TOWARD WAVENET SPEECH SYNTHESIS

2018. 12. 11
DSP & AI Lab
Min-Jae Hwang

PROFILE

Min-Jae Hwang

- Affiliation : 7th semester of MS/Ph.D. student
  DSP & AI Lab., Yonsei Univ., Seoul, Korea
- Adviser : Prof. Hong-Goo Kang
- E-mail : hmj234@dsp.yonsei.ac.kr

Work experience

- 2017.12 ~ 2017.12 – Intern at NAVER corporation
- 2018.01 ~ 2018.11 – Intern at Microsoft Research Asia

Research area

- 2015.8 ~ 2016.12 - Audio watermarking
  - **M. Hwang**, J. Lee, M. Lee, and H. Kang,“사전 분석법을 통한 스프레드 스펙트럼 기반 오디오 워터마킹 알고리즘의 성능 향상,” 한국음향학회 제 33회 음성통신 및 신호처리 학술대회, 2016
  - **M. Hwang**, J. Lee, M. Lee, and H. Kang, “SVD-based adaptive QIM watermarking on stereo audio signals,” *in IEEE Transactions on Multimedia*, 2017
- 2017.1 ~ Present - Deep learning-based statistical parametric speech synthesis
  - **M. Hwang**, E. Song, K. Byun, and H. Kang, “Modeling-by-generation structure-based noise compensation algorithm for glottal vocoding speech synthesis system,” *in ICASSP*, 2018
  - **M. Hwang**, E. Song, J. Kim, and H. Kang, “A unified framework for the generation of glottal signals in deep learning-based parametric speech synthesis systems,” *in Interspeech,* 2018
  - **M. Hwang**, F. Soong, F. Xie, X. Wang, and H. Kang, “LP-WaveNet: linear prediction-based WaveNet speech synthesis,” *in ICASSP*, 2019 [Submitted]

CONTENTS

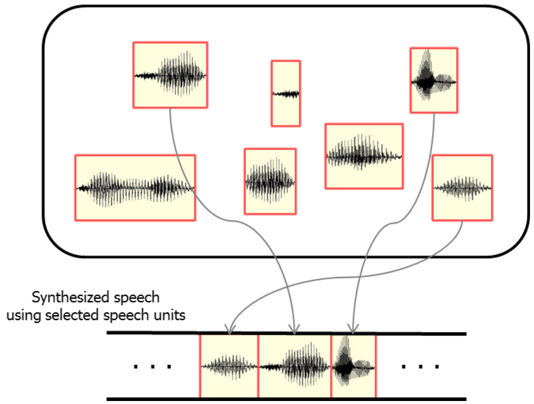
Traditional Text-to-speech	WaveNet	WaveNet Vocoder	MDN-WaveNet	ExcitNet	LP-WaveNet
~2015	*2016*	*2017*	*2017*	*2018*	*2019*

INTRODUCTION

Story of WaveNet vocoder

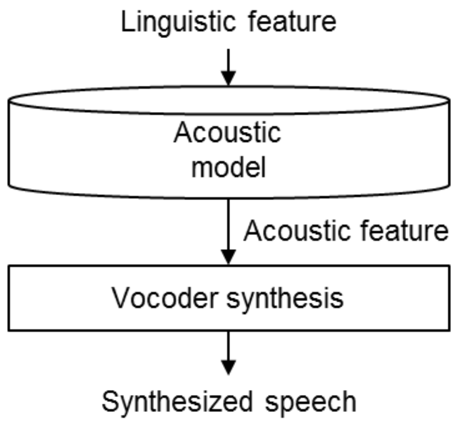
CLASSICAL TTS SYSTEMS

Unit-selection speech synthesis [1]



- Concatenate tiny segment of real speech signal
- Pros) High quality
- Cons) Low controllability, large database

Statistical parametric speech synthesis (SPSS) [2]



- Parameterize speech signal using vocoder technique
- Pros) High controllability, small database
- Cons) Low quality

GENERATIVE MODEL-BASED SPEECH SYNTHESIS

WaveNet [3]

- First generative model for raw audio waveform

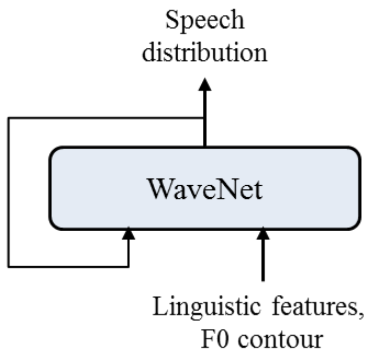
$$p(\mathbf{x}) = \prod_{n=1}^{N} p(x_n \mid \mathbf{x}_{<n})$$

  - Predict the probability distribution of waveform sample auto-regressively

- Generate high quality of audio / speech signal
  - Impact on the task of TTS, voice conversion, music synthesis, etc.

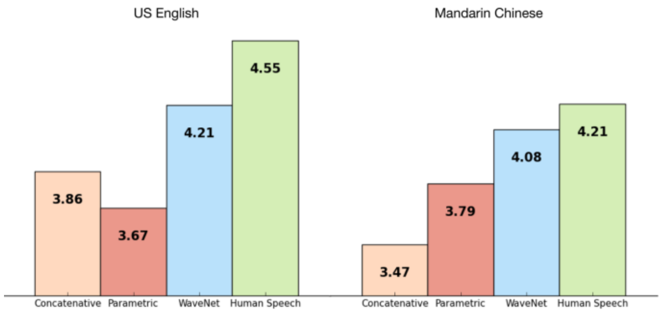
WaveNet in TTS task

- Utilize *linguistic features* and *F0 contour* as a conditional information



[WaveNet speech synthesis]

