Appendices

MIDI Index	MIDI Instrument	MIDI Program	Our Mapping	Our Index
0-3	Piano	Grand/Bright/Honky-tonk Piano	Piano	0
4-5	Piano	Electric Piano 1-2	Electric Piano	1
6	Piano	Harpsichord	Harpsichord	2
7	Piano	Clavinet	Clavinet	3
8-15	Chr. Percussion	Celesta, Glockenspiel, Music box, Vibraphone, Marimba, Xylophone, Tubular Bells, Dulcimer	Chr. Percussion	4
16-20	Organ	Drawbar, Percussive, Rock, Church, Reed Organ	Organ	5
21	Organ	Accordion	Accordion	6
22	Organ	Harmonica	Harmonica	7
23	Organ	Tango Accordion	Accordion	6
24-25	Guitar	Acoustic Guitar (nylon, steel)	Acoustic Guitar	8
26-31	Guitar	Electric Guitar (jazz, clean, muted, overdriven, distorted, harmonics)	Electric Guitar	9
32-39	Bass	Acoustic/Electric/Slap/Synth Bass	Bass	10
40	Strings	Violin	Violin	11
41	Strings	Viola	Viola	12
42	Strings	Cello	Cello	13
43	Strings	Contrabass	Contrabass	14
44	Strings	Tremolo Strings	Strings	15
45	Strings	Pizzicato Strings	Strings	15
46	Strings	Orchestral Harp	Harp	16
47	Strings	Timpani	Timpani	17
48-51	Ensemble	Acoustic/Synth String Ensemble 1-2	Strings	15
52-54	Ensemble	Aahs/Oohs/Synth Voice	Voice	18
55	Ensemble	Orchestra Hit	Strings	15
56	Brass	Trumpet	Trumpet	19
57	Brass	Trombone	Trombone	20
58	Brass	Tuba	Tuba	21
59	Brass	Muted Trumpet	Trumpet	19
60	Brass	French Horn	Horn	22
61-63	Brass	Acoustic/Synth Brass	Brass	23
64-67	Reed	Soprano, Alto, Tenor, Baritone Sax	Saxophone	24
68	Reed	Oboe	Oboe	25
69	Reed	English Horn	Horn	22
70	Reed	Bassoon	Bassoon	26
71	Reed	Clarinet	Clarinet	27
72	Pipe	Piccolo	Piccolo	28
73	Pipe	Flute	Flute	29
74	Pipe	Recorder	Recorder	30
75-79	Pipe	Pan Flute, Blown bottle, Shakuhachi, Whistle, Ocarina	Pipe	31
80-87	Synth Lead	Lead 1-8	Synth Lead	32
88-95	Synth Pad	Pad 1-8	Synth Pad	33
96-103	Synth Effects	FX 1-8	Synth Effects	34
104-111	Ethnic	Sitar, Banjo, Shamisen, Koto, Kalimba, Bagpipe, Fiddle, Shana	Ethnic	35
112-119	Percussive	Tinkle Bell, Agogo, Steel Drums, Woodblock, Taiko Drum, Melodic Tom, Synth Drum	Percussive	36
120-127	Sound Effects	Guitar Fret Noise, Breath Noise, Seashore, Bird Tweet, Telephone Ring, Helicopter, Applause, Gunshot	Sound Effects	37
128	Drums	Drums	Drums	38

Table 7. The instrument mapping used in our experiments. Our mapping is less detailed than the MIDI Program Number, but it is finer than the MIDI Instrument code, thus resulting in 39 different instruments.

		Full length SDR			10s SDR		
feature merge	Model	inst	piece	source	inst	piece	source
sum	T+S (binary)	1.35	3.22	2.97	3.14	5.32	4.31
	T+S (posterior)	1.86	3.55	3.32	3.47	5.60	4.54
	preT+S (binary)	0.20	2.75	2.44	1.86	5.34	4.11
	preT+S (posterior)	2.01	3.72	3.50	**3.69**	5.86	4.81
	T+S+STE (binary)	1.30	3.30	3.06	2.40	5.13	4.16
	T+S+STE (posterior)	1.45	3.31	3.10	2.41	4.96	4.07
	preT+S+STE (binary)	1.67	3.30	3.06	3.29	5.36	4.41
	preT+S+STE (posterior)	1.99	3.42	3.24	3.26	5.17	4.29
concat	T+S (binary)	1.28	3.20	2.95	2.72	5.06	4.20
	T+S (posterior)	1.80	3.53	3.31	3.21	5.50	4.51
	preT+S (binary)	-0.04	2.63	2.31	1.86	5.44	4.09
	preT+S (posterior)	1.92	**3.75**	**3.52**	3.50	**6.03**	**4.90**
	T+S+STE (binary)	1.89	3.60	3.37	3.63	5.70	4.66
	T+S+STE (posterior)	2.01	3.58	3.37	3.59	5.53	4.55
	preT+S+STE (binary)	1.72	3.47	3.22	3.18	5.78	4.61
	preT+S+STE (posterior)	**2.13**	3.66	3.46	1.86	5.44	4.09
spec patch	T+S (binary)	1.74	3.42	3.20	3.22	5.19	4.37
	T+S (posterior)	1.79	3.46	3.24	3.35	5.30	4.43
	preT+S (binary)	0.38	2.72	2.41	1.91	3.60	3.39
	preT+S (posterior)	1.91	3.61	3.40	3.58	5.52	4.65
	T+S+STE (binary)	1.74	3.46	3.24	3.50	5.28	4.42
	T+S+STE (posterior)	1.94	3.48	3.27	3.48	5.15	4.30
	preT+S+STE (binary)	1.71	3.47	3.23	3.28	5.40	4.53
	preT+S+STE (posterior)	2.01	3.58	3.38	3.57	5.30	4.48

Table 8. Full version of Table 4. 'spec patch' represents the model variation where the final learned mask is applied to the summation of the piano roll features and the spectrograms instead of the original spectrograms. We also experimented with two forms of piano rolls: posteriorgram and binary. Posteriorgram provides the probability of the existence of a note in the input audio to the model, which outperforms the model that uses binary version of the piano roll.

Task	Input length	(k, d)	(h_k, h_d)
Downbeat	6 seconds	(128, 96)	(4, 8)
Chord	12 seconds	(64, 256)	(8, 8)
Key	36 seconds	(128, 32)	(8, 4)

Table 9. SpecTNT parameters we used in each task, where k and d denote spectral and temporal feature dimensions; while h_k and h_d represent the number of heads for the spectral and temporal Transformer encoders, respectively.