- the MagnaTagATune dataset (Law & Von Ahn, 2009), which consists of about 200 hours of music audio. Each 29-second clip is annotated with tags from a set of 188, which describe the genre, instrumentation, tempo, volume and mood of the music.
- the YouTube piano dataset, which consists of about 60 hours of solo piano music obtained from YouTube videos. Because it is constrained to a single instrument, it is considerably easier to model.

Although it is difficult to quantitatively evaluate these models, a subjective evaluation is possible by listening to the samples they produce. We found that enlarging the receptive field was crucial to obtain samples that sounded musical. Even with a receptive field of several seconds, the models did not enforce long-range consistency which resulted in second-to-second variations in genre, instrumentation, volume and sound quality. Nevertheless, the samples were often harmonic and aesthetically pleasing, even when produced by unconditional models.

Of particular interest are conditional music models, which can generate music given a set of tags specifying e.g. genre or instruments. Similarly to conditional speech models, we insert biases that depend on a binary vector representation of the tags associated with each training clip. This makes it possible to control various aspects of the output of the model when sampling, by feeding in a binary vector that encodes the desired properties of the samples. We have trained such models on the MagnaTagATune dataset; although the tag data bundled with the dataset was relatively noisy and had many omissions, after cleaning it up by merging similar tags and removing those with too few associated clips, we found this approach to work reasonably well.

### 3.4 Speech Recognition

Although WaveNet was designed as a generative model, it can straightforwardly be adapted to discriminative audio tasks such as speech recognition.

Traditionally, speech recognition research has largely focused on using log mel-filterbank energies or mel-frequency cepstral coefficients (MFCCs), but has been moving to raw audio recently (Palaz et al., 2013; Tüske et al., 2014; Hoshen et al., 2015; Sainath et al., 2015). Recurrent neural networks such as LSTM-RNNs (Hochreiter & Schmidhuber, 1997) have been a key component in these new speech classification pipelines, because they allow for building models with long range contexts. With WaveNets we have shown that layers of dilated convolutions allow the receptive field to grow longer in a much cheaper way than using LSTM units.

As a last experiment we looked at speech recognition with WaveNets on the TIMIT (Garofolo et al., 1993) dataset. For this task we added a mean-pooling layer after the dilated convolutions that aggregated the activations to coarser frames spanning 10 milliseconds ($160\times$ downsampling). The pooling layer was followed by a few non-causal convolutions. We trained WaveNet with two loss terms, one to predict the next sample and one to classify the frame, the model generalized better than with a single loss and achieved $18.8$ PER on the test set, which is to our knowledge the best score obtained from a model trained directly on raw audio on TIMIT.

## 4 Conclusion

This paper has presented WaveNet, a deep generative model of audio data that operates directly at the waveform level. WaveNets are autoregressive and combine causal filters with dilated convolutions to allow their receptive fields to grow exponentially with depth, which is important to model the long-range temporal dependencies in audio signals. We have shown how WaveNets can be conditioned on other inputs in a global (e.g. speaker identity) or local way (e.g. linguistic features). When applied to TTS, WaveNets produced samples that outperform the current best TTS systems in subjective naturalness. Finally, WaveNets showed very promising results when applied to music audio modeling and speech recognition.

## Acknowledgements

The authors would like to thank Lasse Espeholt, Jeffrey De Fauw and Grzegorz Swirszcz for their inputs, Adam Cain, Max Cant and Adrian Bolton for their help with artwork, Helen King, Steven