Speech Compression using Deep Learning

EE 679 Course Project
IIT Bombay

- Mithilesh Vaidya
17D070011

Introduction

- Why? Efficient for storage and transmission!
- 2 classes:

  1. Waveform-based: Generic, may not exploit speech information
     Goal is to minimise MSE
     E.g. PCM, DPCM, delta modulation
  2. Parametric - assume an underlying model
     E.g. CELP, simple LP filtering, ANNs

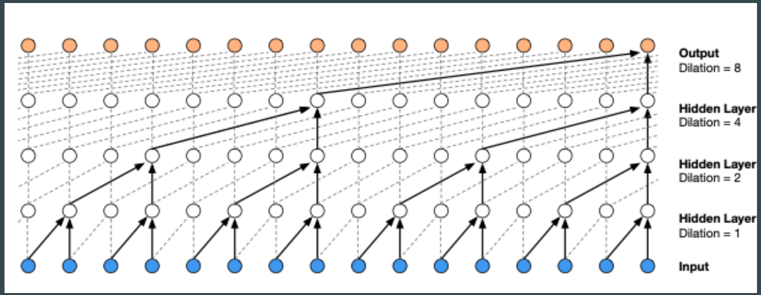
Focus on: Speaker-dependent WaveNet Vocoder [1]

WaveNet [2]

- Breakthrough DNN model which can generate speech sample-by-sample
- Challenging due to large sampling rate
- Stack of convolution layers model the probability:

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$

- Observe the **causal** nature of convolution filters
- Dilation increases temporal receptive field
- No pooling → Output size = Input size



WaveNet (continued)

- µ-law companding tx to get 256 output bins (instead of 16-bit i.e. 65,536)
- Activation function: $\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x}\right) \odot \sigma\left(W_{g,k} * \mathbf{x}\right)$

To summarise:

- Input: Raw waveform of T samples
- WaveNet: Gives a Tx256 vector of probability distributions
- Loss: Simple Log Likelihood

Train using standard ML techniques

Conditional WaveNet

- To condition on any variable of interest: $$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h}).$$

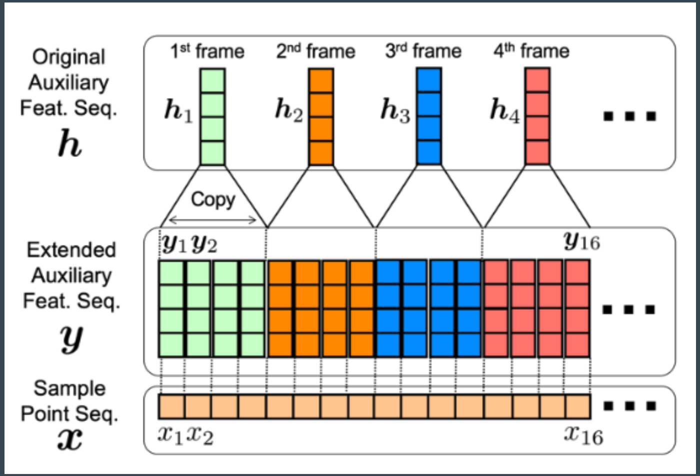
- Augment the activation function: $$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k}^T \mathbf{h}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k}^T \mathbf{h}\right)$$

- Intuition: Guide WaveNet in producing desired characteristics
- Examples:
    - Desired Speaker: *h* is one-hot encoded
      Global conditioning since constant across utterance
    - TTS: Supply information about text to generate e.g. embeddings, F0
      Local since varying with time

Speech Coding in [1]

Encoder (per-frame basis):

- Extract mel-cepstrum (say 25) from either STFT or smoothened envelope using STRAIGHT analysis [3]
- Extract pitch F0 using RAPT [4]
- 25 ms window, 5 ms hop



Decoder:

- Feed above features to a trained conditional WaveNet model
- Generate speech sample-by-sample

Features

No need of separately modelling excitation signal since no encoding regarding:

- Voiced/Unvoiced
- Glottal shape

Although parameters sent on a frame-by-frame basis,

- WaveNet is more *powerful* than a linear time-invariant system i.e. the temporal structure can be fine-tuned
- No assumption about stationarity within a segment

Baseline

Features:

- Plain: Mel-Cepstrum from STFT
- STRAIGHT [3]: Smoothen envelope to reduce periodic redundancies, then extract coefficients

Synthesis:

- MLSA [5]: Pass coefficients through this filter to synthesize speech
- WaveNet: 4 separate models, 1 for each speaker

Comparative Method	**Source of mel-cepstrum**	**Waveform Synthesis**
Plain-MLSA	STFT	MLSA filter
STRAIGHT-MLSA	Spectrum envelop	MLSA filter
Plain-WaveNet	STFT	WaveNet
STRAIGHT-WaveNet	Spectrum envelop	WaveNet

Evaluation Measures

For each frame:

- x(n)/s: synthesized speech
- y(n)/r: original speech
- N: # samples in a frame
- Y(f) and X(f): Fourier Transforms
- F: Number of frequency bins

