LOW BIT-RATE SPEECH CODING WITH VQ-VAE AND A WAVENET DECODER

Cristina Gârbacea[1,2], Aäron van den Oord[2], Yazhe Li[2], Felicia S C Lim[3], Alejandro Luebs[3], Oriol Vinyals[2], Thomas C Walters[2]
[1]University of Michigan, Ann Arbor (this work undertaken at DeepMind), [2]DeepMind, London, [3]Google, San Francisco

DeepMind

BACKGROUND + RELATED WORK

- van den Oord et al. [1] introduce the vector quantized variational autoencoder (VQ-VAE) which can reconstruct speech with high quality after passing through a constrained latent representation by sampling from a WaveNet-like decoder.
- Kleijn et al. [2] use a learned WaveNet decoder to produce audio comparable in quality to that produced by the AMR-WB codec when conditioned on a low-rate representation and 2.4kbps.
- Kuyk et al. [3] compute the true information rate of speech to be less than 100 bps, yet current systems require a rate roughly two orders of magnitude higher than this to produce good quality speech.
- In this work, we attempt to use the VQ-VAE with WaveNet deocder as an end-to-end learned speech codec, to better compress speech.

THIS WORK

Can VQ-VAE be used as an end-to-end speech codec?

- VQ-VAE extracts a compact and semantically meaningful representation of the input, capturing high-level speech features.
- The model can generate very high-quality speech even when using a latent representation that is many times smaller than the original waveform.

However, various aspects of the original architecture need modifying to make a good codec:

Maintaining Speaker Identity

VQ-VAE conditions the encoder and decoder on speaker.

- We remove explicit conditioning on speaker identity and replaced it with a latent representation that does not vary over time and takes its input from the whole utterance.

Constraining Prosody

VQ-VAE produces utterances where the semantic content is preserved but the prosody can differ.

- For consistency of prosody between the source and reconstructed waveforms, we need to ensure that pitch and timing information are preserved and passed to the decoder.
- We add a second decoder in parallel with the WaveNet decoder which can predict the f_0 track of the utterance.
- We add an f_0 prediction term with tunable weight to the loss function causing the latent representation to pass pitch and timing information through the bottleneck.

