DeepMind

Speech bandwidth extension with WaveNet

Archit Gupta, Brendan Shillingford, Yannis Assael, Thomas C. Walters

Introduction

• We tackle the problem of bandwidth extension, with the goal of improving the quality of speech at low sample rate, which may have passed through a low-rate codec.

• We propose a model based on WaveNet, a deep generative model of audio waveforms [1].

• WaveNet has been shown to be extremely effective at synthesizing high quality speech when conditioned on linguistic features.

• The WaveNet architecture has also been conditioned on log-mel spectrograms for text-to-speech and other low-dimensional latent representations for speech coding.

• Here we demonstrate that a WaveNet model conditioned on log-mel spectrograms from low sample rate speech can be used to generate high sample-rate speech of good quality, when compared with audio passed through the 'HD-voice' AMR-WB codec.

• There has been recent interest [2, 3, 4] in predicting upsampled waveforms directly from low-rate speech waveforms.

Model Architecture and Training

WaveNet is a generative model that models the joint probability of a waveform $x = \{x_1, \ldots, x_T\}$ as a product of conditional probabilities given the samples at previous timesteps. A conditional WaveNet model takes an additional input variable h and models this conditional distribution as:

$$p(\mathbf{x}|\mathbf{h}) = \prod_{t=1}^{T} p\left(x_t|x_1, \ldots, x_{t-1}, \mathbf{h}\right)$$

A conditional WaveNet model is used in this task.

The model is trained using maximum likelihood to predict the 24kHz waveform from log mel-spectrograms computed from the 8kHz band-limited waveform. There are two types of inputs to the model during training, the autoregressive inputs containing the sample from the previous timestep, and the conditioning inputs.

The autoregressive inputs during training are teacher-forced, and were therefore fed high quality 24kHz audio samples. We compute log-mel spectrograms from the lower bandwidth audio as conditioning inputs. In other words, the WaveNet described previously then models:

$$p(\mathbf{x}_{\text{hi}}|\mathbf{x}_{\text{lo}}) = \prod_{t=1}^{T} p\left(x_{\text{hi},t}|x_{\text{hi},1}, \ldots, x_{\text{hi},t-1}, \mathbf{x}_{\text{lo}}\right)$$

where x_{hi} is the autoregressively modelled 24kHz waveform, and x_{lo} is the 8kHz band-limited version, represented as a log-mel spectrogram. The x_{lo} is used as input in the WaveNet conditioning stack.

This WaveNet model, while it has a lot of modelling power, is very slow at inference time, due to its autoregressive nature. In further work we wish to investigate architectures which are more amenable to fast inference, such as WaveGlow and WaveRNN.

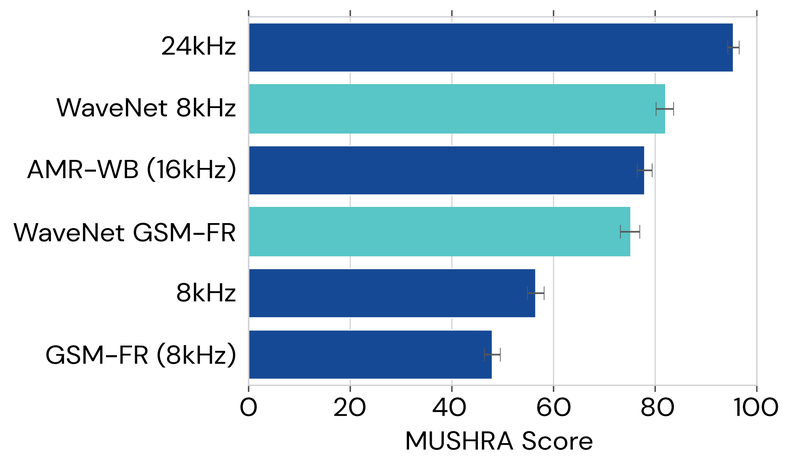
Results

• We evaluate models using the MUSHRA listening test methodology.

