets / 2 hours of 3-class data (RBMA, IDMT/SMT)

~33,000 onsets / ~1.5 hours of > 3-class data (ENST/MDB)

ADT Datasets



Building the S********* Drum Dataset (SDDS)
The Sounds

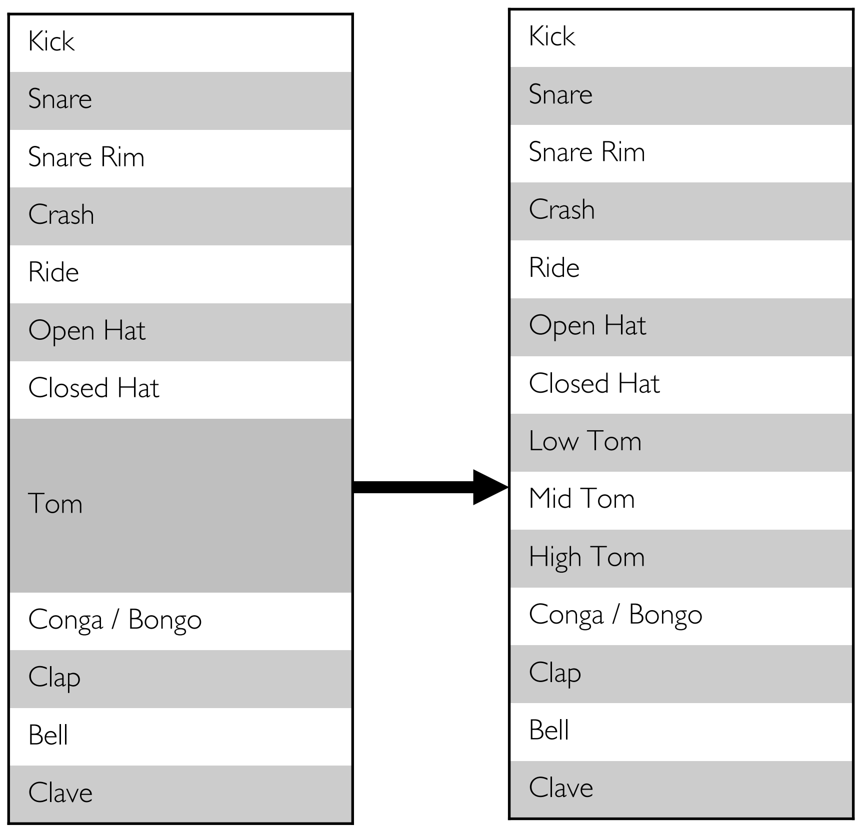
- Labeled samples from 8 drum sample libraries
- Split toms into low, mid, high based on pitch, spectral centroid
- Split dataset into 4000 train / 2000 test samples

Kick
Snare
Snare Rim
Crash
Ride
Open Hat
Closed Hat
Tom
Conga / Bongo
Clap
Bell
Clave

Kick
Snare
Snare Rim
Crash
Ride
Open Hat
Closed Hat
Low Tom
Mid Tom
High Tom
Conga / Bongo
Clap
Bell
Clave

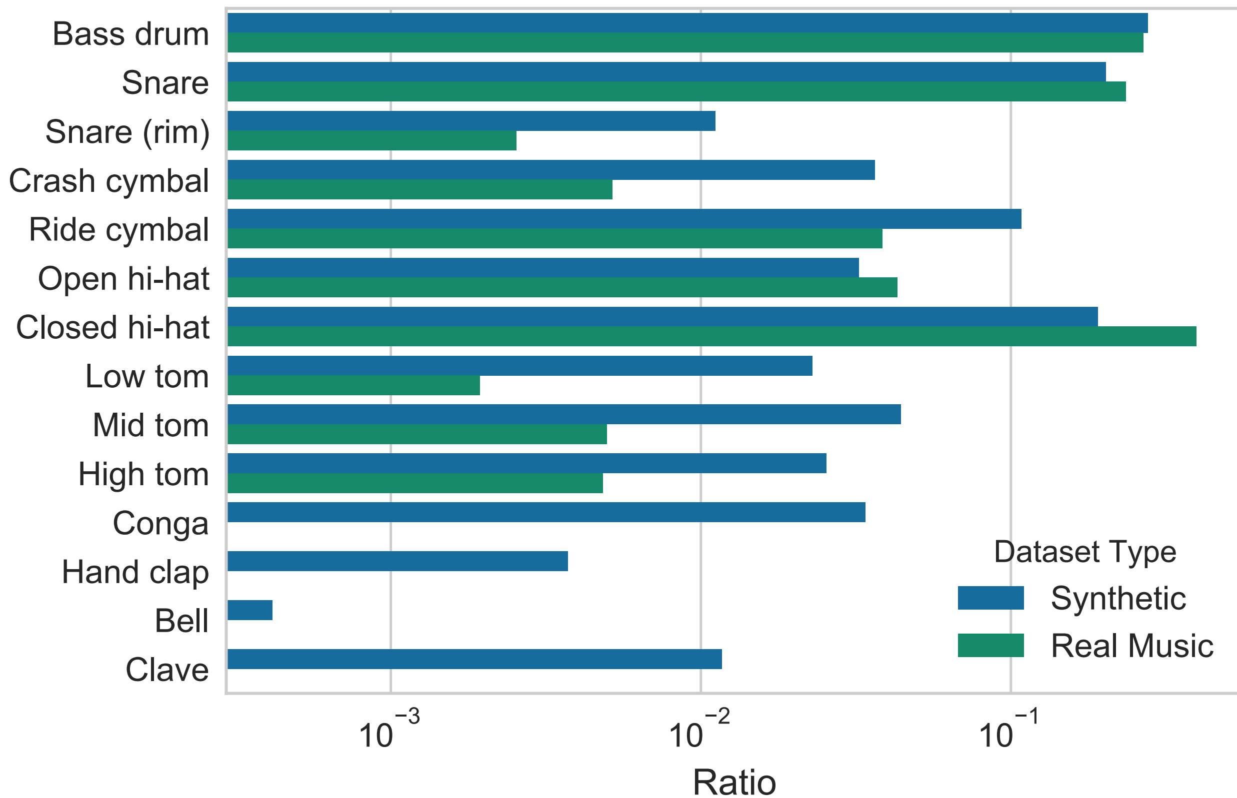
Building the Synthetic Drum Dataset (SDDS)
The Sounds

