An Investigation of Noise Shaping with Perceptual Weighting for WaveNet-based Speech Generation

K. Tachibana[1], T. Toda[2,1], Y. Shiga[1] and H. Kawai[1]

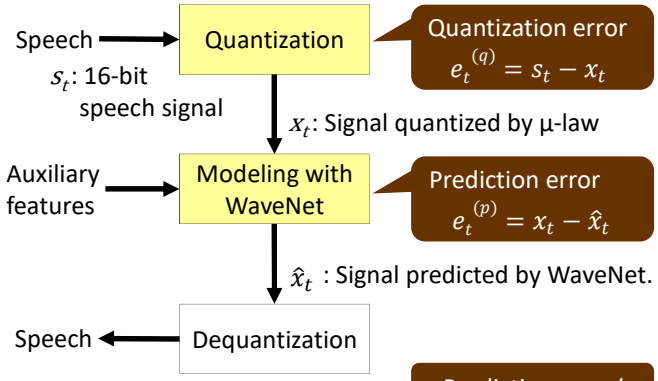
[1] National Institute of Information and Communications Technology, Japan
[2] Information Technology Center, Nagoya University, Japan

Towards Improvements of WaveNet

- WaveNet [van den Oord *et al.*, 2016]
  - CNN used as an autoregressive model
  - Modeling of quantized waveform signals as a discrete symbol sequence (*i.e.*, Markov modeling)
  - No need to use source filter model
- Our contribution towards further improvements
  - C1. Analyze noise signals generated in WaveNet
  - C2. Propose noise shaping to perceptually reduce them

C1. Analysis of noise generated in WaveNet

- Analyze quantization error & prediction error

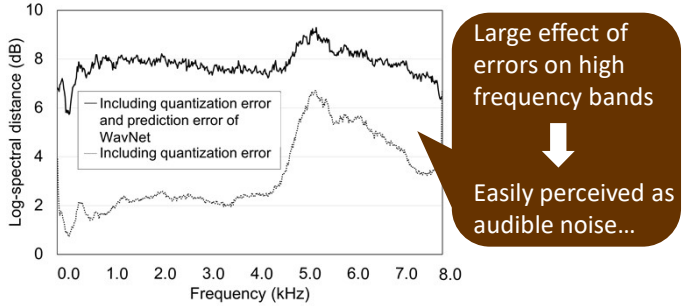


- Which error is dominant?

Noise signal	SNR [dB]	MCD [dB]
$e_t^{(q)}$	33.78 ± 0.38	4.12 ± 0.02
$e_t^{(q)} + e_t^{(p)}$	1.63 ± 0.01	2.90 ± 0.24

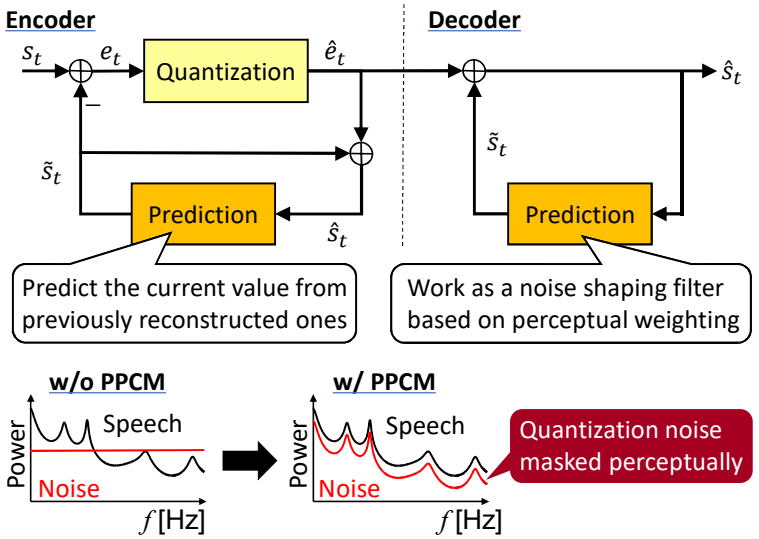
Prediction error!

- Which frequency components suffer from errors?



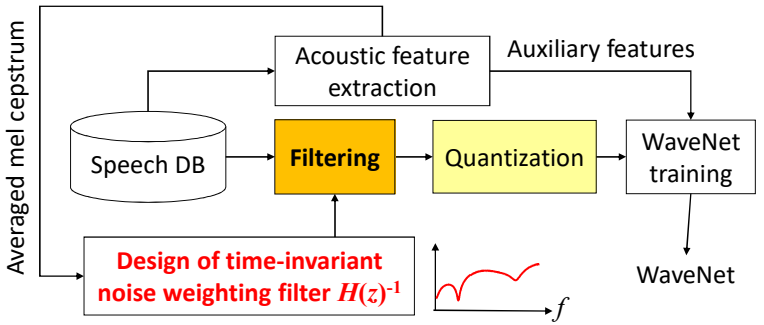
C2. Proposed noise shaping method for WaveNet

- Basic idea: noise shaping by PPCM [Atal *et al.*, 1978]

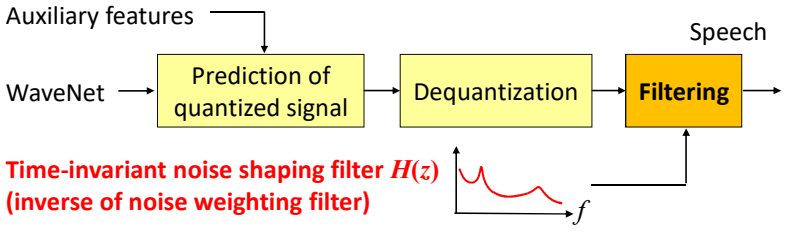


- **Proposed noise shaping procedure for WaveNet**

Training procedure (for WaveNet vocoder [Tamamori *et al.*, 2017])



Generation procedure



Experimental Evaluations

- Experimental conditions

Speech data	One Japanese female (16 kHz sampling) Training data: 7,365 sentences (4 hours) Test data: 30 sentences
Noise shaping/ weighting filter	MLSA filter (0^{th} through 39^{th} coefficients) β: 1.0 (strongly shaping) to 0.0 (no shaping)
Network architecture	Dilated causal convolution layers: 30 Convolution channels: 256 Skip channels: 2,048 Batch size: 20,000 samples Iteration times: 200,000
Objective evaluation measures	Signal-to-noise ratio (SNR) Log spectral distance (LSD) Mel-cepstral distortion (MCD)
Subjective evaluation	Preference test on naturalness Number of listeners: 15

- w/o noise shaping vs. w/ noise shaping

Result of objective evaluation

	SNR [dB]	LSD [dB]	MCD [dB]
w/o	1.79 ± 0.19	10.45 ± 0.47	3.80 ± 0.02
w/ (β=0.1)	$\mathbf{2.27 \pm 0.21}$	9.64 ± 0.36	3.52 ± 0.02
w/ (β=0.5)	2.02 ± 0.20	8.70 ± 0.28	$\mathbf{3.21 \pm 0.02}$
w/ (β=1.0)	1.59 ± 0.18	$\mathbf{8.66 \pm 0.29}$	3.34 ± 0.02

Can reduce distortion!

Result of subjective evaluation

w/o	w/	No pref	p-value
23.7	**45.3**	30.8	$<10^{-6}$

Can improve naturalness!

Effectiveness of noise shaping