A Hybrid DSP/Deep Learning Approach to Real-Time Fullband Speech Enhancement

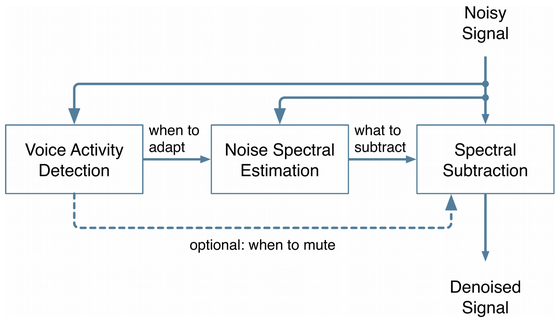
Jean-Marc Valin
jmvalin@jmvalin.ca

Conventional Noise Suppression

Building blocks (estimators):

- Voice Activity Detection (VAD)
- Noise estimation
- Spectral estimation

Estimators are hard to tune



Deep Learning and RNNs

Learn estimators to avoid manual tuning

- Recurrent Neural Networks (RNNs) can model temporal behaviour

Common drawbacks of deep learning:

- High complexity
- Large memory footprint (weights)

Hybrid Approach

System overview:

- 48 kHz input speech (0-20 kHz)
- 10 ms frame size (20-ms window)
- Low latency (10 ms look-ahead)

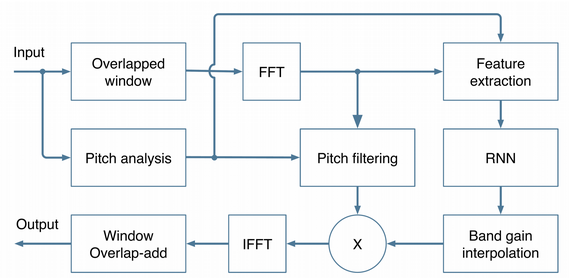
DSP for straightforward parts:

- Overlapping windows (FFT)
- Bark-like band structure (22 bands)
- Pitch filtering for harmonic structure

Deep learning to replace tricky estimators:

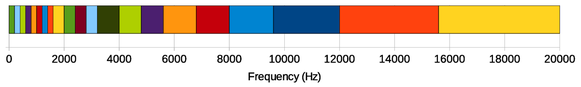
- All three estimators in the same network
- Estimating gains rather than spectrum
- Using gated recurrent units (GRU)
- Small network, low complexity

Block Diagram



Bands

Use the same Bark scale (critical bands) approximation as the Opus codec



Reduces complexity compared to per-bin gain

Gains

Each band has a gain $0 < g < 1$

- Ideal ratio between clean and noisy magnitudes
- Sigmoid activation guarantees range

Pitch filtering

Per-band comb filter

- Attenuates noise between harmonics
- Avoids the need for per-bin gains
- Computed in frequency-domain

Adaptive attenuation based on periodicity and amount of noise

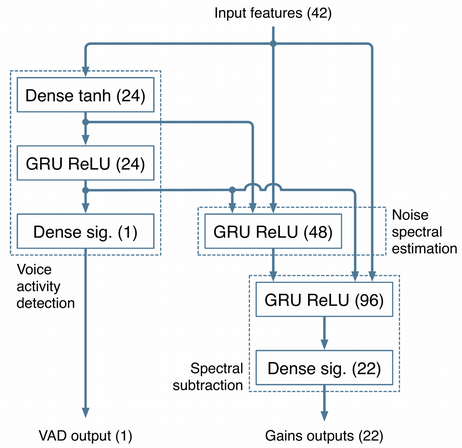
Features

Total of 42 features per frame:

- 22 cepstral coefficients
- 6 delta coefficients
- 6 delta-delta coefficients
- 6 pitch gain DCT coefficients
- 1 pitch period
- 1 non-stationarity metric

Architecture

Requires only 215 units over 5 layers



Data

Synthetic (speech+noise) noisy speech is needed for ground truth

Using data augmentation by varying

- SNR
- Frequency response (signal and noise)
- Signal gain
- Bandwidth (low-pass)
- ±20% resampling

Combining 6 hours of clean speech, 4 hours of noise into 140 hours of noisy speech

Training

Perceptual loss function

- MSE over sqrt(gain)
- Related to loudness

Complexity

Neural network:

- 87,503 weights (fits in L2 cache)
- 17.5 MFLOPS

Total complexity:

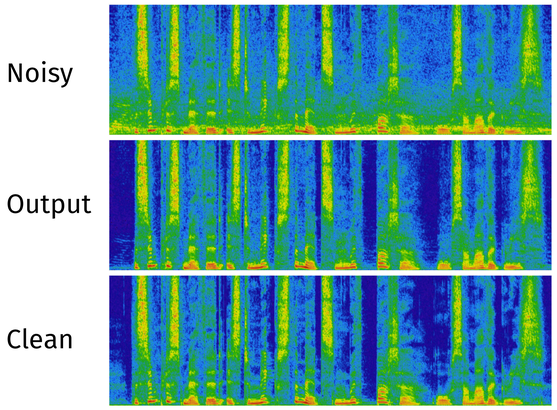
- ~40 MFLOPS
- 14% CPU on 1.2 GHz Raspberry Pi 3 (unoptimized C code)

Software

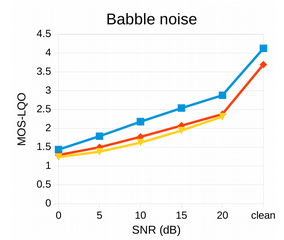
Open-source (BSD) C implementation at https://github.com/xiph/rnnoise

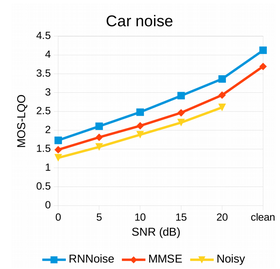
Results

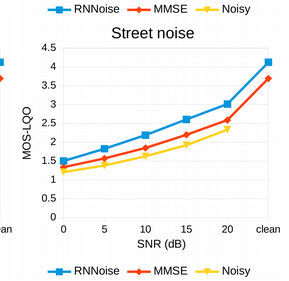
Interactive demo, samples, noise data https://people.xiph.org/~jm/demo/rnnoise/



Objective evaluation (PESQ-WB)







Contribution

Hybrid system combining

- Low complexity of conventional systems
- Quality improvements from deep learning

Other Applications

- Residual echo cancellation
- Microphone array post-filtering