17th May 2019.

Investigations of real-time Gaussian FFTNet and parallel WaveNet neural vocoders with simple acoustic features

Takuma Okamoto[1], *Tomoki Toda*[2,1], *Yoshinori Shiga*[1], and *Hisashi Kawai*[1]

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

1. Introduction

- **Target: Real-time high-fidelity text-to-speech (TTS) and voice conversion (VC)**
  - Conventional: DNN-based acoustic model with source-filter vocoders
  - State-of-the-art: Raw waveform generation-based speech synthesis with neural vocoders conditioned on mel-spectrograms
    - End-to-end TTS system Tacotron 2 with autoregressive (AR) WaveNet vocoder: Human speech quality synthesis
    - Entire end-to-end TTS system ClariNet (Deep voice 3 + single Gaussian (SG) parallel WaveNet vocoder): Real-time high-fidelity synthesis
- **Existing TTS and VC systems**
  - Introducing simple acoustic features (SAF) rather then mel-spectrograms
    - SAF: Fundamental frequency f_o and mel-cepsta for source-filter vocoders
- **Purpose: Following four investigations of neural vocoders with SAF**

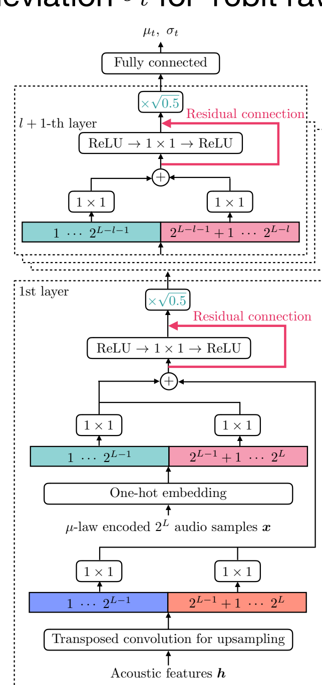
1. SG AR WaveNet and FFTNet neural vocoders with SAF
2. SG parallel WaveNet vocoder with SAF
3. Noise shaping effect in SG neural vocoders with SAF
4. Bandwidth extension effect in SG neural vocoders with SAF

2. Single Gaussian WaveNet and FFTNet vocoders

- **Single Gaussian AR WaveNet (ClariNet teacher)**
  - Single Gaussian conditional probability distribution rather than categorical one
    - Predicting continuous valued mean μ_t and standard deviation σ_t for 16bit raw audio prediction
    - Training criterion: Maximum likelihood estimation

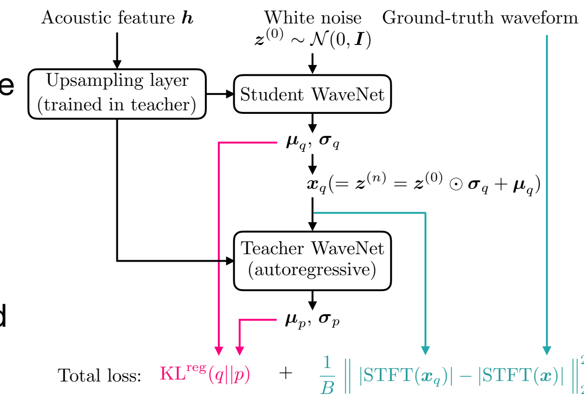
$$-\log p\left(x_t|x_{<t}\right) = \frac{1}{2}\log 2\pi + \frac{1}{2}\log \sigma_t^2 + \frac{\left(x_t-\mu_t\right)^2}{2\sigma_t^2}$$

- **Proposed single Gaussian FFTNet**
  - FFTNet: Real-time AR neural vocoder
  - SG modeling can be directly applied to FFTNet
    - With additional residual connections
- **Noise shaping considering auditory perception** (K. Tachibana *et al*. ICASSP 2018)
  - Improving synthesis quality by reducing spectral distortion due to prediction error in categorical WaveNet and FFTNet (T. Okamoto *et al*. SLT 2018)
- **Investigations**
  - Can SG AR WaveNet and FFTNet be trained with SAF?
  - Can noise shaping improve synthesis quality of SG neural vocoders?



