WAVENET BASED LOW RATE SPEECH CODING

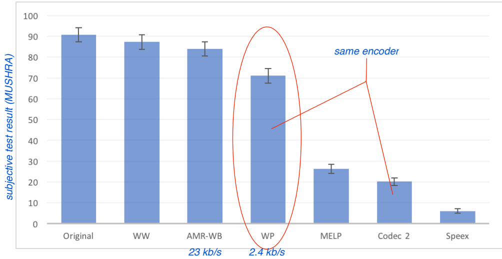
W. Bastiaan Kleijn [Vict,Google] Felicia S. C. Lim [Google] Alejandro Luebs [Google] Jan Skoglund [Google]
Florian Stimberg [DeepMind] Quan Wang [Google] Thomas C. Walters [DeepMind]

Problem

- Objective: low-rate coding of speech:
  - Rate: 2.4 kb/s.
  - Quality: as coders at 10 times the rate.
  - Wide-band (16 kHz sampling rate).
  - Good speaker identifiability.
- Based on generative modeling.
- Information-theoretical analysis.

Contributions

- Order of magnitude improvement rate-quality trade-off.



- Rate analysis: rate for model versus waveform.
  - Waveform coders cannot be improved further.
- Current objective quality estimators are inadequate.

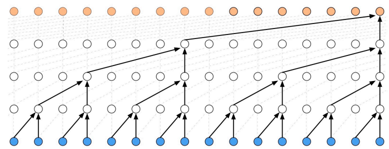
Background

- Speech coding applications:
  - Secure communications.
  - Mobile and internet communications.
- Now effectively subject to minimum quality threshold:
  - Because rate relatively cheap.
  - Quality threshold enforces *waveform coding*:
    - *Parametric coding* is inadequate.
    - Generative models are inadequate.
- However, significant rate reduction still attractive:
  - Particularly for scenarios with poor infrastructure.
- Relevant: true information rate 100 b/s [1];
  - Other attributes are negligible (mood, speaker).

- Generative models of speech:
  - Traditional:
    - Autoregressive.
    - Hidden Markov.
    - Kernel density estimation Hidden Markov.
  - New and good: deep neural network based:
    - WaveNet [2].
    - Only tried with known talkers.

WaveNet Decoder

- WaveNet conditioned on decoded bit stream.
- Mostly standard WaveNet configuration:
  - Multi-layer structure with dilated convolution.
  - Output: conditional dist for 8-bit ITU-T G.711 μ-law.
  - Signal samples drawn from conditional distribution.
  - Conditioning variables updated at 100 Hz.
  - Cross entropy loss function.
- Not standard: no talker identity provided.



Encoder

- Parametric coders: transmit only conditioning variables:
  - Condition the generative model: $p(s|\theta)$.
- Choices for WaveNet conditioning variables:
  - A trained network based encoding.
  - Use parameters of existing low-rate coder.
- Advantages conventional encoder:
  - Low computational complexity for encoder.
  - Illustrative of underlying principle.
- Codec 2

Variable	*Bits per update*	*Update Rate (Hz)*
spectrum	35	50
pitch	7	50
voicing	2	50
energy	5	50

Rate Analysis

- What is the rate benefit of *generating* the waveform?
- $\{S_i\}_{i \in \mathscr{A}}$: generated sequence.
- $\{\Theta_i\}_{i \in \mathscr{A}}$: conditioning sequence.
- Overall rate of generated signal over segment $\mathscr{A}$ is

$$\frac{1}{|\mathscr{A}|} H(\{S_i\}, \{\Theta_i\}) = \frac{1}{|\mathscr{A}|} H(\{S_i\}|\{\Theta_i\}) + \frac{1}{|\mathscr{A}|} H(\{\Theta_i\}).$$

- Rate $\frac{1}{|\mathscr{A}|} H(\{\Theta_i\})$ upper bounded by encoded rate.
- Assuming ergodicity, the generated signal rate is

$$\lim_{|\mathscr{A}| \to \infty} \frac{1}{|\mathscr{A}|} H(\{S_i\}|\{\Theta_i\}) = H(S_i | S_{i-1}, S_{i-2}, \cdots; \Theta_i)$$

$$\approx \frac{1}{|\mathscr{A}_0|} \sum_{i \in \mathscr{A}_0} H(S_i | s_{i-1}, s_{i-2}, \cdots; \theta_i),$$

- Mean information rate generated for a sample i is:

$$H(S_i | s_{i-1}, s_{i-2}, \cdots; \theta_i) = -\sum_{n \in \mathscr{N}} q_n^{(i)} \log_2 q_n^{(i)}.$$

