

LIST OF FIGURES

1	The automatic music transcription setup to be used in this thesis. Using per-instrument piano-roll representations is easier for machines to process, and avoids variability and subjectivity that may arise from symbolic and textual notations.	2
2	A generative model has to know all of necessary information required to reconstruct the audio data, including pitch, timbre, loudness, and duration. Generative models can be jointly trained with an encoder that finds those semantic information, giving a transcriber-synthesizer pair.	3
3	The full score notation of the music used to build the piano rolls in Figure 1. To fully recover this level of notations from the audio, the transcriber has to make many additional decisions than for the piano rolls, such as determining the key signature, time signature, clefs, dynamics, trills, bowing instructions, etc.	7
4	Increasingly realistic qualities of the generated faces using generative adversarial networks as shown in (Brundage et al., 2018); images are taken from (Goodfellow et al., 2014), (Radford et al., 2015), (Liu & Tuzel, 2016), (Karras et al., 2018), and (Karras et al., 2019).	12
5	The standard pipeline for music feature extraction. An appropriate set of feature extraction methods needs to be heuristically selected depending on the task.	17
6	The input and output representations of the CREPE model. The input is 1024 time-domain samples, and the output is a 360-dimensional vector that contains a Gaussian curve centered at the ground-truth frequency (see Equation 16). For details on the convolutional architecture used in the model, see Figure 7.	52
7	The architecture of the CREPE pitch tracker. The six convolutional layers predicts the Gaussian curve centered at the ground-truth frequency. The output is then used to extract the exact pitch estimate as in Equation 14-15.	53
8	Pitch tracking performance when additive noise signals are present. The error bars are centered at the average raw pitch accuracies and span the first standard deviations. With brown noise being a notable exception, CREPE shows the highest noise robustness in general.	60

9	Fourier spectra of the first-layer filters sorted by the frequency of the peak magnitude. Histograms on the right show the distribution of ground-truth frequencies in the corresponding dataset.	61
10	The raw pitch accuracy (RPA) of CREPE’s predictions on each of the 230 tracks in MDB-stem-synth with respect to the instrument, sorted by the average frequency.	61
11	The first layer filters of the CREPE model trained with six different datasets, visualized using the same method as in Figure 9. Compared to the previous cases where the filters are specialized to one dataset, the peak-frequency curve is smoother and covers the full frequency range.	64
12	Visualization of the target activation, which can be saved as in image file by using the <code>--save-activation</code> option in the CREPE command-line interface. The audio clip used contains an excerpt of male singing voice.	65
13	An example of an inaccurate target output where there exist multiple peaks around the harmonics of the ground-truth frequency. Equation 18 ensures that the predicted pitch is calculated using only the frequency bins around the highest peak.	67
14	The overall architecture of the proposed Mel2Mel model. The note sequences and instrument embeddings are combined to predict the Mel spectrogram, which is then fed to the WaveNet vocoder.	77
15	Pearson correlations between the reconstructed and original audio. (a) each stage of degradation. (b) per-instrument breakdown of the green curve on the left. The curves are smoothed and drawn with respect to each octave for readability.	81
16	Visualization of the embedding space trained in 2 dimensions, using spectral centroids and mean energy. The continuous colors are obtained for each pixel in the 320-by-320 grid and are related to the spectral and temporal envelopes of the timbres.	83
17	A <i>t</i> -SNE visualization of the 10-dimensional timbre embedding space learned using 100 instruments, color-coded according to the spectral centroid. In the web demo, users can rotate the dots to navigate and click on the dots to play the corresponding audio segments.	84
18	The Onset and Frames model. CNN denotes the convolutional acoustic model taken from (Kelz et al., 2016), FC denotes a fully connected layer, and σ denotes sigmoid activation. Dotted lines mean stop-gradient, i.e. no backpropagation.	91
19	A computation graph showing how a discriminator is appended to the original model. The appended parts are shown as dotted components.	95

20	Comparisons of the frame activations predicted by the baseline and our model ($\ell = \text{BCE}$, $\alpha = 0.3$), on three example segments. The input Mel spectrograms and the target piano rolls are shown together. The GAN version produces more confident predictions compared to the noisy baselines, leading to more accurate predictions.	102
21	Distribution of the F1 score improvements over the baseline, tested on the MAESTRO test tracks.	103
22	Distribution of frame activation values. Our model outputs more confident predictions, as indicated by the lower relative frequency in (0.1, 0.9).	103
23	Learning curves showing the generalization gaps; training curves are drawn as dotted lines, and test curves are drawn as solid lines.	104
24	The synthesizer architecture. The temporal and spectral characteristics of each notes are respectively modeled by the envelope estimator and the waveshaper. This structure enforces each instrument to have consistent spectral and temporal envelopes across the pitch.	116
25	An example of the input, target, and output representations of the synthesizer. The output resembles the target Mel spectrogram but shows a regular pattern as constrained by the synthesizer model.	119
26	The learned timbre embedding space mapping the 11 instruments in the MusicNet dataset into the corresponding points in the space. Types of instruments (keyboards, winds, and strings) are color-coded.	120
27	Examples of the frame predictions by the baseline and the proposed model compared to the ground-truth, showing their qualitative differences in performance.	122
28	An example of multi-instrument transcription of a violin piece accompanied by piano (MusicNet track 2628), where the predictions of keyboard and string instruments are color-coded in cyan and magenta, respectively.	124
29	The MusicNet Inspector web interface. Users can select among the 330 tracks in the MusicNet dataset, and the selected track is played together with the CQT, piano roll, and amplitude visualizations.	125
30	A combination of music language model and music synthesizer can be used as an infinite source of accurately annotated training data for a transcription model.	131
31	The <i>transcriptional Turing test</i> , to test whether an automatic music transcription algorithm has reached human-level. When an automatically generated transcription is indistinguishable by an expert human listener from one produced by a skilled musician, it can be said that AMT is solved.	132

“What I cannot create, I do not understand.”

-Richard Phillips Feynman