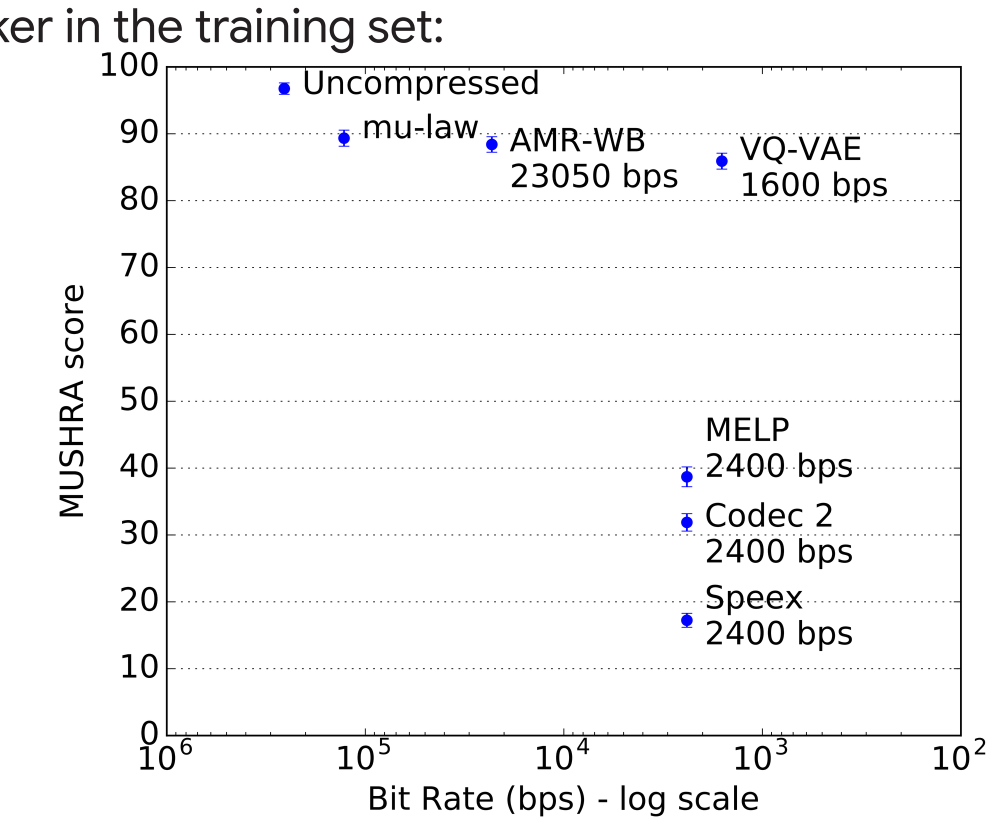
EXPERIMENTS AND RESULTS

We trained VQ-VAE models on two different corpora of speech. 'Studio' - high-quality recordings, and LibriSpeech - user-recorded audiobooks.

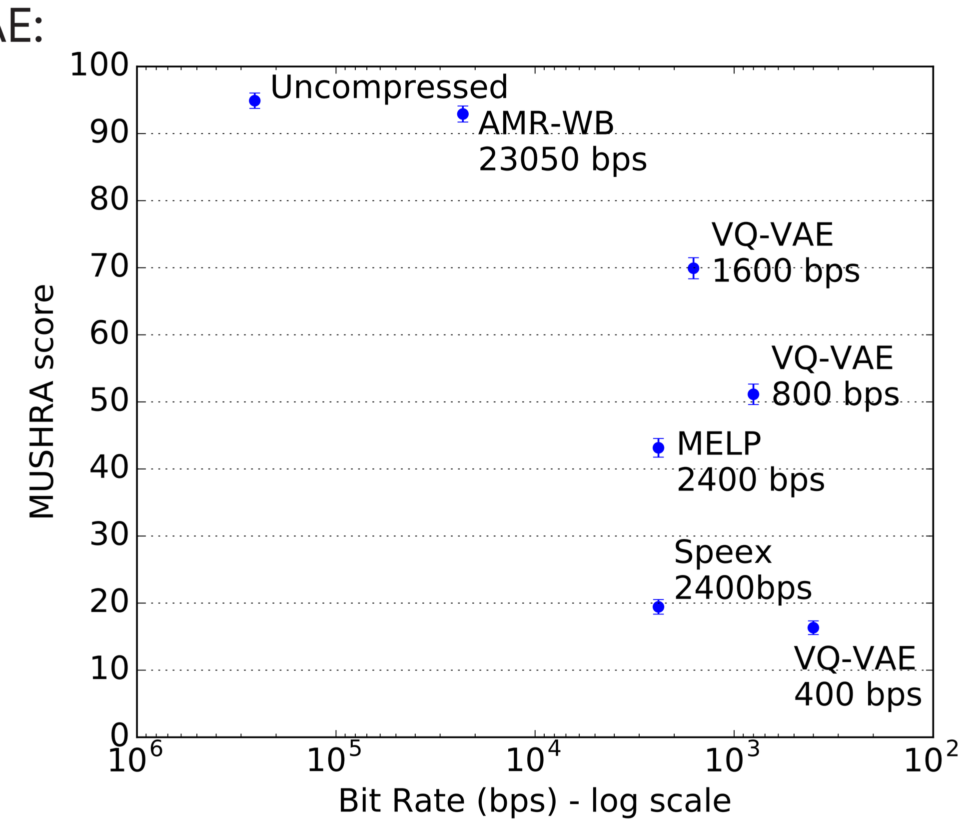
Initial experiments determined a good set of hyperparameters that led to low-rate coding and good reconstruction quality (see 'Model Architecture', right).

Model quality was evaluated using MUSHRA tests, comparing against other codecs at high and low rates.