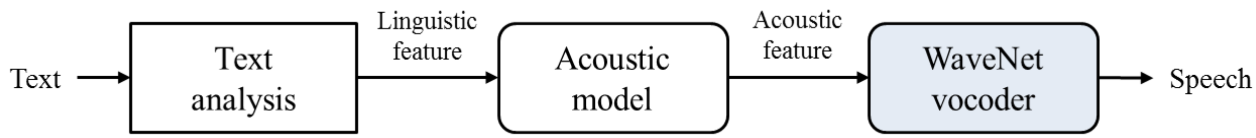
- Present higher quality than the conventional TTS systems



[Mean opinion scores]

WaveNet vocoder-based speech synthesis

Utilize WaveNet as parametric vocoder [4]



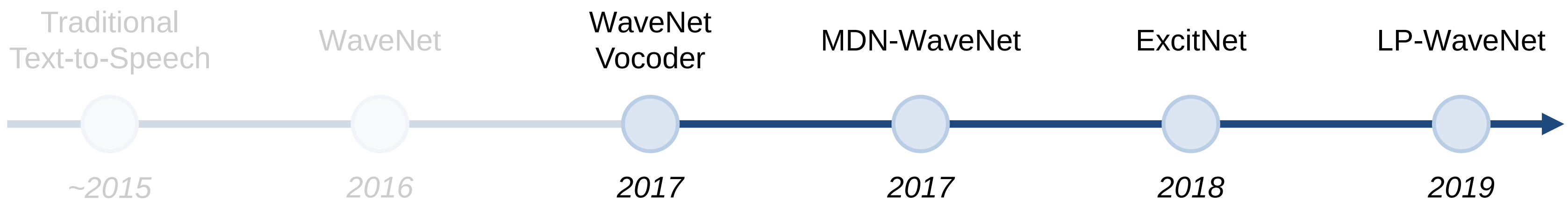
[WaveNet vocoder based parametric speech synthesis]

- Use acoustic features as conditional information

Advantages

- **Higher quality synthesized speech** than the conventional vocoders
  - Don't require hand-engineered processing pipeline
- **Higher controllability** than the case of linguistic features
  - Controlling acoustic features
- **Higher training efficiency** than the case of linguistic features
  - Linguistic feature: 25~35 hour database
  - Acoustic feature: 1 hour database

Recorded	Conventional vocoder	WaveNet vocoder



Various types of WaveNet vocoders

1. **SoftMax-WaveNet**
2. **MDN-WaveNet**
3. **ExcitNet**
4. **Proposed LP-WaveNet**

Basic of WaveNet

Auto-regressive generative model

- Predict probability distribution of speech samples

$$p(\mathbf{x}) = \prod_{n=1}^{N} p(x_n \mid \mathbf{x}_{n-R:n-1})$$

- Use past waveform samples as a condition information of WaveNet

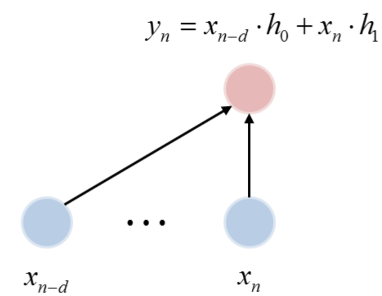
Problem: Long-term dependency nature of speech signal

- Highly correlated speech signal in high sampling rate, e.g. 16000 Hz
  - E.g. 1) At least 160 (=16000/100) speech samples to represent 100Hz of voice correctly
  - E.g. 2) Averagely 6,000 speech samples to represent single word[1]

➡ ***Effective embedding method of long receptive field is required***

Solution

- Put the speech samples as an input of dilated causal convolution layer
- Stack the dilated causal convolution layer
  - Result in exponentially growing receptive field

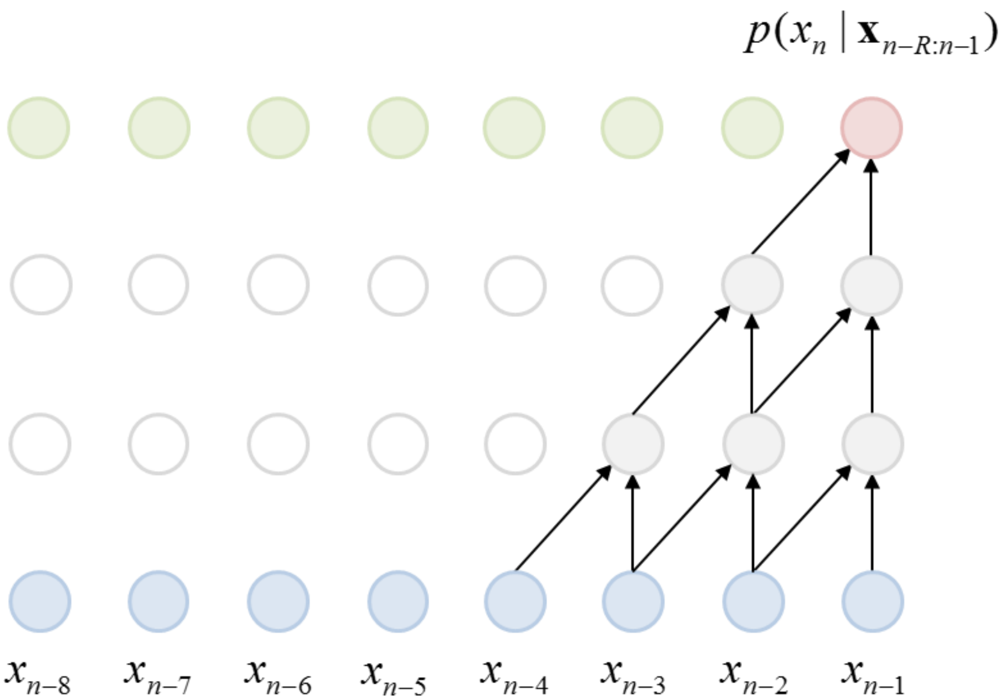


[Dilated causal convolution]

[1] Statistics based on the average speaking rate of a set of TED talk speakers
http://sixminutes.dlugan.com/speaking-rate/

