- In order to deal with long-range temporal dependencies needed for raw audio generation, we develop new architectures based on dilated causal convolutions, which exhibit very large receptive fields.
- We show that when conditioned on a speaker identity, a single model can be used to generate different voices.
- The same architecture shows strong results when tested on a small speech recognition dataset, and is promising when used to generate other audio modalities such as music.

We believe that WaveNets provide a generic and flexible framework for tackling many applications that rely on audio generation (e.g. TTS, music, speech enhancement, voice conversion, source separation).

2 WaveNet

In this paper we introduce a new generative model operating directly on the raw audio waveform. The joint probability of a waveform $\mathbf{x} = \{x_1, \ldots, x_T\}$ is factorised as a product of conditional probabilities as follows:

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}) \tag{1}$$

Each audio sample x_t is therefore conditioned on the samples at all previous timesteps.

Similarly to PixelCNNs (van den Oord et al., 2016a;b), the conditional probability distribution is modelled by a stack of convolutional layers. There are no pooling layers in the network, and the output of the model has the same time dimensionality as the input. The model outputs a categorical distribution over the next value x_t with a softmax layer and it is optimized to maximize the log-likelihood of the data w.r.t. the parameters. Because log-likelihoods are tractable, we tune hyper-parameters on a validation set and can easily measure if the model is overfitting or underfitting.

2.1 Dilated Causal Convolutions

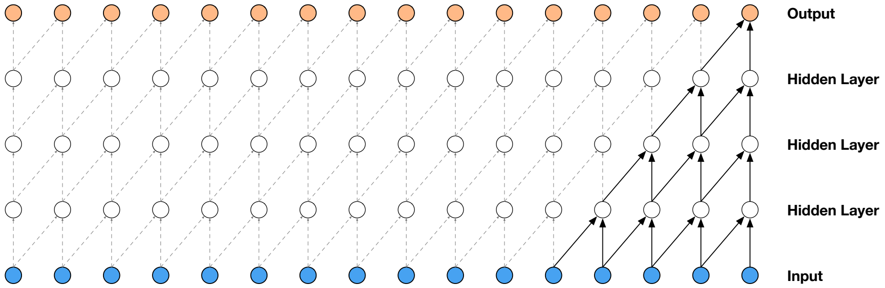

Figure 2: Visualization of a stack of causal convolutional layers.

The main ingredient of WaveNet are causal convolutions. By using causal convolutions, we make sure the model cannot violate the ordering in which we model the data: the prediction $p(x_{t+1} \mid x_1, ..., x_t)$ emitted by the model at timestep t cannot depend on any of the future timesteps $x_{t+1}, x_{t+2}, \ldots, x_T$ as shown in Fig. 2. For images, the equivalent of a causal convolution is a masked convolution (van den Oord et al., 2016a) which can be implemented by constructing a mask tensor and doing an elementwise multiplication of this mask with the convolution kernel before applying it. For 1-D data such as audio one can more easily implement this by shifting the output of a normal convolution by a few timesteps.

At training time, the conditional predictions for all timesteps can be made in parallel because all timesteps of ground truth $\mathbf{x}$ are known. When generating with the model, the predictions are sequential: after each sample is predicted, it is fed back into the network to predict the next sample.