HIGH-QUALITY SPEECH CODING WITH SAMPLE RNN

Janusz Klejša, Per Hedelin, Cong Zhou, Roy Fejgin, Lars Villemoes

1Dolby Sweden AB, Stockholm, Sweden, 2Dolby Laboratories, San Francisco, CA, USA

Introduction

Background
- Deep generative schemes achieve realistic-sounding speech
- But can they provide transparent quality in speech coding applications?

Objective
- High-quality wideband speech with bitrate competitive to state-of-the-art codecs

Highlights
- Coding scheme based on a conditioned SampleRNN model
- Rigorous MUSHRA-like testing, showing that the quality of state-of-the-art codecs can be achieved at less than half the bit-rate
- Robustness study

System Overview

Vocoder

Operating points of the encoder (10 ms frames)

<table>
<thead>
<tr>
<th>Rate (kb/s)</th>
<th>LPC filter order</th>
<th>Spectral distortion (dB)</th>
<th>RMS level of LPC residual [bits]</th>
<th>Voicing level [bits]</th>
<th>Pitch, F0 [bits]</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.0</td>
<td>22</td>
<td>0.754</td>
<td>1+9</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>6.4</td>
<td>16</td>
<td>0.782</td>
<td>1+8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>5.6</td>
<td>16</td>
<td>1.33</td>
<td>1+8</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

Categorization Techniques Employed
- Predictive and entropy coding
- LPC model coded as line spectral pairs (LSP)
- Vector quantization (VQ)
- Quantization cell probabilities are provided by a Gaussian mixture model (GMM) trained on the WSJ dataset

SampleRNN with Conditioning

Loss function: negative log-likelihood (NLL)
Training: truncated back propagation through time with ADAM optimizer
16-bit outputs: using logistic mixtures
Sampling: directly from the conditional distribution

Experiments

Dataset
- Wall St. Journal dataset (WSJ0 / CSR3)
- 35478 utterances; 16 kHz, multiple speakers
- Test speakers were excluded from the training set

Experiment 1: 8 kb/s

Experiment 2: 6.4 kb/s

Experiment 3: Quality-Bitrate Tradeoff

Note: (emb) i.e. embedding, refers to systems trained for 8 kb/s but tested with lower bitrate conditioning. That is, we test a new bitrate without retraining.

Experiment 4: Objective evaluation with POLQA

Out-of-distribution Performance

- Diverse training dataset provides robust performance

Experiment 5: Out-of-distribution Performance

Experiment 6: Multi-frame Conditioning 6.4 kb/s (post-submission)

- Frames t-2 through t+2 were provided
- In a follow-up experiment, we found that the same quality can be achieved using only frame t+2

Conclusions

- 2.5x performance gain compared to the state-of-the-art codecs
- Speaker and language independence
- Out-of-distribution degradation mitigated by dataset expansion

Listen to the samples on IEEE SigPort

https://sigport.org/documents/high-quality-speech-coding-sample-rnn