Basic of WaveNet

WaveNet without dilation

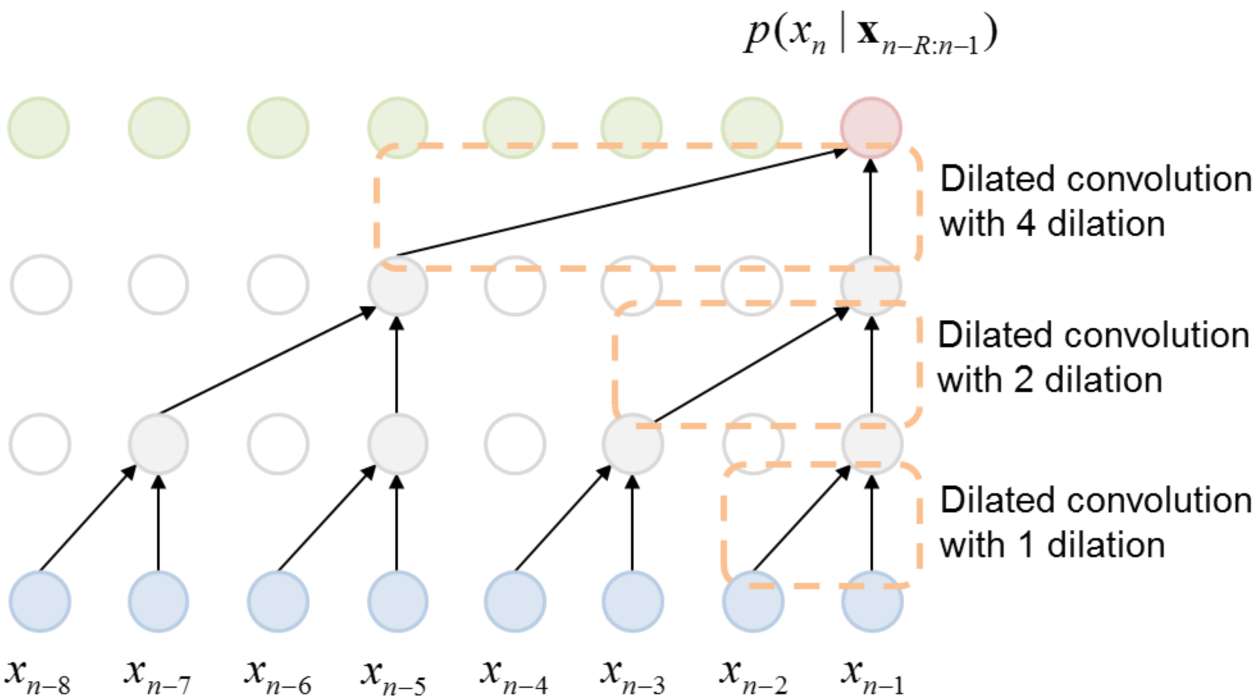


[WaveNet without dilated convolution]

Receptive field: # layer - 1

Basic of WaveNet

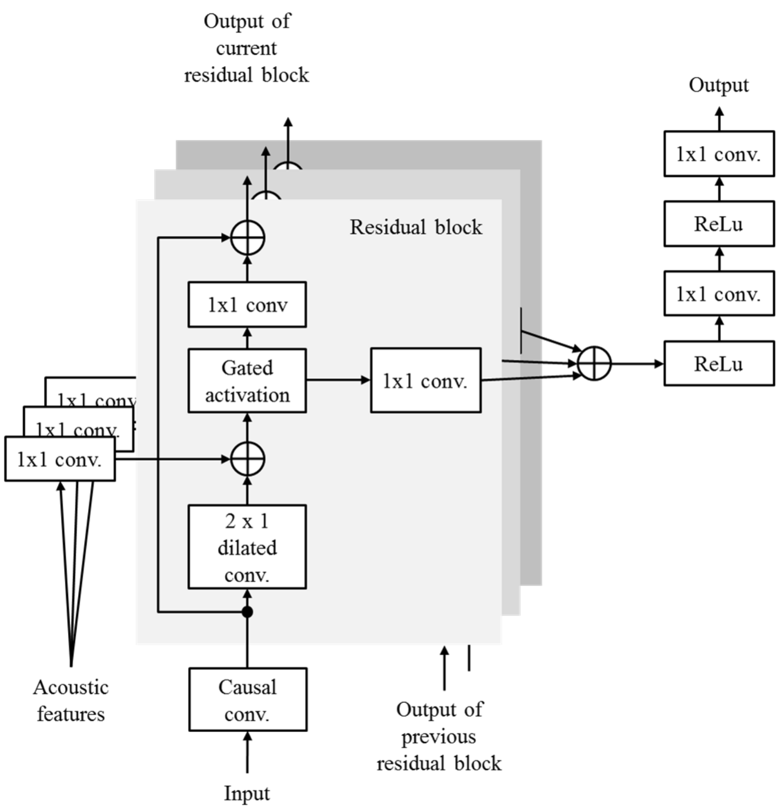
WaveNet with dilation



[WaveNet with dilated convolution]

Receptive field: $2^{\#\text{ layer}} - 1$

Basic of WaveNet



[Basic WaveNet architecture]

Multiple stack of residual blocks

1. Dilated causal convolution

$$y[d,n] = \sum_{k=0}^{M-1} h[k]x[n-d\cdot k]$$

• Exponentially increase the receptive field

2. Gated activation

$$\mathbf{z} = \tanh\left(W_f * \mathbf{x} + V_f * \mathbf{h}\right) \odot \mathrm{sigmoid}\left(W_g * \mathbf{x} + V_g * \mathbf{h}\right)$$

• Impose non-linearity on the model
• Enable conditional WaveNet

3. Residual / skip connections
• Speed up the convergence
• Enable deep layered model training

SoftMax-WaveNet

Use WaveNet as multinomial logistic regression model [4]

