A Real-Time Wideband Neural Vocoder at 1.6 kb/s Using LPCNet

Jean-Marc Valin*, Amazon
Jan Skoglund, Google LLC

*work performed at Mozilla

Neural Speech Coding

Using a deep neural network to synthesize speech from conditioning features quantized at low bitrate

Demonstrated in the past using WaveNet

- Autoregressive model
- Dilated convolutions
- Outputs distribution of next μ-law value
- High output quality
- Very high complexity (> 100 GFLOPS)

Goal: Real-time wideband neural coding

WaveRNN

Replaces WaveNet dilated convolutions with a GRU

- Lower complexity (~10 GFLOPS)
- Two-stage, 16-bit output (8+8)

LPCNet Overview

Preemphasis

- Reduces HF noise caused by μ-law
- Avoids having to use 16-bit output

Sparse matrices

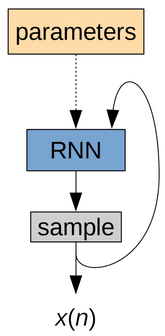
- Better quality for equivalent size

Input embedding

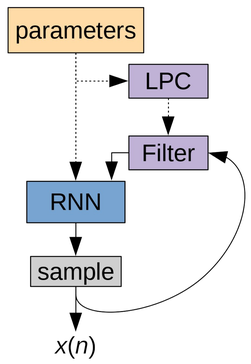
Linear prediction input to RNN

- Let LPC model vocal tract
- Neurons used to model excitation

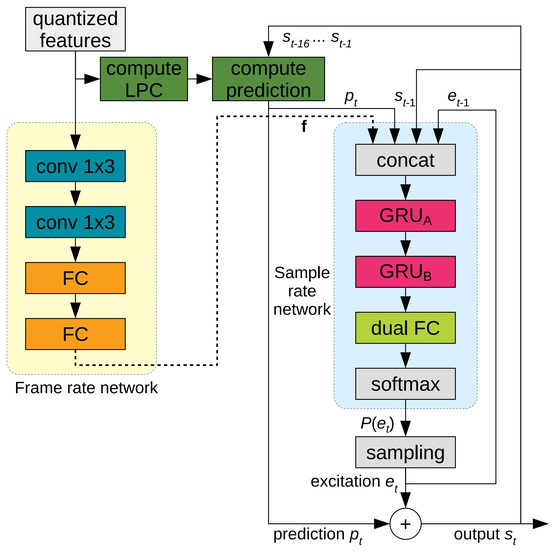
WaveRNN



LPCNet



Block Diagram



Features

Conditioning features: 10 ms

- Cepstrum
- Pitch period
- Pitch correlation

Packets: 40 ms

- Packing 4 frames

Pitch

Detection:

- Cross-correlation on LPC residual
- 5 ms sub-frames
- Range: 62.5 Hz to 500 Hz

Dynamic programming search

- Improves robustness
- Prevents large changes within packet

Quantization:

- Log-scale pitch over packet (6 bits)
- Linear pitch modulation (3 bits)
- Pitch correlation (2 bits)

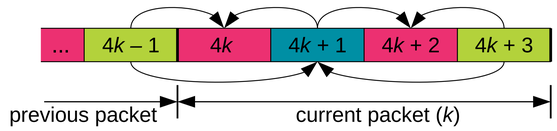
Cepstrum

Cepstral coefficients on 18 Bark bands

- 20-ms windows (50% overlap)

Quantization using B-frame-like structure

- Prediction can be 1) past sub-frame, 2) future sub-frame, or 3) average



Independent 3-stage VQ (30 bits)

Prediction + VQ (13 bits)

Prediction, no VQ (3 bits for both vectors)

Bit Allocation

Parameter	Bits
Pitch period	6
Pitch modulation	3
Pitch correlation	2
Energy (C0)	7
Cepstrum VQ (40 ms)	30
Cepstrum delta (20 ms)	13
Cepstrum interpolation (10 ms)	3
Total	**64**

Data

Train on NTT Multilingual Speech Database for Telephonometry (21 languages, 4 hours)

Use data augmentation (to 14 hours) to improve robustness by varying

- Frequency response
- Signal gain

Training

Add noise to input data to reduce effects of teacher forcing

Two-step training:

- Net trained with unquantized features
- Frame rate net adapted with quantized features (sample rate network is frozen)

Software

Open-source (BSD) C implementation at https://github.com/mozilla/LPCNet/

Complexity

Sample rate network: 72k weights

Decoder complexity: 3 GFLOPS

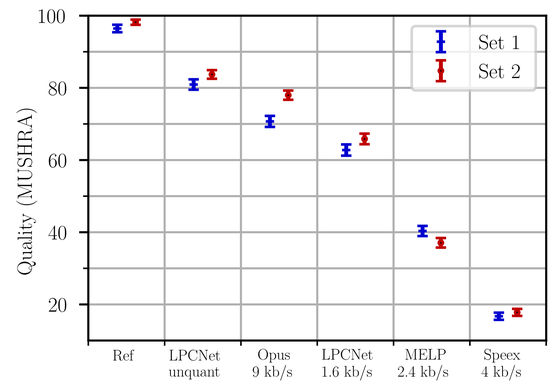
Real-time operation (one core) on phone

CPU	Clock	% Core
*AMD 2990WX (Threadripper)	3.0 GHz	14%
*Xeon E5-2640 v4 (Broadwell)	2.4 GHz	20%
Snapdragon 855 (**Galaxy S10**)	2.8 GHz	31%
Snapdragon 845 (**Pixel 3**)	2.5 GHz	68%
Cortex-A72 (Raspberry Pi 4)	1.5 GHz	110%

*turbo enabled

Results

Interactive demo and samples at https://people.xiph.org/~jm/demo/lpcnet_codec/



Contribution

Demonstrating usable neural vocoder technology

Significant improvement over existing vocoders

Future Work

Improve robustness to input noise

Synthesize from existing waveform codec bitstream (e.g. Opus)