- Labeled samples from 8 drum sample libraries
- Split toms into low, mid, high based on pitch, spectral centroid
- Split dataset into 4000 train / 2000 test samples



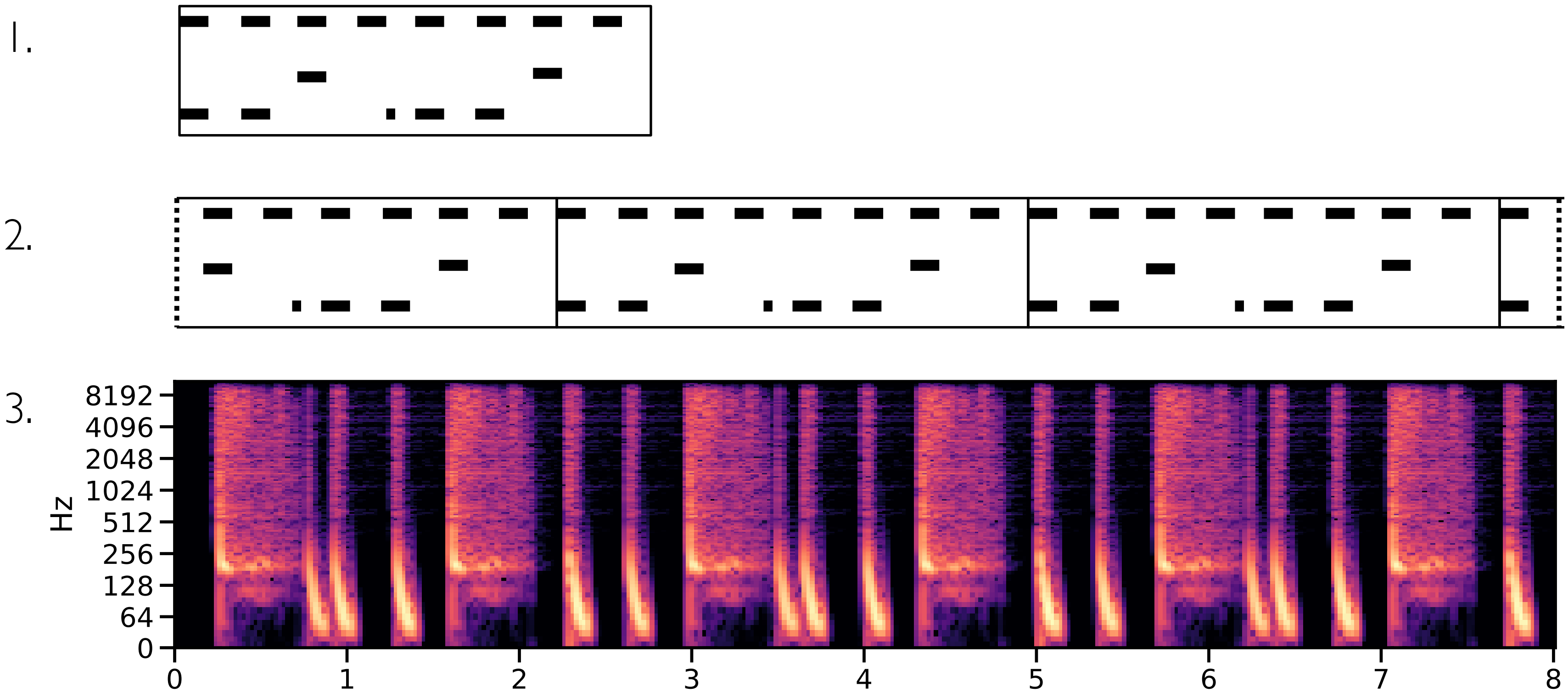
Building the Synthetic Drum Dataset (SDDS)
The Sounds

- Labeled samples from 8 drum sample libraries
- Split toms into low, mid, high based on pitch, spectral centroid
- Split dataset into 4000 train / 2000 test samples