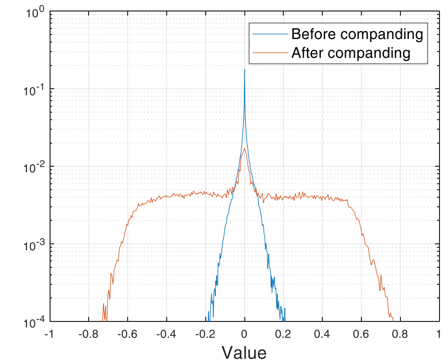
- μ-law companding for evenly distributed speech sample

$$y = sign(x) \cdot \frac{\ln\left(1+\mu|x|\right)}{\ln(1+\mu)}, \ \mu=255$$

- 8-bit one-hot encoding

$$p = OneHot_{8bit}(y)$$

  - 256 symbols



[Distribution of speech samples]

Estimate each symbol using WaveNet

$$\mathbf{z}^q = WaveNet(\mathbf{q}_{<n} \mid \mathbf{h})$$

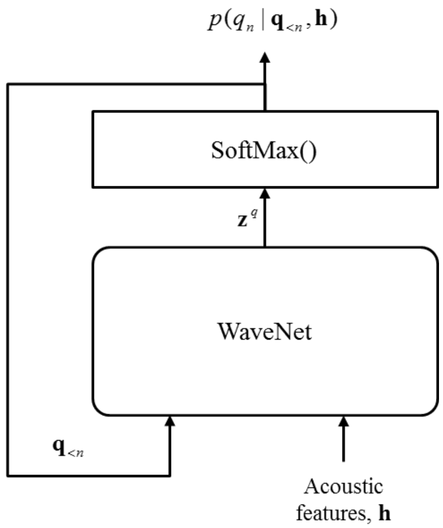
$$q_n = \frac{\exp(z_n^q)}{\sum_i \exp(z_i^q)}$$

- Predict the sample by SoftMax distribution

Optimize network by cross-entropy (CE) loss

$$L = \sum_n \left[-p_n \log q_n\right]$$

- Minimize the probabilistic distance between p_n and q_n

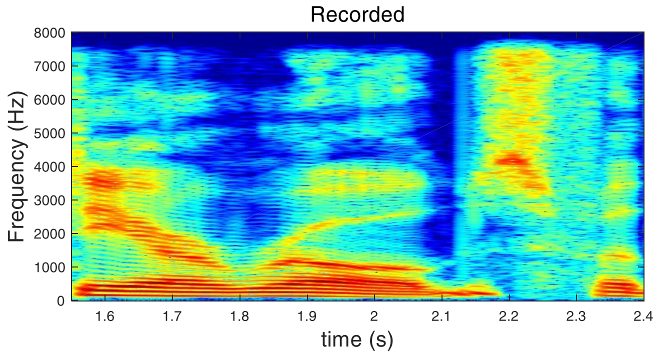


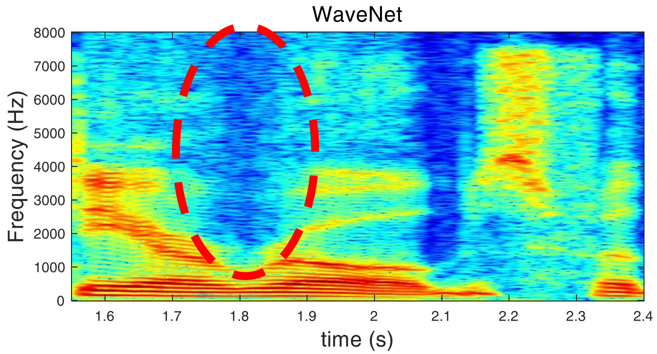
[SoftMax-WaveNet]

SoftMax-WaveNet

Limitation by the usage of 8-bit quantization

- Noisy synthetic speech due to insufficient number of quantization bits





Intuitive solution

- Expand the SoftMax dimension to 65,536 corresponds to 16-bit quantization
  → High computational cost & difficult to train

Mixture density network (MDN)-based solution [5]

- Train the WaveNet to predict the parameter of pre-defined speech distribution

σ : logistic function

Mixture density network

MDN-WAVENET

Define the distribution of waveform sample as parameterized form [5]

- Discretized mixture of logistic (MoL) distribution

$$p(x)=\sum_{n=1}^{N}\pi_n\left[\sigma\left(\frac{x+\Delta/2-\mu_n}{s_n}\right)-\sigma\left(\frac{x-\Delta/2-\mu_n}{s_n}\right)\right]$$

  - Discretized logistic mixture with 16bit quantization ($\Delta=1/2^{16}$)

Estimate mixture parameters by WaveNet

$$[\mathbf{z}^{\pi},\ \mathbf{z}^{\mu},\ \mathbf{z}^{s}]=WaveNet(\mathbf{x}_{<n},\ \mathbf{h})$$

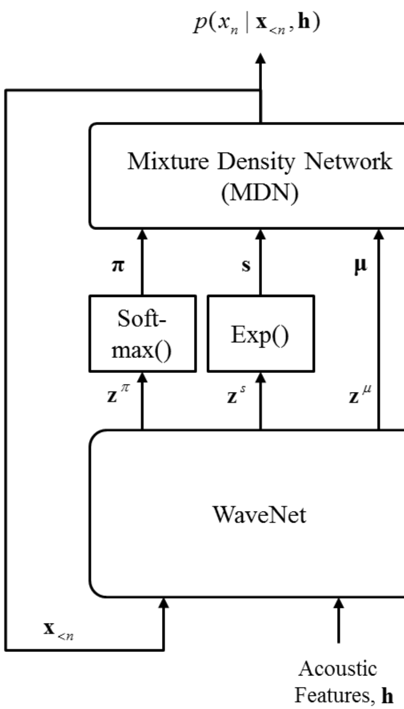
$\boldsymbol{\pi}=\mathrm{softmax}(\mathbf{z}^{\pi})$, for unity-summed mixture gain

$\boldsymbol{\mu}=\mathbf{z}^{\mu}$

$\mathbf{s}=\exp(\mathbf{z}^{s})$, for positive value of mixture scale

Optimize network by negative log-likelihood (NLL) loss

$$L=\sum_n[-\log p(x_n\mid x_{<n},\mathbf{h})]$$



[MDN-WaveNet]

