APPENDIX: FAST AND FLEXIBLE NEURAL AUDIO SYNTHESIS

1. LOUDNESS CALCULATIONS

To calculate loudness, we use librosa's perceptual_weighting() function on the square of the STFT. This produces a spectrum in dB, which we convert back into a linear scale and compute a mean over frequency bins. This value is then scaled via log compression with a small offset $eps = 1e - 5$ to prevent overflow. Like any log scaling, this step squashes loud transients together, whilst expanding the dynamic range available for medium to low volume portions of the sound. The loudness vector, like energy, has length 1000 for 4 seconds of audio and is centered using the mean and standard deviation of the entire dataset. The range of values is approximately $[-1.6, 1.6]$.

2. DETAILS OF PITCH AND INSTRUMENT FAMILY CLASSIFIER

We train a multi-task classification model to do pitch and instrument family classification on the entire NSynth dataset. We use a 2D CNN architecture on log magnitude spectrograms of the audio, with a fft size of 1024 and hop size of 512 and hann window. The full model structure can be seen in Figure 1, where we use a a softmax-crossentropy loss for the labels as they are mutually exclusive.

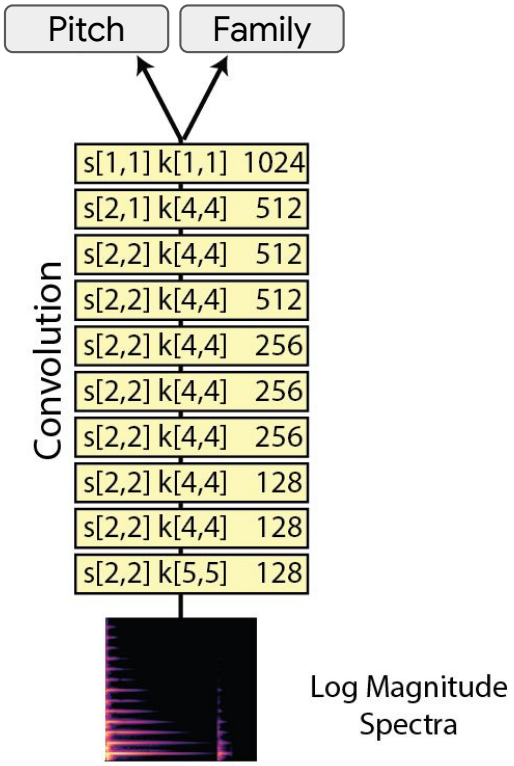

Figure 1. Model architecture for pitch and instrument family classification. Stride is denoted by s and kernel size is given by k. Each convolution layer is followed by batch normalization and a Leaky-ReLU nonlinearity (0.1 off-slope).

	Signal Processing	CREPE		NSynth Classifier	
	Loudness (L_1)	**F0** (L_1)	**F0 Outliers**	**Pitch Error**	**Family Error**
Loudness	**0.14**	6.20	0.57	0.98	**0.75**
Pitch/Cents	0.47	**0.93**	**0.07**	**0.14**	0.82
Energy + Pitch/Cents	0.23	1.67	0.09	0.18	0.72
Loudness + Pitch/Cents	**0.10**	**0.94**	**0.06**	**0.16**	**0.53**

Table 1. Comparing resynthesis metrics for Energy vs. Loudness and just Loudness or Pitch conditioning.