Building the Synthetic Drum Dataset (SDDS)
The Rhythms

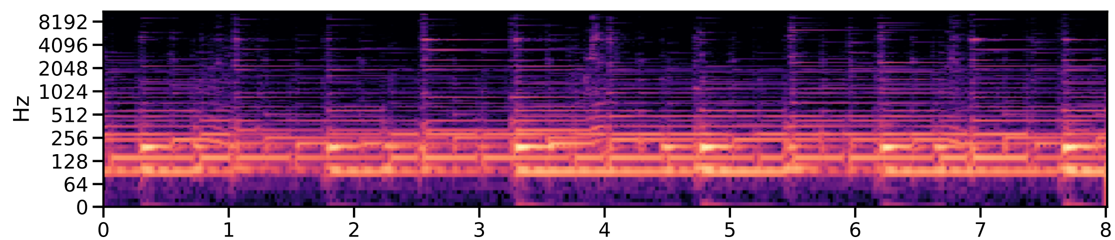
- 60k measures of percussion MIDI files (50k train / 5k test / 5k validate)



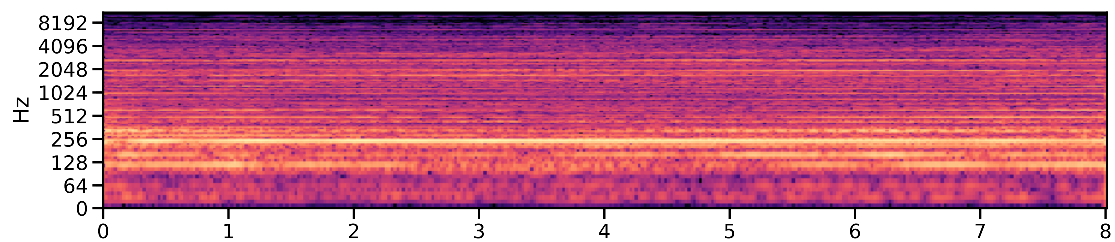
Building the Synthetic Drum Dataset (SDDS)
Augmentation and "Accompaniment"

- Augment to 210k (200k train / 5k test / 5k validate) w/ small pitch shifts, added pink noise, and "*harmonic noise* accompaniment":

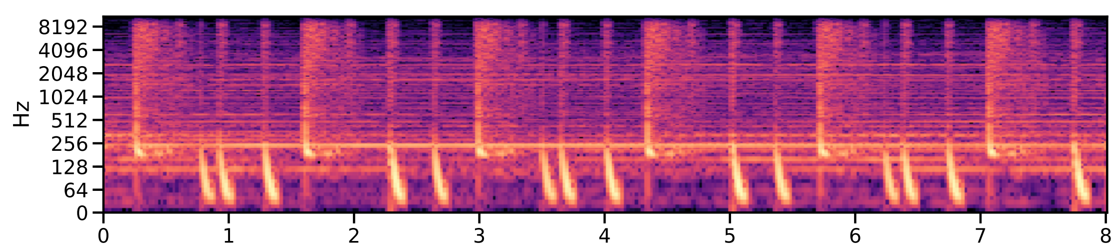
- Solo harmonic instrument recording



- "Smear" in time (fwd / bkwd reverb)



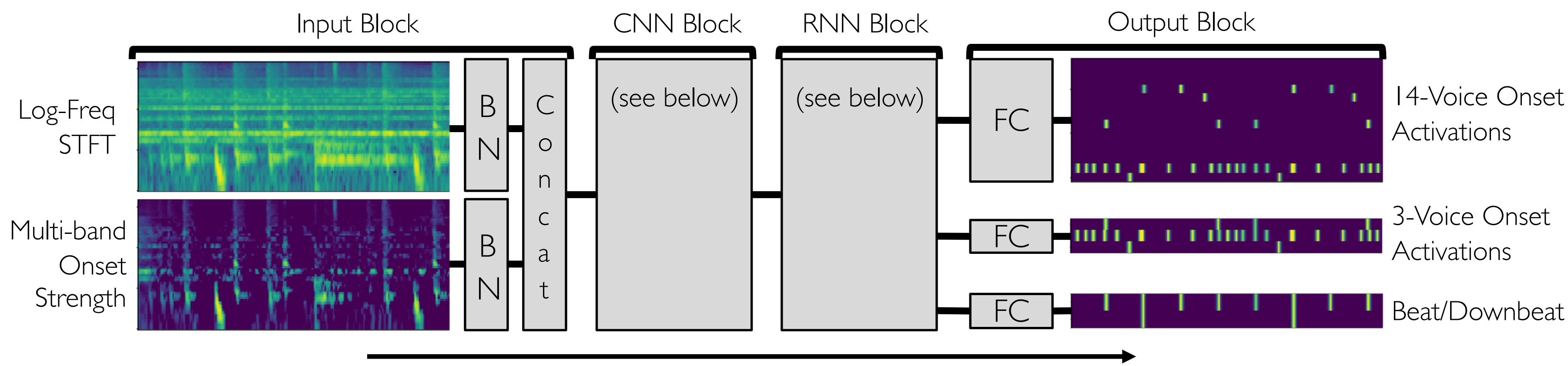
- Mix with drum track



Combined Datasets

	RBMA	SMT	ENST	MDB	SDDS
Hours	1.67	0.51	1.28	0.23	467
Accomp.	X	X		X	sort of
14-voice Onsets			X	X	X
3-voice Onsets	X	X	X	X	X
Beats	X				X