Mixture density network

MDN-WAVENET

Higher quality than SoftMax-WaveNet

- Enable to model the speech signal by 16-bit
- Overcome difficulty of spectrum modeling

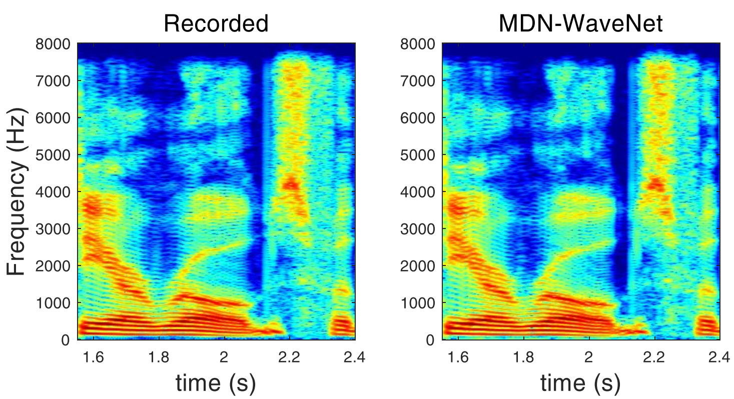
Recorded	SoftMax WaveNet	MDN WaveNet

Difficulty of WaveNet training

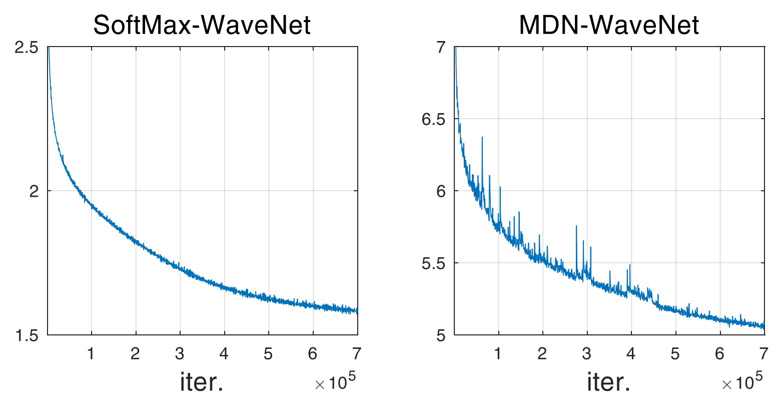
- Due to increased quantization bit from 8-bit to 16-bit

Solution based on the human speech production model [6]

- Model the vocal source signal, whose physical behavior is much simpler than the speech signal



[Spectrogram comparison]



[Loss comparison]

Speech production model

Concept

- Modeling the speech as the filtered output of vocal source signal to vocal tract filter

$$S(z) = [G(z) \cdot V(z) \cdot R(z)] \cdot E(z)$$

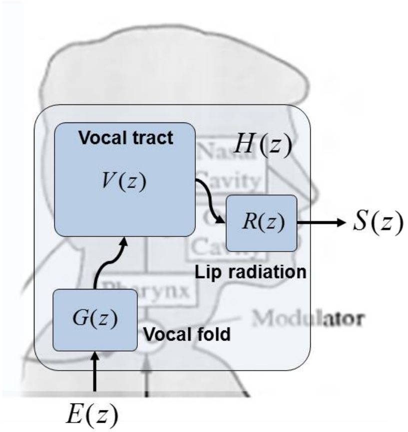
Speech = [vocal fold × vocal tract × lip radiation] × excitation

Methodology: Linear prediction (LP) approach

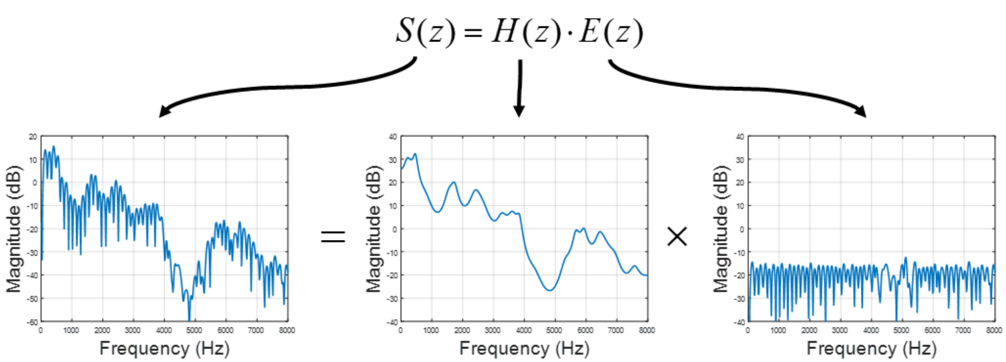
- Define speech signal as linear combination of past speech samples

$$s_n = \sum_{i=1}^{P} \alpha_i s_{n-i} + e_n \quad \Leftrightarrow \quad S(z) = H(z) \cdot E(z), \text{ where } H(z) = \frac{1}{1 - \sum_{i=1}^{p} \alpha_i z^{-i}}$$

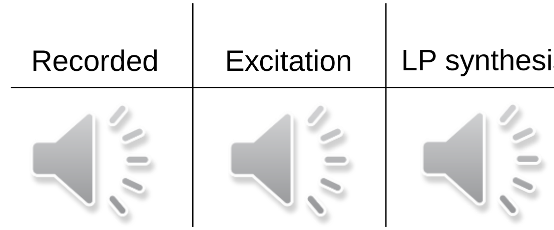
Spectrum part = LP coefficients
Excitation part = Residual signal of LP analysis



[Speech production model]



[Spectral deconvolution by LP analysis]



ExcitNet

Model the excitation signal by WaveNet, instead of speech signal [6]