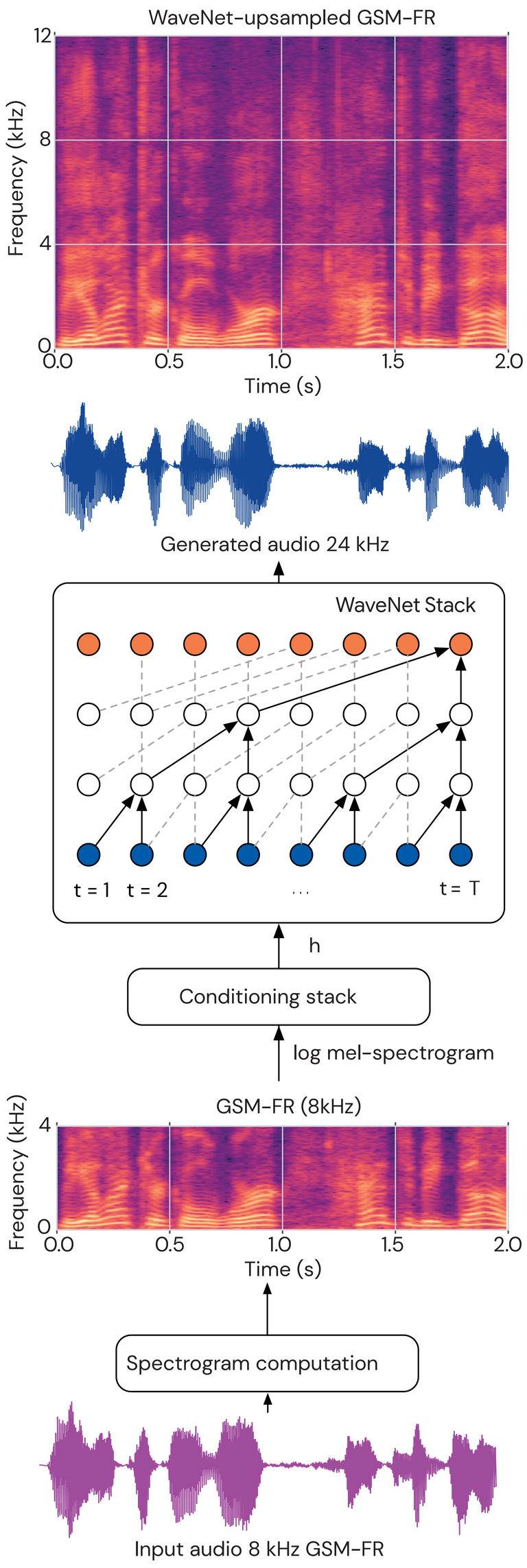
• Human listeners rated samples from the WaveNet models alongside low rate, compressed and reference high sample rate audio on a 100-point scale.

• The model trained to predict to 24kHz from 8kHz audio directly performs better than the AMR-WB codec.

• The model predicting 24kHz from GSM encoded 8kHz performs only slightly worse than AMR-WB.



Experiment Architecture



Training Data

The LibriTTS corpus is used for training and evaluaton. This corpus is derived from user sourced audio in the LibriSpeech dataset. The source audio is sampled at 24kHz and consists of a range of English voices. The models evaluated in this paper were trained on the train-clean-100 subset of the corpus and evaluated on utterances drawn from the test-clean subset.

Subset	Hours	Male Speakers	Female Speakers
test-clean	8.56	19	20
train-clean-100	53.78	123	124

References

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio." in SSW, 2016, p. 125.

[2] Kuleshov, S. Z. Enam, and S. Ermon, "Audio super resolution using neural networks,"arXiv preprint arXiv:1708.00853, 2017.

[3] Z.-H. Ling, Y. Ai, Y. Gu, and L.-R. Dai, "Waveform modeling and generation using hierarchical recurrent neural networksfor speech bandwidth extension,"IEEE/ACM Transactions onAudio, Speech, and Language Processing, vol. 26, no. 5, pp.883–894, 2018.

[4] Y. Gu and Z.-H. Ling, "Waveform modeling using stacked di-lated convolutional neural networks for speech bandwidth extension." in INTERSPEECH, 2017, pp. 1123–1127.