Model



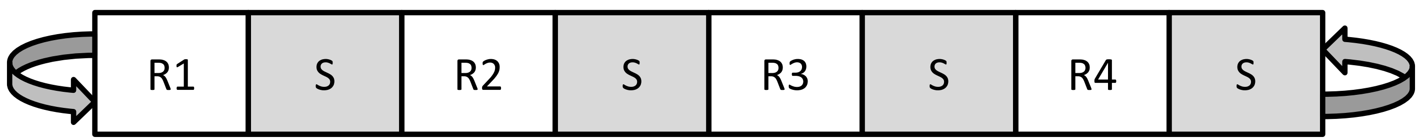
Input Features
22050 sampling rate
Log-magnitude, log-frequency STFT: 1024 samples 64 bands (8 octaves, 8 bands per octave) 40 Hz – 10kHz 0.01 sec hop
Multi-band onset strength: Computed from log-f STFT Half-rectified difference between current frame and mean of past 200 ms

CNN Block
32 (3x3) Conv
32 (3x3) Conv
Batch Norm
ReLU
30% Dropout
64 (3x3) Conv
64 (3x3) Conv
BatchNorm
ReLu
30% Dropout
64 (1x64) Conv
BatchNorm
ReLU

RNN Block
(-6:+6) Context Windowing
64-unit BLSTM
64-unit BLSTM
64-unit BLSTM

Training with Heterogeneous Outputs

- Mask outputs not in use for each example
- Use round-robin sampling with Pescador[1] to ensure all outputs used in each mini-batch of 8:



- Minimize weighted combination of binary cross-entropy losses based using weights computed by activation entropy
- 3-fold CV splits for the small real music small datasets
  25% validation / 75% testing in each split

1. https://github.com/pescadores/pescador

Experiments

Variables:

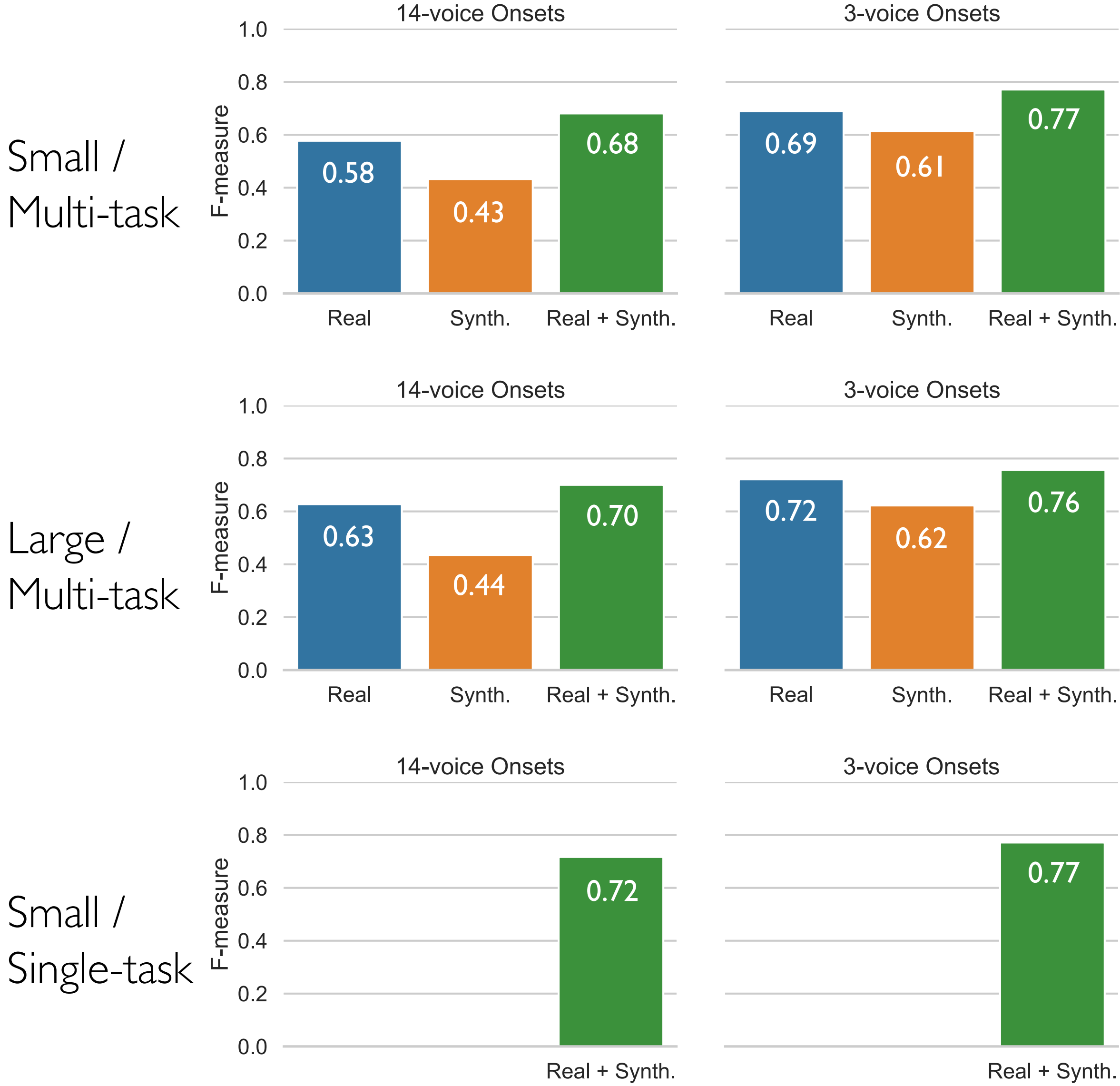
- Training data
  - Real music (RBMA, SMT, ENST, MDB)
  - Synthetic (SDDS)
  - Recorded + Synthetic