Averaged over all frames

RMSE: actual spectral distortion
MCD: envelope distortion

$$SNR = 10 \ln_{10} \left(\frac{\sum_{n=1}^{N} y(n)^2}{\sum_{n=1}^{N} (x(n) - y(n))^2} \right)$$

$$RMSE = \sqrt{\frac{1}{F} \sum_{f=1}^{F} \left(20 \log_{10} \frac{|Y(f)|}{|X(f)|} \right)^2}$$

$$MCD = \frac{10}{\log 10} \sqrt{2 \sum_{m=1}^{M} (c_r(m) - c_s(m))^2}$$

Mean Cepstral Distance

$$RMSE(f_o) = 1200 \sqrt{(\log_2(F_r) - \log_2(F_s))^2}$$

Results

(a) *SNR (dB); distortion in time domain*

Method	slt	bdl	clb	rms
MLSA (P)	-0.24 ± 0.31	-2.7 ± 0.19	-0.044 ± 0.35	-2.2 ± 0.52
MLSA (ST)	3.7 ± 0.32	-2.6 ± 0.16	-1.9 ± 0.31	-2.3 ± 0.45
WaveNet (P)	$\mathbf{4.1 \pm 0.23}$	$\mathbf{3.6 \pm 0.21}$	$\mathbf{3.8 \pm 0.38}$	$\mathbf{4.0 \pm 1.0}$
WaveNet (ST)	3.7 ± 0.32	2.2 ± 0.28	3.7 ± 0.32	2.6 ± 0.94

(b) *RMSE (dB); distortion in frequency domain*

Method	slt	bdl	clb	rms
MLSA (P)	$\mathbf{7.9 \pm 0.13}$	$\mathbf{7.9 \pm 0.21}$	$\mathbf{7.8 \pm 0.23}$	$\mathbf{8.1 \pm 0.97}$
MLSA (ST)	8.3 ± 0.31	8.6 ± 0.48	7.9 ± 0.43	8.4 ± 0.53
WaveNet (P)	8.8 ± 0.21	8.6 ± 0.21	9.2 ± 0.30	9.0 ± 1.3
WaveNet (ST)	9.0 ± 0.35	9.4 ± 0.30	9.1 ± 0.28	9.5 ± 1.3

Table 3: *Comparison of distortion between acoustic features of natural speech and synthesized speech*

(a) *Mel-cepstrum (MCD; dB)*

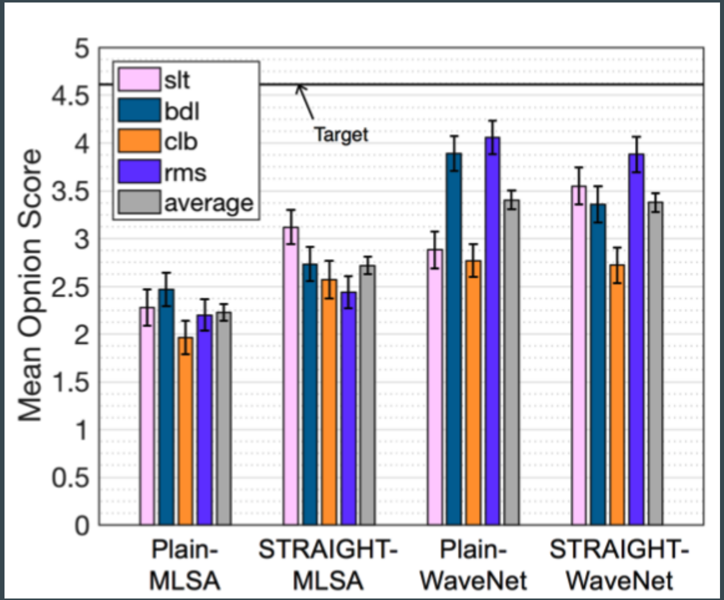
Method	slt	bdl	clb	rms
MLSA (P)	3.8 ± 0.027	3.8 ± 0.050	4.6 ± 0.050	3.6 ± 0.054
MLSA (ST)	$\mathbf{2.4 \pm 0.047}$	$\mathbf{2.3 \pm 0.054}$	$\mathbf{2.5 \pm 0.049}$	$\mathbf{2.5 \pm 0.059}$
WaveNet (P)	5.5 ± 0.052	5.5 ± 0.050	6.8 ± 0.11	4.9 ± 0.053
WaveNet (ST)	5.7 ± 0.045	5.7 ± 0.053	6.8 ± 0.045	5.1 ± 0.052

(b) *Fundamental frequency (RMSE; cent)*

Method	slt	bdl	clb	rms
MLSA (P)	2.9 ± 0.21	9.4 ± 1.6	2.4 ± 0.19	6.4 ± 0.63
MLSA (ST)	2.7 ± 0.18	8.7 ± 1.6	2.1 ± 0.13	6.2 ± 0.79
WaveNet (P)	$\mathbf{1.9 \pm 0.22}$	$\mathbf{7.5 \pm 1.6}$	$\mathbf{1.1 \pm 0.087}$	$\mathbf{3.7 \pm 1.4}$
WaveNet (ST)	2.3 ± 0.13	9.7 ± 2.0	1.1 ± 0.13	5.6 ± 1.5

- WaveNet has low SNR, F0 distortion
- MLSA has low RMSE, MCD

Results



- slt, clb: female ; bdl, rms: male
- Subjective evaluation tells a different story than objective evaluation
- Average MOS for WaveNet > MLSA
- As expected, performance for female speakers is worse

Conclusion

Limitations:

- Speaker-dependent: 1 hour of data per speaker
- Slow: 2 days to train WaveNet for one speaker
  6 minutes to synthesize a 3-second speech => nowhere close to real-time

Future Scope:

- Try out other features
- Make it speaker-independent
- Reduce time taken for synthesis

References

[1] Speaker-dependent WaveNet vocoder (2017)

[2] WaveNet: A generative model for raw audio (2016)

[3] STRAIGHT (1999)

[4] RAPT

[5] MLSA

Recent papers:

[6] LOW BIT-RATE SPEECH CODING WITH VQ-VAE AND A WAVENET DECODER (2019)