Generating Detailed Music Datasets with Neural Audio Synthesis

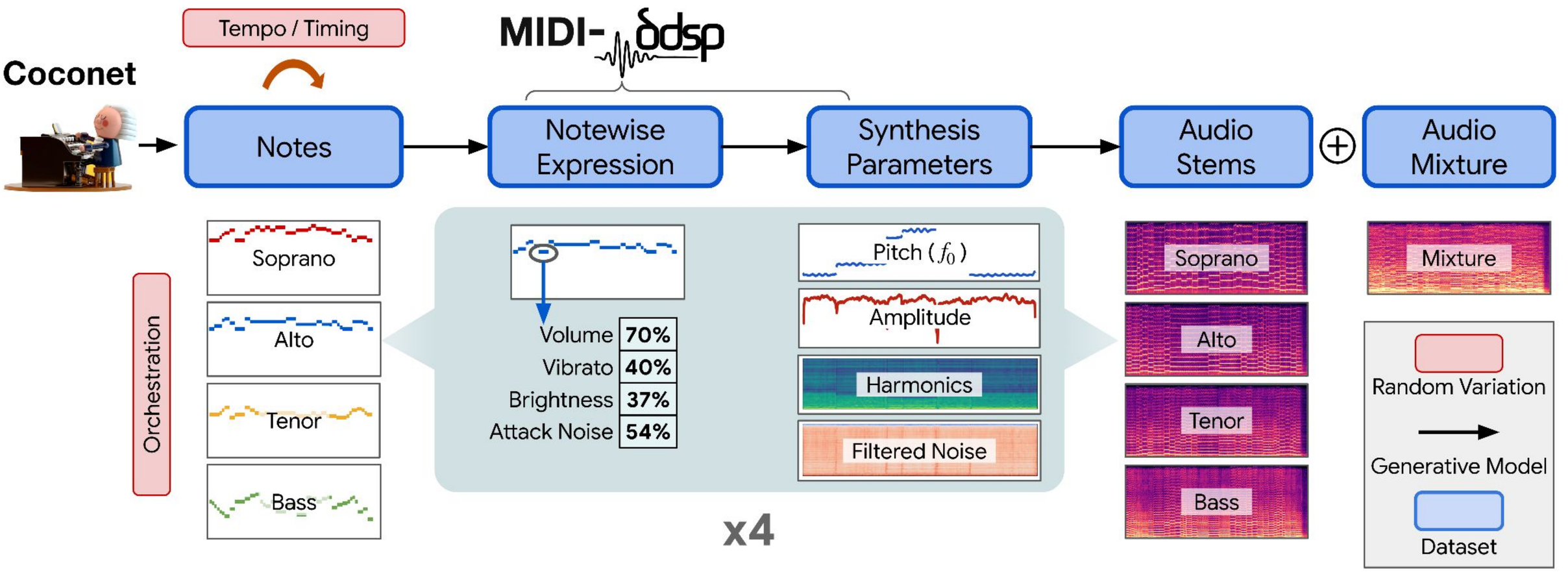
Yusong Wu, Josh Gardner, Ethan Manilow, Ian Simon,
Curtis Hawthorne, Jesse Engel

Structured Generative Models + Generation + Variation → Large Dataset with Detailed Labels

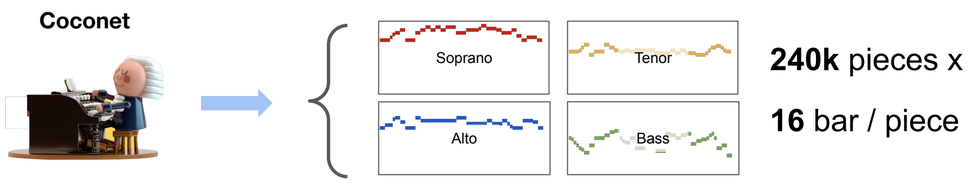
🎵 Coconet + 🎻🎷🎺 MIDI-DDSP + ⚙️ Variation → CocoChorales

- Coconet: Trained with 382 pieces
- MIDI-DDSP: Trained with 3.75 hours of solo (1 hours of mix)
- Variation: Instrumentation, Tempo, Expressive Timing
- **x 240,000**
- **x 1411 hours**
- \+ **Aligned Label**

Website | Code



MIDI Generation

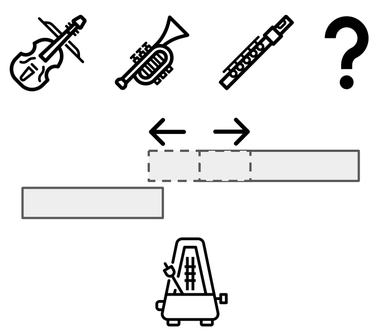


240k pieces x

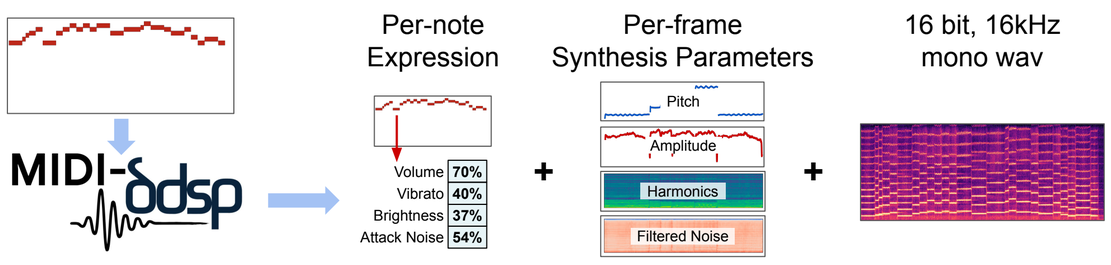
16 bar / piece

MIDI Variation

- Instrumentation Variation
  4 ensembles: String, Brass, Woodwind, Random
- Expressive Timing Variation
  Uniform([50,150]) BPM
- Tempo Variation
  N(0, 15^2) between [-50,50] ms