- Model capacity
  - "Small" (as described)
  - Large (more conv filters, more BLSTM units)

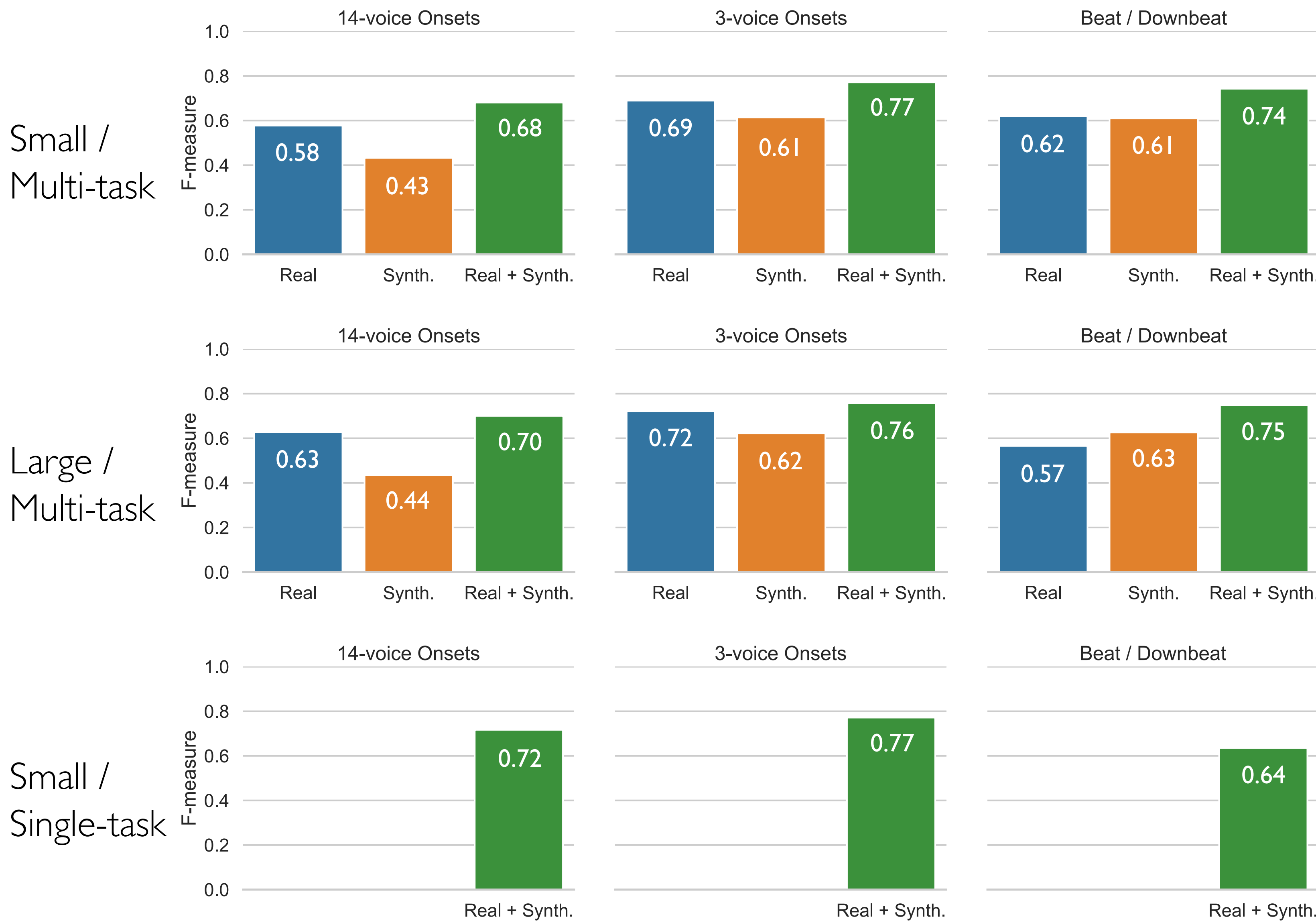
- Outputs
  - Multi-task
  - Single-task (w/ limited data)

