HIGH–QUALITY SPEECH CODING WITH SAMPLE RNN

Janusz Klejsa[1], Per Hedelin[1], Cong Zhou[2], Roy Fejgin[2], Lars Villemoes[1]

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

Introduction

Background

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

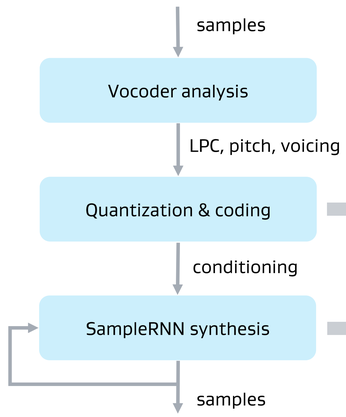
Objective

- 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



Application

- Speaker and language-independent coding of dry speech

Vocoder

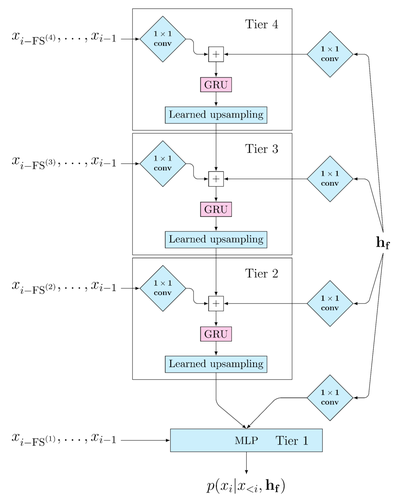
Operating points of the encoder (10 ms frames)

Rate [kb/s]	LPC filter order	Spectral distortion [dB]	RMS level of LPC residual [bits]	Voicing level vector [bits]	Pitch, *f0* [bits]
8.0	22	0.754	1+9	9	10
6.4	16	0.782	1+8	9	10
5.6	16	1.33	1+8	9	10

Coding 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 WSJ0 training set

SampleRNN with Conditioning



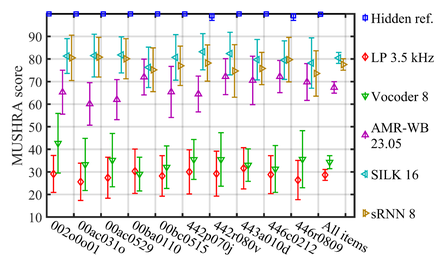
Loss function: negative log likelihood (NLL)
Training: truncated back propagation through time with ADAM optimizer
16-bit outputs: using logistic mixtures
Sampling: directly from the conditional distribution

Experiments

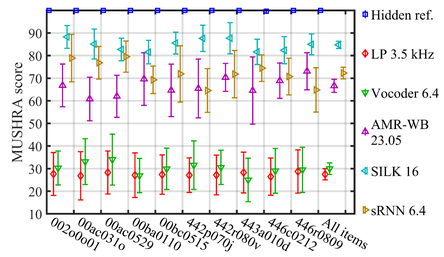
Dataset

- Wall St. Journal dataset (WSJ0 / CSR-1)
- 35,478 utterances; 16 kHz; multiple speakers
- Test speakers were excluded from the training set

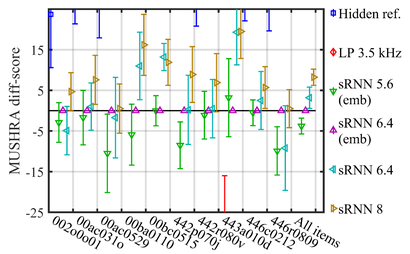
Experiment 1: 8 kb/s



Experiment 2: 6.4 kb/s



Experiment 3: Quality–Bitrate Tradeoff

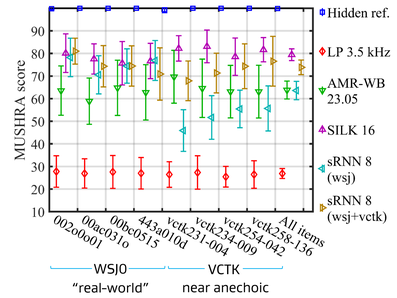


Note: (emb) i.e. embedding, refers to systems trained for 8 kb/s but tested with lower-bitrate conditioning. That is, we test a new bitrate without retraining.

Experiment 4: Objective evaluation with POLQA

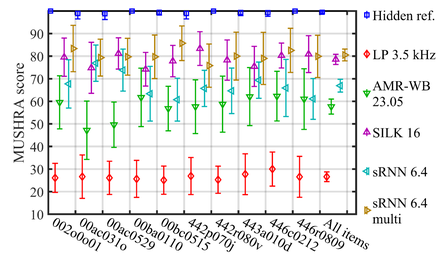
	Vocoder		AMR-WB	SILK	SampleRNN	
Rate [kb/s]	6.4	8.0	23.05	16.0	6.4	8.0
MOS-LQO	3.43	3.67	4.39	4.41	3.27	3.48

Experiment 5: Out-of-distribution Performance



- Diverse training dataset provides robust 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