3. Single Gaussian parallel WaveNet (ClariNet)

- **Knowledge distillation (teacher-student training) based on Gaussian inverse autoregressive flow (IAF)**
- **Loss functions for non-AR student WaveNet**
  - Regularized Kullback-Leibler (KL)-divergence
  - Spectrogram frame loss for avoiding whisper voice problem
- **Comparison with conventional mixture of logistics (MoL)-based parallel WaveNet**
  - KL-divergence can be analytically calculated
  - Only initial sampling $z^{(0)}$ is sufficient
- **Investigation**
  - Can SG parallel WaveNet be trained with SAF instead of mel-spectrograms?

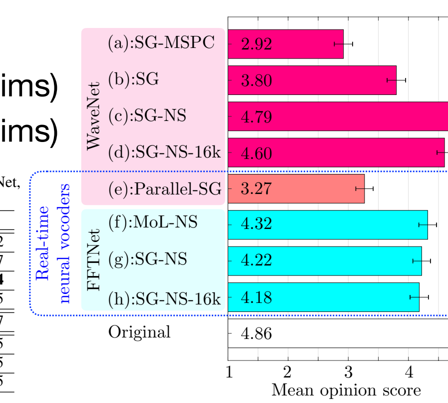


4. Experiments

- **Corpus: Japanese male speech (3.7 hours, $fs = 24$ kHz)**
- **Acoustic features**
  - MSPC: 80-dim. mel-spectrograms (125 to 7600 Hz)
  - SAF 24k Hz: $\log f_o$ + vuv + 35-dim. mel-cepstra (37-dims)
  - SAF 16k Hz: $\log f_o$ + vuv + 25-dim. mel-cepstra (27-dims)

Table 1. Results of objective evaluations of 20 test set utterances. "NS" and "Parallel WN" denote noise shaping and parallel WaveNet, respectively.

Method	Network	Type	Input features	Real-time	NS	SNR [dB]	SD [dB]	MCD [dB]
(a):WN-SG-MSPC	WaveNet	SG	Mel-spectrogram			3.30 ± 0.39	9.34 ± 0.20	3.71 ± 0.12
(b):WN-SG-SAF	WaveNet	SG	SAF 24 kHz			$\mathbf{5.40 \pm 0.44}$	8.02 ± 0.08	2.55 ± 0.07
(c):WN-SG-SAF-NS	WaveNet	SG	SAF 24 kHz		✓	3.90 ± 0.73	$\mathbf{7.57 \pm 0.08}$	$\mathbf{2.20 \pm 0.04}$
(d):WN-SG-SAF16k-NS	WaveNet	SG	SAF 16 kHz		✓	3.70 ± 0.69	8.26 ± 0.07	2.89 ± 0.05
(e):PWN-SG-SAF	Parallel WN	SG	SAF 24 kHz	✓		5.20 ± 0.41	8.09 ± 0.06	2.73 ± 0.07
(f):FN-MoL-SAF-NS	FFTNet	MoL	SAF 24 kHz	✓	✓	3.20 ± 0.66	7.96 ± 0.08	2.69 ± 0.05
(g):FN-SG-SAF-NS	FFTNet	SG	SAF 24 kHz	✓	✓	2.90 ± 0.66	8.01 ± 0.08	2.80 ± 0.05
(h):FN-SG-SAF16k-NS	FFTNet	SG	SAF 16 kHz	✓	✓	3.10 ± 0.66	8.53 ± 0.07	3.36 ± 0.05



5. Extended investigations

- **Using a larger amount of training data (27 hours)**
  - Synthesized quality can be improved
- **Multi-resolution frame loss (MRFL) in parallel WaveNet**

$$\sum_{i=1}^{3}\frac{1}{B_i}\left\| |\mathrm{STFT}\left(x_q\right)| - |\mathrm{STFT}(x)| \right\|_2^2$$

$$B_1 = 1025,\ B_2 = 513,\ B_3 = 257$$

  - Synthesized quality can be slightly improved
- **WaveRNN and WaveGlow neural vocoders with SAF**
  - Successfully synthesize high-quality speech waveforms
    - Demo samples are available in the poster session (8:30-11:30 17th May)