- Class-weighting
  - No weighting
  - Weighted by activation entropy

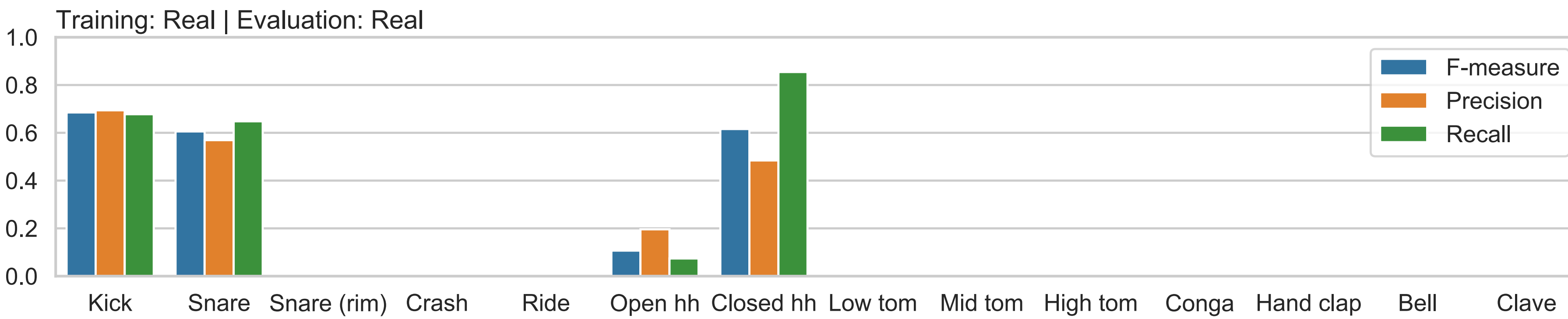
Results


Small /
Multi-task
14-voice Onsets
3-voice Onsets
F-measure
1.0
0.8
0.6
0.4
0.2
0.0
0.58
0.43
0.68
0.69
0.61
0.77
Real
Synth.
Real + Synth.
Large /
Multi-task
0.63
0.44
0.70
0.72
0.62
0.76
Small /
Single-task
0.72
0.77

Results

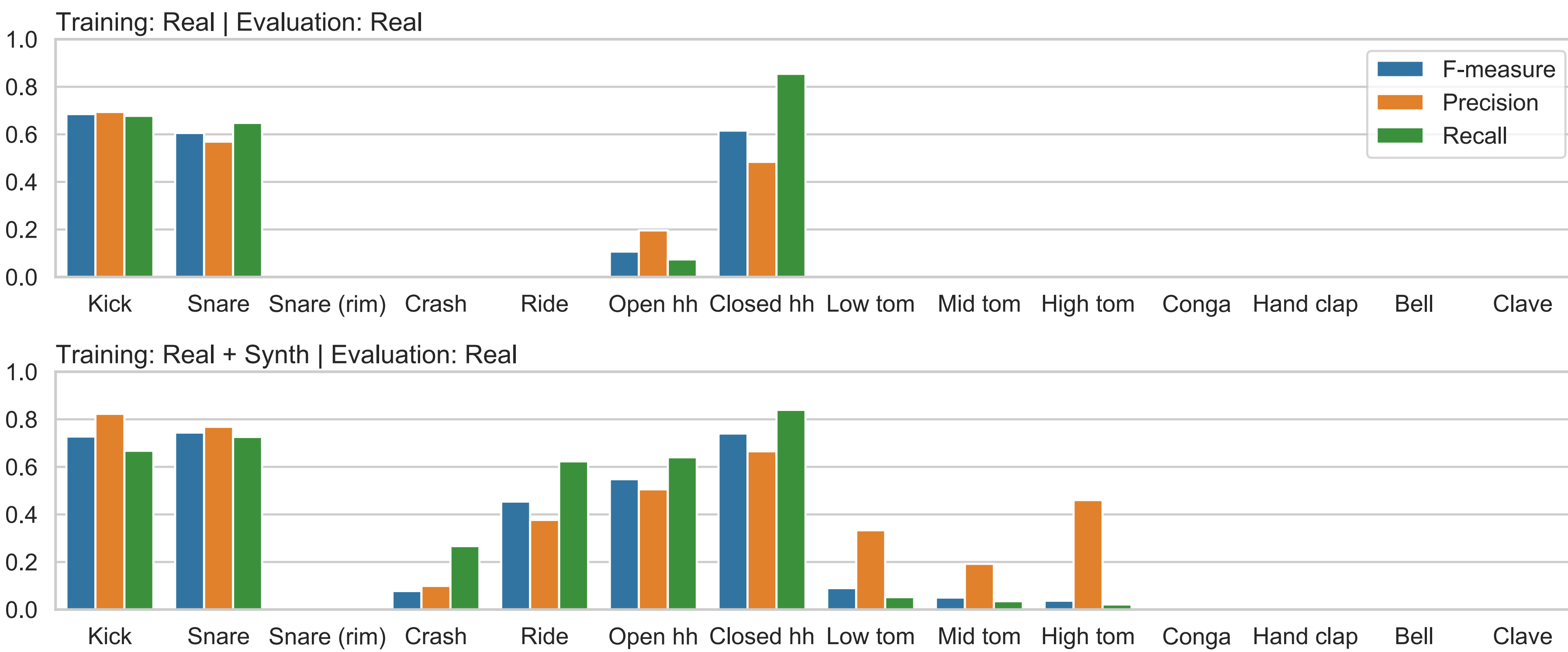

Small /
Multi-task
Large /
Multi-task
Small /
Single-task
F-measure
14-voice Onsets
3-voice Onsets
Beat / Downbeat
1.0
0.8
0.6
0.4
0.2
0.0
Real
Synth.
Real + Synth.
0.58
0.43
0.68
0.69
0.61
0.77
0.62
0.61
0.74
0.63
0.44
0.70
0.72
0.62
0.76
0.57
0.63
0.75
0.72
0.77
0.64