Training stage

- Extract excitation signal by linear prediction (LP) analysis

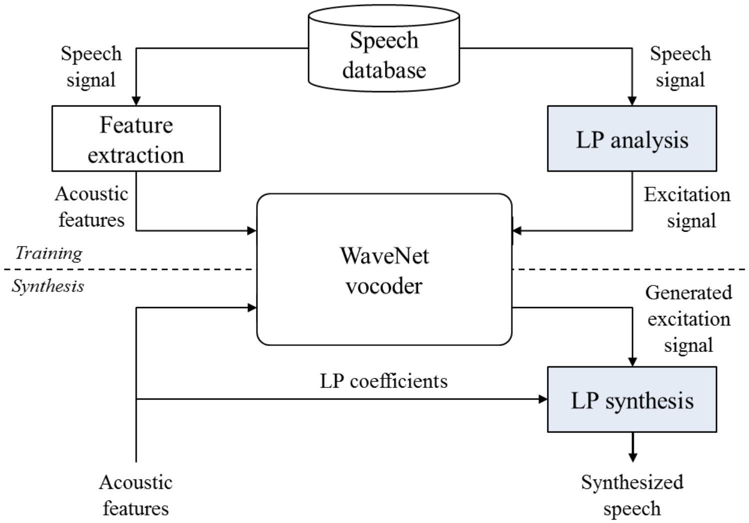
$$e_n = s_n - \sum_{i=1}^{P} \alpha_i s_{n-i}$$

  - Periodically updated filter coefficients matched to frame rate of acoustic feature
- Train WaveNet to model excitation signal

Synthesis stage

- Generated excitation signal by WaveNet
- Synthesize speech signal by LP synthesis filtering

$$s_n = \sum_{i=1}^{p} \alpha_i s_{n-i} + e_n$$



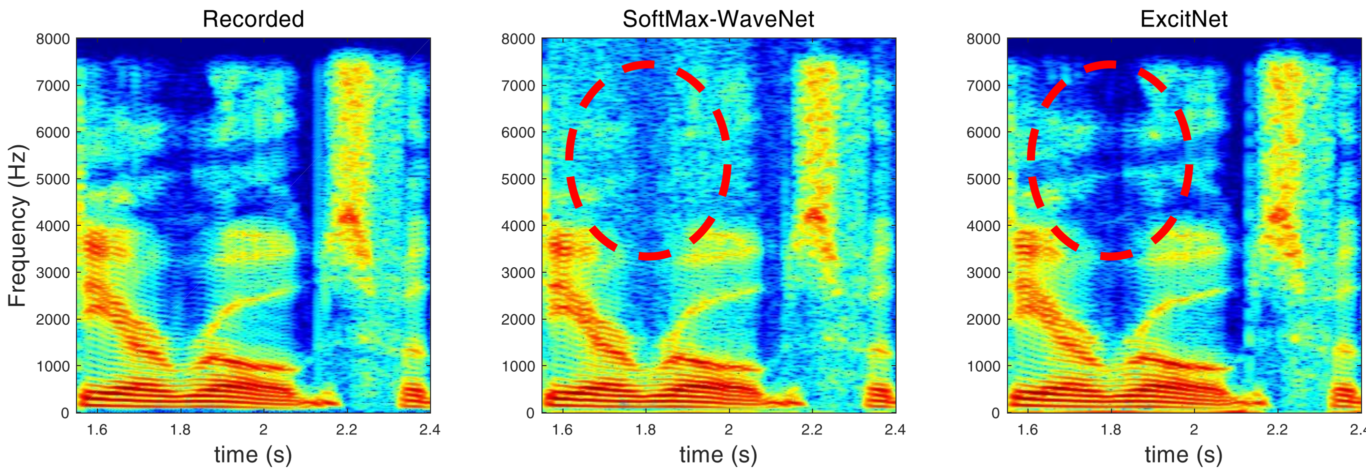
[ExcitNet]

Recorded	SoftMax WaveNet	ExcitNet

ExcitNet

High quality of synthesized speech even though 8-bit quantization is used

- Overcome the difficulty of spectrum modeling by using external spectral shaping filter



Limitation

- Independent modeling of excitation and spectrum parts results in unpredictable prediction error

$$S(z) = \frac{1}{\underbrace{1-\sum_{i=1}^{p}\alpha_i z^{-i}}_{\text{Contains error}}} \cdot \underbrace{E(z)}_{\text{Contains error}}$$

Objective

Represent LP synthesis process during WaveNet training / generation

LP-WaveNet

Motivation from the assumption of WaveNet vocoder

1. Previous speech samples, $\mathbf{x}_{<n}$, are given
2. LP coefficients, $\{\alpha_i\}$, are given

Their linear combination, $\hat{x}_n = \sum_{i=1}^{p} \alpha_i x_{n-i}$ ***, are also given***

Probabilistic analysis

$$x_n = e_n + \hat{x}_n$$

$$\begin{aligned} X_n \mid (\mathbf{x}_{<n}, \mathbf{h}) &= (E_n + \hat{x}_n) \mid (\mathbf{x}_{<n}, \mathbf{h}) \\ &= E_n \mid (\mathbf{x}_{<n}, \mathbf{h}) + \hat{x}_n \end{aligned}$$

- Difference between the random variables X_n and E_n is only a constant value of $\hat{x}_n$