Ideal conditions. Train on the Studio corpus with the test speaker in the training set:

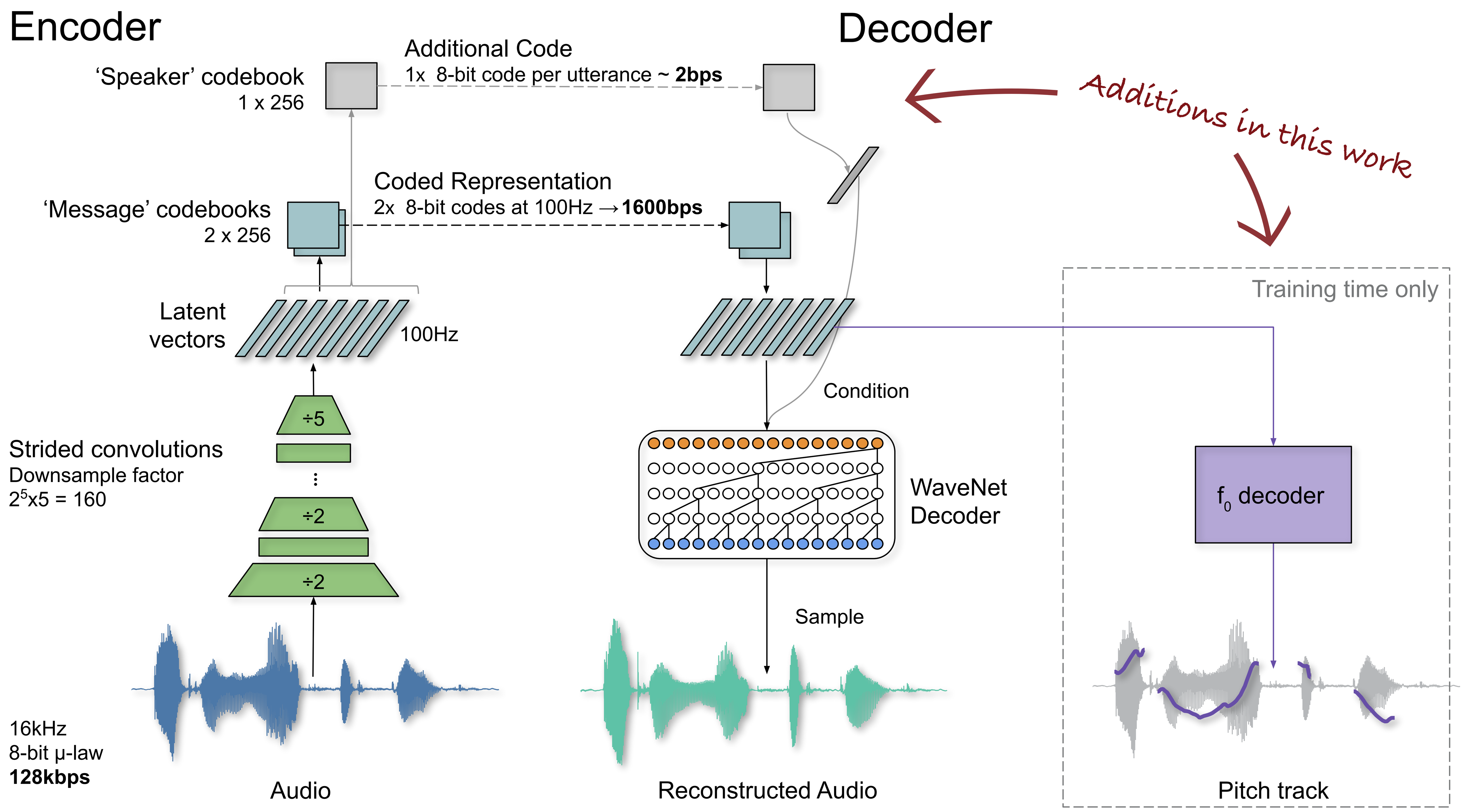


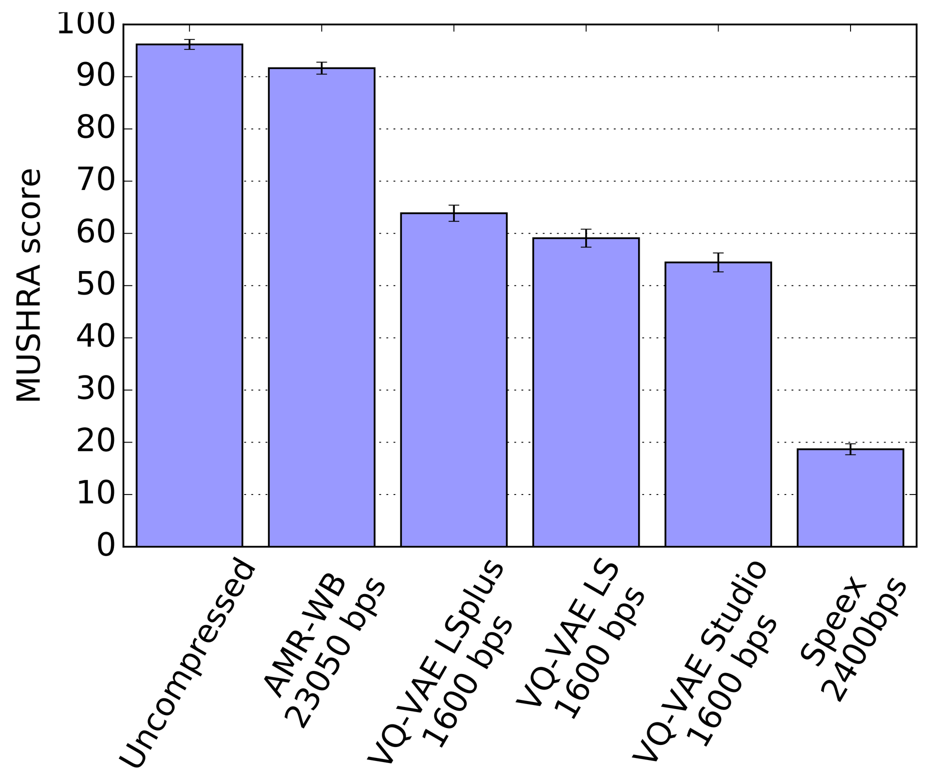
More realistic conditions. Train on the LibriSpeech corpus, with the test speaker held out. Evaluate different bit-rates for VQ-VAE:



Evaluate the effect of training corpus. Evaluate on LibriSpeech test set voices. Model trained on LibriSpeech plus test set voice (LSPlus), just LibriSpeech (LS) or Studio corpus:

MODEL ARCHITECTURE





Finally, evaluate mean opinion scores for speaker similarity:

Codec	Speaker Similarity (MOS)
VQ-VAE LSPlus 1600bps	3.794 ± 0.451
VQ-VAE LS 1600 bps	3.703 ± 0.716
MELP 2400 bps	3.138 ± 0.324
Speex 2400 bps	2.534 ± 0.233

CONCLUSIONS

- The VQ-VAE neural network with a WaveNet decoder can perform very low rate speech coding with high reconstruction quality.
- VQ-VAE coding speech at 1.6 kbps can produce output of similar perceptual quality to that generated by AMR-WB at 23.05 kbps when trained and tested on studio quality data.
- A prosody-transparent and speaker-independent model trained on the LibriSpeech corpus coding audio at 1.6 kbps exhibits perceptual quality which is around halfway between the MELP codec at 2.4 kbps and AMR-WB codec at 23.05 kbps.
- Speaker identity is perserved at least as well as other low-rate codecs.

REFERENCES

[1] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu, "Neural discrete representation learning," NIPS 2017.
[2] W Bastiaan Kleijn, Felicia S C Lim, Alejandro Luebs, Jan Skoglund, Florian Stimberg, Quan Wang and Thomas C Walters, "Wavenet based low rate speech coding," ICASSP 2018.
[3] Steven Van Kuyk, W Bastiaan Kleijn, and Richard C Hendriks, "On the information rate of speech communication," ICASSP 2017.