Results

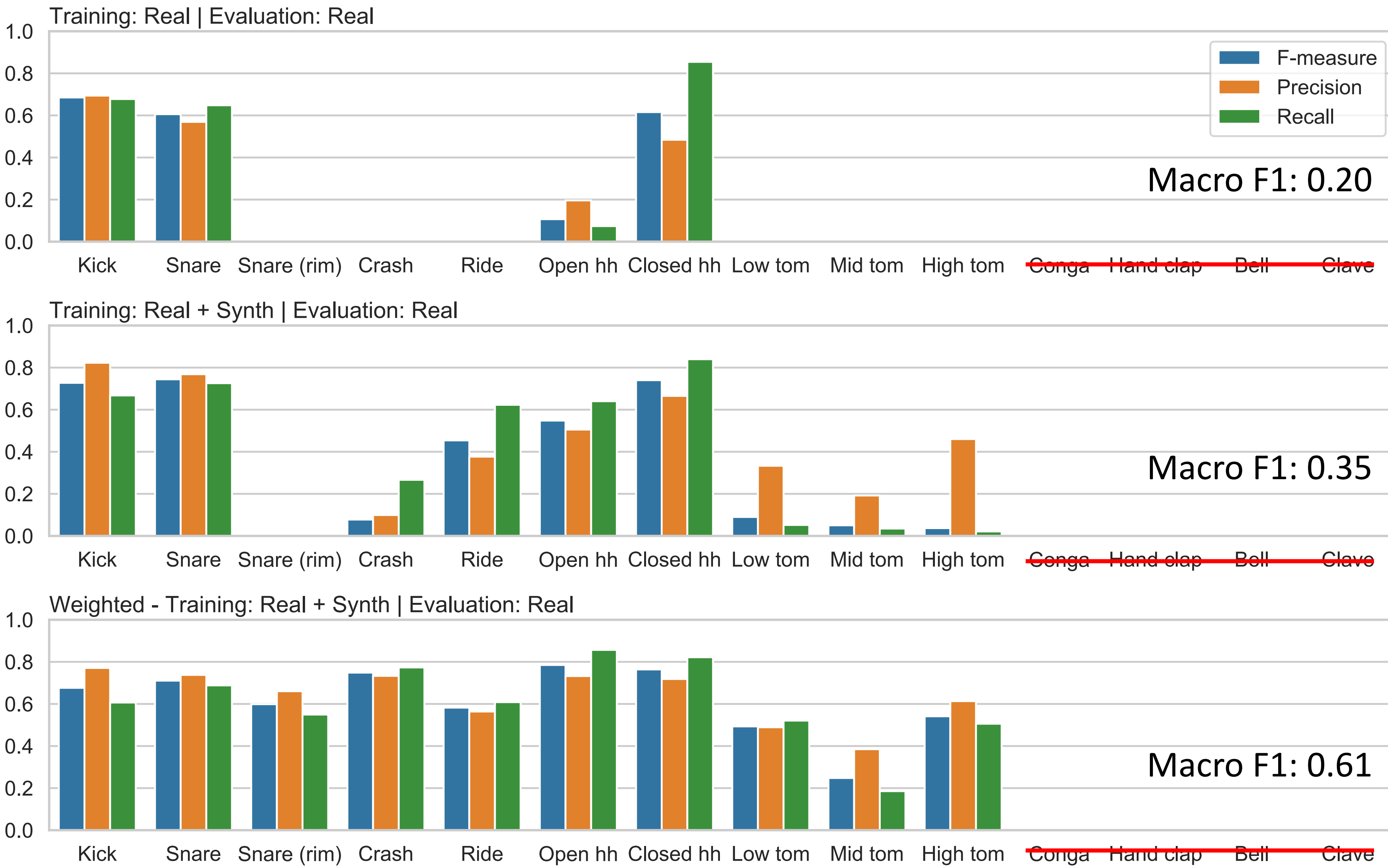


Training: Real | Evaluation: Real
1.0
0.8
0.6
0.4
0.2
0.0
F-measure
Precision
Recall
Kick
Snare
Snare (rim)
Crash
Ride
Open hh
Closed hh
Low tom
Mid tom
High tom
Conga
Hand clap
Bell
Clave