Assume the discretized MoL distributed speech

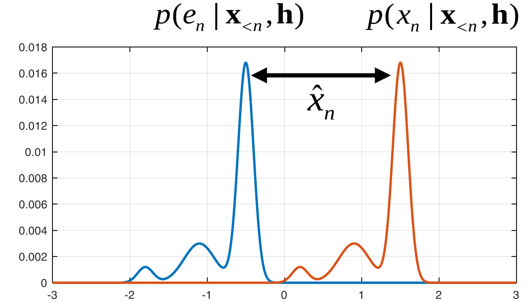
- Shifting property of 2nd order random variable

$$\pi_i^x = \pi_i^e$$

$$\mu_i^x = \mu_i^e + \hat{x}_n$$

$$s_i^x = s_i^e$$

Only mean parameters are different

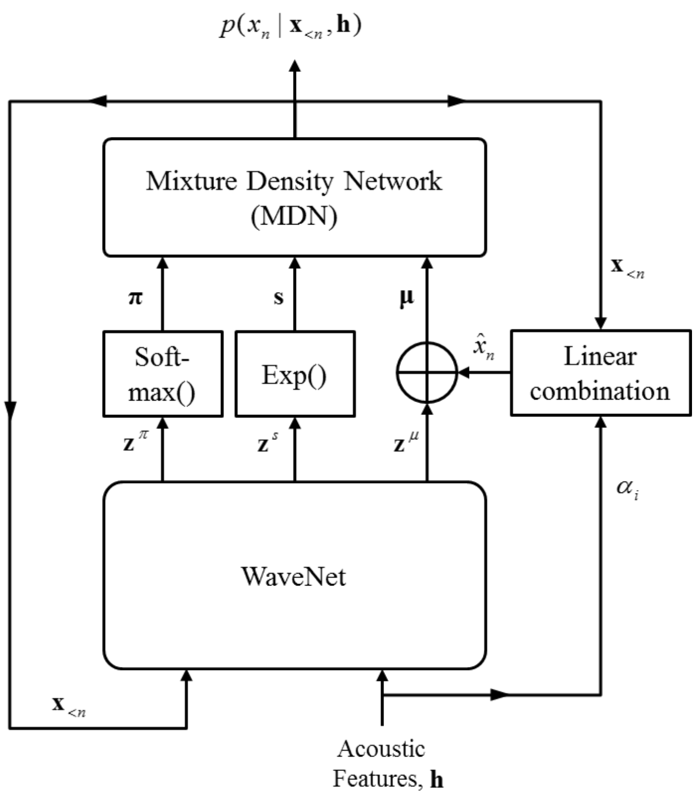


[Difference of distributions between speech and excitation]

LP-WaveNet

Recorded	ExcitNet	LP-WaveNet

Utilize the causality of WaveNet and the linearity of LP synthesis processes [7]



[LP-WaveNet]

1. Mixture parameter prediction

$$\left[\mathbf{z}^{\pi}, \mathbf{z}^{\mu}, \mathbf{z}^{s}\right] = WaveNet\left(\mathbf{x}_{<n}, \mathbf{h}\right)$$

2. Mixture parameter refinement

$$\boldsymbol{\pi} = \mathrm{softmax}(\mathbf{z}^{\pi})$$
$$\boldsymbol{\mu} = \mathbf{z}^{\mu} + \hat{x}_n$$
$$\mathbf{s} = \exp(\mathbf{z}^{s})$$

3. Discretized MoL loss calculation

$$p(x_n \mid \mathbf{x}_{<n}, \mathbf{h}) = \sum_{i=1}^{N} \pi_i \cdot \left[\sigma\left(\frac{x + \Delta/2 - \mu_i}{s_i} \right) - \sigma\left(\frac{x - \Delta/2 - \mu_i}{s_i} \right) \right]$$

$$E = \sum_n \left[-\log p(x_n \mid \mathbf{x}_{<n}, \mathbf{h}) \right]$$

TUNING OF WAVENET

1. Solution to waveform divergence problem
2. Waveform generation methods

WAVEFORM DIVERGENCE PROBLEM

Waveform divergence problem

- Waveform divergence by a stacked error during WaveNet's autoregressive generation

Cause – Overfitting on the silence region

- Unique solution in silence region

$$x_{curr} = WaveNet\left(\mathbf{x}_{prev}, \mathbf{h}\right) \Rightarrow 0 = WaveNet\left(\mathbf{0}, \mathbf{h}_{sil}\right)$$

- Easier to be happen when the portion of silence region in training set is larger
- Be sensitive to tiny error in silence region during waveform generation

Solution – Noise injection

- Inject negligible amount of noise

$$\hat{\mathbf{x}} = \mathbf{x} + \varepsilon \cdot \mathbf{n}, \text{ where } \varepsilon = 2/2^{16}$$

  - Allowing only 1 bit error

- Increase robustness to the prediction error in silence region

$$\begin{aligned}
x_0 &\sim p(x_0 \mid \mathbf{x}_{<0}) & & err_0 \\
x_1 &\sim p(x_1 \mid \mathbf{x}_{<1}) & & err_0 + err_1 \\
x_2 &\sim p(x_2 \mid \mathbf{x}_{<2}) & & \vdots \\
&\vdots & & \sum_i err_i \\
x_n &\sim p(x_n \mid \mathbf{x}_{<n}) & &
\end{aligned}$$

[Error stacking during waveform generation]

Waveform generation methods

Random sampling

$$\hat{x}_{rand}(n) \sim p(x(n) \mid x(k<n), \mathbf{h})$$

Argmax sampling

$$\hat{x}_{\max}(n) = \arg\max_{x(n)} p(x(n) \mid x(k<n), \mathbf{h})$$

Greedy sampling [8]

$$\hat{x}_{greedy}(n) = vuv \cdot \hat{x}_{\max}(n) + (1-vuv) \cdot \hat{x}_{rand}(n)$$

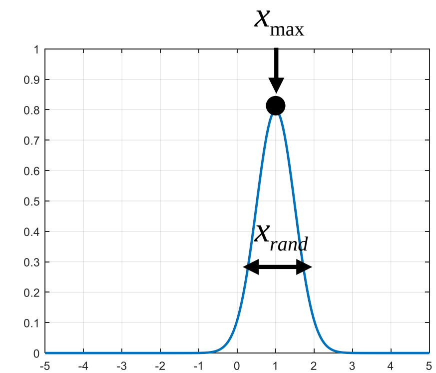
Mode sampling [7]

$$\hat{x}_{mode}(n) = vuv \cdot \hat{x}_{rand,nar}(n) + (1-vuv) \cdot \hat{x}_{rand}(n)$$