WaveNet Waveform Coding

- Waveform coding is robust, hence commonly used:
  - Cost of poor model: Kullback-Leibler divergence.
- WaveNet waveform coder has two goals:
  - WaveNet can detect when model is poor; switch to waveform coding.
  - Analysis of existing predictive coding systems.
- WaveNet waveform coding:
  - Lossless coding of the μ-law quantized signal:
    - Quantization encoder $Q: \mathbb{R} \to \mathscr{N}$.
    - Quantization decoder $Z: \mathscr{N} \to \mathbb{R}$.
    - $\hat{x}_i = Z(n_i) = Z(Q(x_i))$
  - Predictive distr. known at encoder and decoder:
    - Have copy of WaveNet decoder at encoder.
    - Past signal is past reconstructed signal.
    - WaveNet $\to q_n^{(i)}$.
  - Use known predictive distribution for entropy coder.
- Estimate of the rate of waveform entropy coder is

$$\bar{H} = -\frac{1}{|\mathscr{A}_0|} \sum_{i \in \mathscr{A}_0} \sum_{n \in \mathscr{N}} q_n^{(i)} \log_2 q_n^{(i)}. \qquad (1)$$

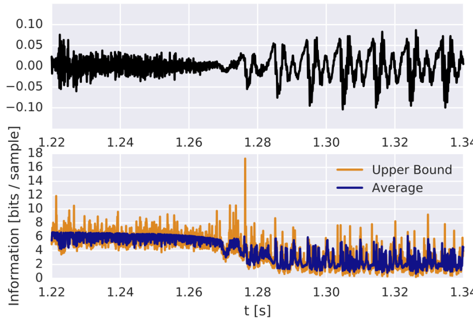
- Lower bound on real-world rate is

$$R = -\frac{1}{|\mathscr{A}_0|} \sum_{i \in \mathscr{A}_0} \log_2 q_{n_i}^{(i)}. \qquad (2)$$

- We expect R and $\bar{H}$ to be close.
- Required rate for conditioning variables:
  - Optimal rate for the conditioning variables independent on the mean signal distortion [3].
  - Vary only quantizer step size to vary rate.
- Can add perceptual weighting (pre- and post-filtering).

Measured Rates

- Mean encoding rate for waveform coding is high.



- Model of WaveNet waveform coding accurate:
  - Current waveform coders very good!
  - Waveform coders exploit perception.

Experimental Setup

- Encoder Codec 2 at 8 kHz and 2.4 kb/s.
- Decoder speech 16 kHz.
- Data bases:
  - Training set 32580 utterances, 123 speakers.
  - Testing set 2907 utterances, 8 speakers.

Quality Results

- Conventional objective quality estimators malfunction:
  - POLQA mean opinion scores (MOS)

	Codec 2	*MELP*	*Speex*	*AMR-WB*	*WW*	*WP*
Rate	2.4	2.4	2.4	23	42	2.4
MOS	2.7	2.9	2.2	4.6	4.7	~~2.9~~

- Subjective MUSHRA-type listening test:
  - 21 participants and 8 utterances.
  - Results in figure in column 1.
  - Two distinctive groups emerged:
    - Low quality: speex, Codec 2 and MELP.
    - High quality: AMR-WB, waveform WaveNet = μ-law, parametric WaveNet.

Speaker Identification Results

- Neural network based speaker identification model [4]:
  - Verification equal error rate (EER) results:
    - 8.4% for μ-law coded speech.
    - 15.8% for parametric WaveNet coded speech.
- Listening test:
  - Triangle test with 15 listeners, 16 trials.
  - Distinguish between two models:
    - Standard with test talkers not included.
    - Special-case with test talkers included.
  - Subjects correctly identify distinction at 41% rate (indistinguishable is 33%).

Conclusions

- High quality multi-talker generative models now exist.
- Coder efficiency improvement by order of magnitude.
- Implicit bandwidth extension is easy.
- Speaker identifiability slightly reduced:
  - Likely can be improved (bit stream, training).
- Waveform coders have reached their performance limit.
- Current objective quality estimators very poor;
  - Nonintrusive likely better.

References

[1] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "On the information rate of speech communication," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 5625–5629.
[2] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *ArXiv e-prints*, Sep. 2016.
[3] W. B. Kleijn and A. Ozerov, "Rate distribution between model and signal," in *Proc. IEEE Workshop on Applic. Signal Process. Audio and Acoust. (WASPAA)*, Nov. 2007, pp. 243–246.
[4] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," *arXiv preprint arXiv:1710.10467*, 2017.