Results

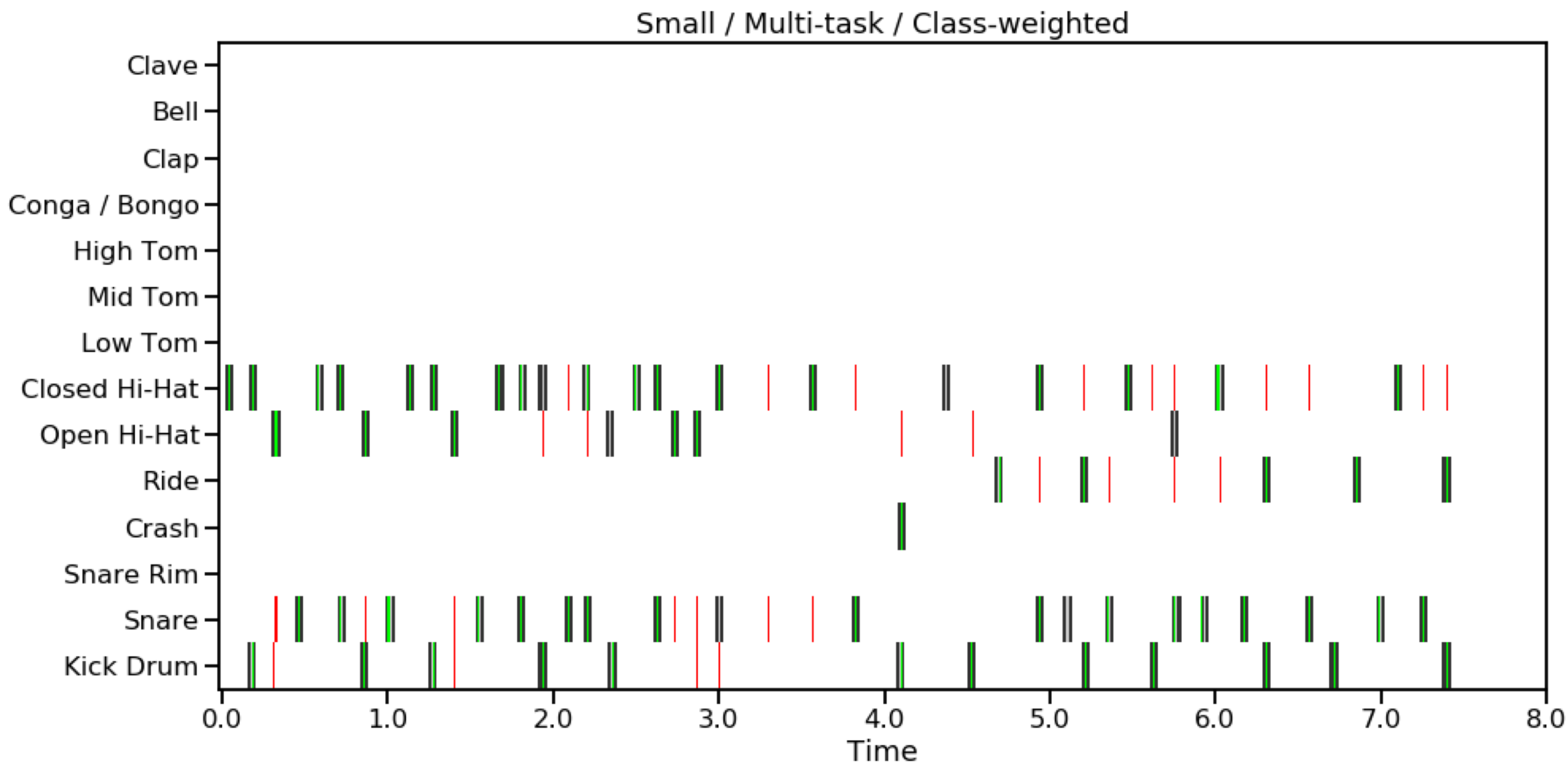

Training: Real | Evaluation: Real
1.0
0.8
0.6
0.4
0.2
0.0
F-measure
Precision
Recall
Kick
Snare
Snare (rim)
Crash
Ride
Open hh
Closed hh
Low tom
Mid tom
High tom
Conga
Hand clap
Bell
Clave
Training: Real + Synth | Evaluation: Real
1.0
0.8
0.6
0.4
0.2
0.0
Kick
Snare
Snare (rim)
Crash
Ride
Open hh
Closed hh
Low tom
Mid tom
High tom
Conga
Hand clap
Bell
Clave

Results



Training: Real | Evaluation: Real
F-measure
Precision
Recall
Macro F1: 0.20
Training: Real + Synth | Evaluation: Real
Macro F1: 0.35
Weighted - Training: Real + Synth | Evaluation: Real
Macro F1: 0.61
1.0
0.8
0.6
0.4
0.2
0.0
Kick
Snare
Snare (rim)
Crash
Ride
Open hh
Closed hh
Low tom
Mid tom
High tom
~~Conga~~
~~Hand clap~~
~~Bell~~
~~Clave~~

Results


Small / Multi-task / Class-weighted
Clave
Bell
Clap
Conga / Bongo
High Tom
Mid Tom
Low Tom
Closed Hi-Hat
Open Hi-Hat
Ride
Crash
Snare Rim
Snare
Kick Drum
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
Time

Conclusion

- Lots of mediocre synthetic data can improve performance on both overall performance and uncommon classes
- However, it must be used in conjunction with real music data
- Multi-task learning doesn't seem to help for large-vocab transcription, but does help in the auxiliary task of downbeat/beat tracking

Download the trained models at
https://github.com/mcartwright/dafx2018_adt