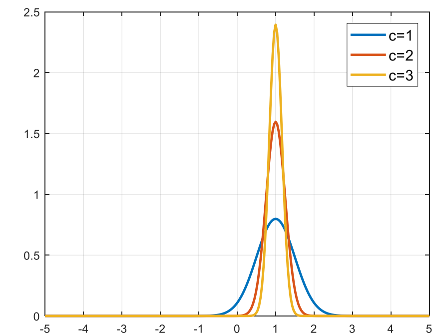
- Sharper distribution on voiced region

$$\hat{x}_{rand}(n) \sim \sum_{i=1}^{I} \pi_i DistLogic(\mu_i, s_i)$$

$$\hat{x}_{rand,narr}(n) \sim \sum_{i=1}^{I} \pi_i DistLogic\left(\mu_i, \frac{s_i}{c}\right)$$



[Random & argmax sampling]



[Scale parameter control in mode sampling]

Recorded	Random sampling	Argmax sampling	Greedy sampling	Mode sampling

EXPERIMENTS

EXPERIMENTS

Network architecture

Database	Korean male: MBC YBS database
Training / validation / Test	2,500 (~3.2h) / 200 / 200
Minibatch size	4 GPU x 20000 samples
Dilation	3 * [1, 2, 4, 8, 16, 32, 64, 128, 256, 512]
Layer	30
Receptive field	3070 samples
Residual chn.	128
Skip chn.	128
Quantization	65536 uniform
Number of mixture	10 mixture components --> 30 output channels
Conditioning features	Voicing flag, LSF, F0 (log), Eng (log), BAP
Normalization	Gaussian normalization
Learning rate	1.00E-04
Initialization / optimizer	Xavier / Adam
Sample generation method	Mode sampling

Systems

- WN_{S}: MDN-WaveNet that models the speech signal
- WN_{E}: MDN-WaveNet that models the excitation signal
- WN_{LP}: Proposed LP-WaveNet

LEARNING CURVE

Training speed

$\mathbf{WN}_{\boldsymbol{LP}} = \mathrm{WN}_E > \mathrm{WN}_S$

A/S: analysis / synthesis
SPSS: synthesis by predicted acoustic features

OBJECTIVE EVALUATION

Performance measurements

- VUV: Voicing error rate (%)
- F0 RMSE: F0 root mean square error (Hz) in voiced region
- LSD: Log-spectral distance of AR spectrum (dB)
- F-LSD: LSD of synthesized speech in frequency domain (dB)

Results

Table 1. Objective evaluation results of the various WaveNet vocoders with analysis and synthesis (A/S) and statistical parametric speech synthesis (SPSS) systems. The system with highest performance is represented in bold typeface.

	System	VUV (%)	F0 RMSE (Hz)	LSD (dB)	F-LSD (dB)
A/S	WN_S	3.62	3.98	2.22	7.7
	WN_E	3.29	3.31	**1.98**	6.97
	WN_{LP}	**3.15**	**3.30**	2.05	**6.87**
SPSS	WN_S	**6.33**	15.55	5.01	11.35
	WN_E	6.35	15.23	**4.94**	11.39
	WN_{LP}	6.56	**15.17**	4.95	**11.28**

A/S: analysis / synthesis
SPSS: synthesis by predicted acoustic features

SUBJECTIVE EVALUATION

Mean opinion score (MOS) test

- 20 random synthesized utterances from test set
- 12 native Korean speakers
- Include STRAIGHT (STR)-based speech synthesis system as baseline [9]

Results

Table 2. Subjective mean opinion score (MOS) test result with a 95% confidence interval for various speech synthesis systems. The system with highest score is represented in bold typeface. The MOS result of recorded speech was 4.81.

	STR	WN_S	WN_E	WN_{LP}
A/S	2.83±0.19	4.78±0.08	4.58±0.08	**4.84±0.11**
SPSS	2.80±0.12	**4.14±0.16**	3.67±0.20	4.04±0.12

A/S: analysis / synthesis
SPSS: synthesis by predicted acoustic features

SAMPLES

Recorded

STRAIGHT – A/S

STRAIGHT – SPSS

WN$_{S}$ – A/S

WN$_{S}$ – SPSS

WN$_{E}$ – A/S

WN$_{E}$ – SPSS

WN$_{LP}$ – A/S

WN$_{LP}$ – SPSS

SUMMARY & CONCLUSION

Investigated various types of WaveNet vocoder

- SoftMax-WaveNet
- MDN-WaveNet
- ExcitNet

Proposed an LP-WaveNet

- Explicitly represent linear prediction structure of speech waveform into WaveNet framework

Introduced tuning methods of WaveNet training / generation

- Noise injection
- Sample generation methods

Performed experiment in objective and subjective manner

- Achieve faster convergence speed than conventional WaveNet vocoders, whereas keep similar speech quality

REFERENCES

[1] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in *Proc. ICASSP*, 1996

[2] H. Zen, et. al. “Statistical parametric speech synthesis using deep neural networks,” in *Proc. ICASSP*, 2013.

[3] A. van den Oord, et. al.,“WaveNet: A generative model for raw audio generation,” in *CoRR*, 2016.

[4] A.Tamamori, et. al., “Speaker-dependent wavenet vocoder,” in *INTERSPEECH*, 2017.

[5] A. van den Oord, et. al., "Parallel WaveNet: Fast High-fidelity speech synthesis," in *CoRR*, 2017

[6] E. Song, K. Byun, and H.-G. Kang, “A neural excitation vocoder for statistical parametric speech synthesis systems,” *in Arxiv,* 2018.

[7] M. Hwang, F. Soong, F. Xie, X. Wang, and H. Kang, “LP-WaveNet: Linear Prediction-based WaveNet Speech Synthesis,” *Submitted to Proc. ICASSP*, 2019.

[8] W. Xin, et. al., “A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis,” in *Proc. ICASSP*, 2018.

[9] H. Kawahara, “STRAIGHT, exploration of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds,” *Acoustical Science and Technology*, 2006.

Thank you!