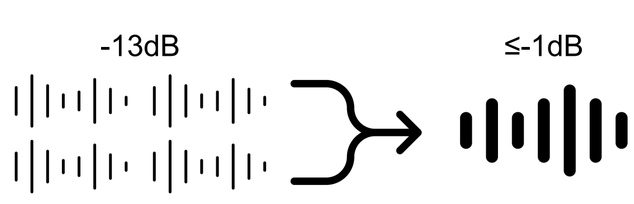


Audio Synthesis

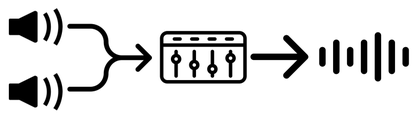


Mixing

- Simple sum of stems with loudness normalization, while ensuring mix doesn't clip.



More Variations and Applications Enabled by Structured Model

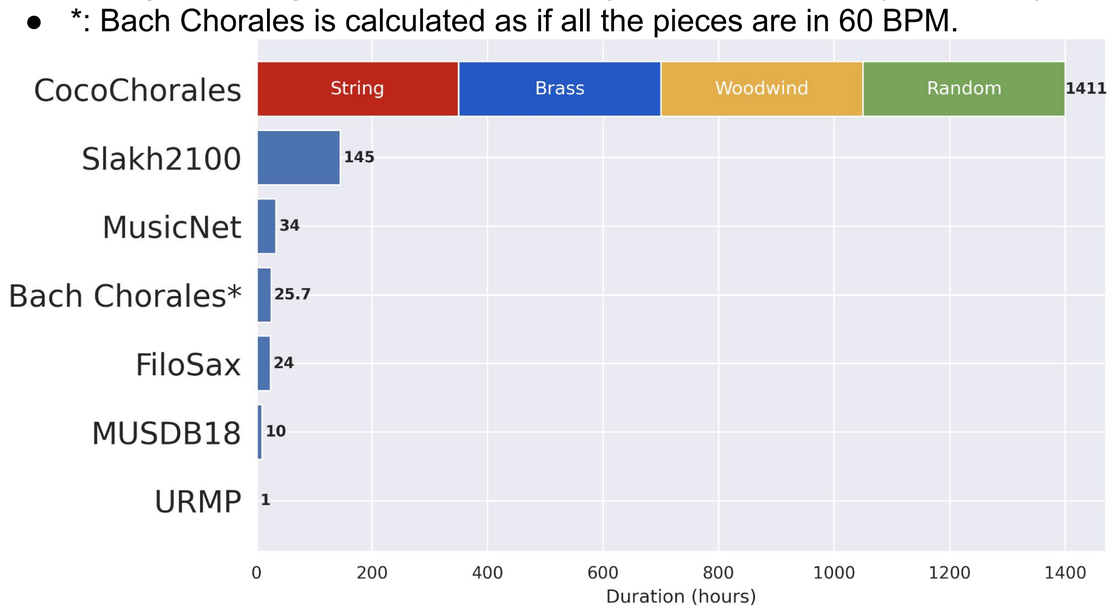
- Random Note Expression 
- Pitch augmentation 
- Random Mixing & Audio Effects 
- **Future applications**: multi-f0 transcription, performance analysis, random ensembles separation, similar soundings separation, etc.

Resources

- Sample website: https://chamber-ensemble-generator.github.io
- Code: https://github.com/lukewys/chamber-ensemble-generator
- Full dataset coming soon

CoCoChorales: 240k tracks (1400 hours) with stem, note, expression, and synthesis labels

- A magnitude larger than the previous synthesized dataset (Slakh2100).
- *: Bach Chorales is calculated as if all the pieces are in 60 BPM.



Transcription Experiments

- Multi-instrument transcription model (MT3) training on CocoChorales
- 😃 Significant performance improvement when training with only one dataset.

Training Dataset(s)	On/Off F1	Multi-Inst. F1
URMP	0.28	0.22
URMP + CocoChorales	**0.55**	**0.48**

- 😐 Performance gains does not transfer to other datasets when having enlarged source dataset (URMP).

Model	MAESTRO	Cerberus4	GuitarSet	MusicNet	Slakh2100	URMP
			Onset-Offset F1			
MT3 Datasets	0.83	0.80	0.78	0.33	0.57	0.61
+ CocoChorales	0.83	0.80	0.79	0.34	0.57	**0.66**
			Multi-Instrument F1			
MT3 Datasets	0.83	0.79	0.78	0.30	0.58	0.50
+ CocoChorales	0.83	0.75	0.79	0.30	0.57	**0.56**

Source Separation Experiments

- 😃 Successfully trained two source separation models for woodwind quartet (Flute, Oboe, Clarinet, Bassoon) which only has **several minutes** of data in URMP.

Network	Flute	Oboe	Clarinet	Bassoon
Demucs v2 [19]	18.7	17.3	21.0	20.7
MI+TR [62]	12.5	6.9	10.7	6.9

SI-SDR (dB)