where $V_{*,k}$ is a learnable linear projection, and the vector $V_{*,k}^T \mathbf{h}$ is broadcast over the time dimension.

For local conditioning we have a second timeseries h_t, possibly with a lower sampling frequency than the audio signal, e.g. linguistic features in a TTS model. We first transform this time series using a transposed convolutional network (learned upsampling) that maps it to a new time series $\mathbf{y} = f(\mathbf{h})$ with the same resolution as the audio signal, which is then used in the activation unit as follows:

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y}\right),$$

where $V_{f,k} * \mathbf{y}$ is now a 1×1 convolution. As an alternative to the transposed convolutional network, it is also possible to use $V_{f,k} * \mathbf{h}$ and repeat these values across time. We saw that this worked slightly worse in our experiments.

2.6 CONTEXT STACKS

We have already mentioned several different ways to increase the receptive field size of a WaveNet: increasing the number of dilation stages, using more layers, larger filters, greater dilation factors, or a combination thereof. A complementary approach is to use a separate, smaller *context* stack that processes a long part of the audio signal and locally conditions a larger WaveNet that processes only a smaller part of the audio signal (cropped at the end). One can use multiple context stacks with varying lengths and numbers of hidden units. Stacks with larger receptive fields have fewer units per layer. Context stacks can also have pooling layers to run at a lower frequency. This keeps the computational requirements at a reasonable level and is consistent with the intuition that less capacity is required to model temporal correlations at longer timescales.

3 EXPERIMENTS

To measure WaveNet's audio modelling performance, we evaluate it on three different tasks: multi-speaker speech generation (not conditioned on text), TTS, and music audio modelling. We provide samples drawn from WaveNet for these experiments on the accompanying webpage:
`https://www.deepmind.com/blog/wavenet-generative-model-raw-audio/`.

3.1 MULTI-SPEAKER SPEECH GENERATION

For the first experiment we looked at free-form speech generation (not conditioned on text). We used the English multi-speaker corpus from CSTR voice cloning toolkit (VCTK) (Yamagishi, 2012) and conditioned WaveNet only on the speaker. The conditioning was applied by feeding the speaker ID to the model in the form of a one-hot vector. The dataset consisted of 44 hours of data from 109 different speakers.

Because the model is not conditioned on text, it generates non-existent but human language-like words in a smooth way with realistic sounding intonations. This is similar to generative models of language or images, where samples look realistic at first glance, but are clearly unnatural upon closer inspection. The lack of long range coherence is partly due to the limited size of the model's receptive field (about 300 milliseconds), which means it can only remember the last 2–3 phonemes it produced.

A single WaveNet was able to model speech from any of the speakers by conditioning it on a one-hot encoding of a speaker. This confirms that it is powerful enough to capture the characteristics of all 109 speakers from the dataset in a single model. We observed that adding speakers resulted in better validation set performance compared to training solely on a single speaker. This suggests that WaveNet's internal representation was shared among multiple speakers.

Finally, we observed that the model also picked up on other characteristics in the audio apart from the voice itself. For instance, it also mimicked the acoustics and recording quality, as well as the breathing and mouth movements of the speakers.