21th Dec. 2018.

# Improving FFTNet vocoder with noise shaping and subband approaches


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

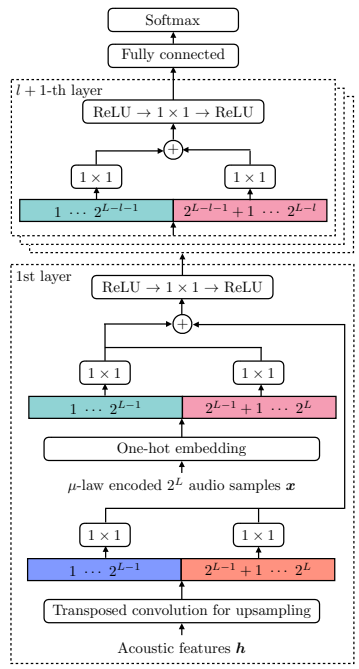
[1]National Institute of Information and Communications Technology, Japan, [2]Nagoya University, Japan

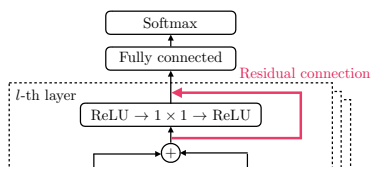
## 1. Introduction

- **Target: High-quality statistical parametric speech synthesis and voice conversion**
  - Raw audio generative models: WaveNet, SampleRNN, FFTNet and WaveRNN
    - Outperforming conventional concatenative and source filter vocoder syntheses
    - Synthesis time problem due to autoregressive modeling
- **Raw audio generative models with real-time synthesis**
  - Parallel WaveNet and WaveRNN
    - High quality but network structures not disclosed
  - FFTNet vocoder (Z. Jin et al., ICASSP 2018)
    - High speed synthesis but not so high synthesis quality
- **Purpose: Improving FFTNet neural vocoder**
  - Realizing high quality synthesis while keeping network model size

## 2. FFTNet neural vocoder

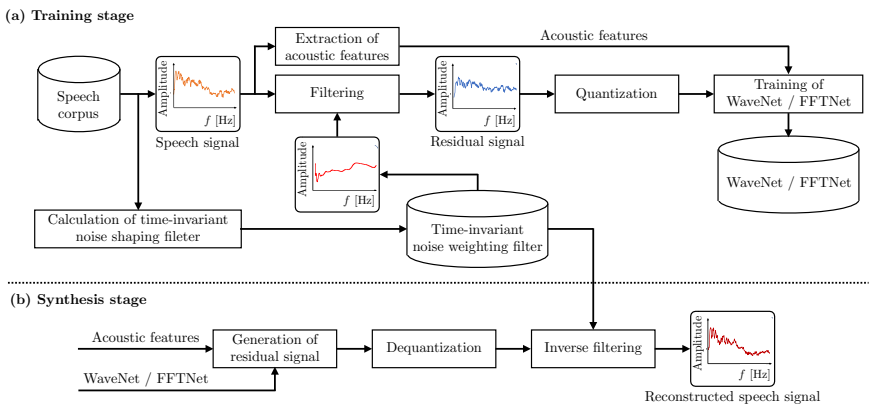
- **Network structure**
  - Input: 256-way one hot vectors and acoustic features
  - Output: 256-way one hot vector representing 8 bit mu-law audio
  - Simpler structure than WaveNet
    - Mainly with 1x1 conv and ReLU
- **Network modifications while keeping network model size**
  - Skip connections: Not effective
  - Residual connections: Effective



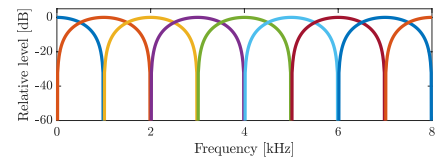

## 3. FFTNet vocoder with noise shaping

- **Noise shaping method considering auditory perception** (K. Tachibana et al., ICASSP 2018)
  - Improving synthesis quality by reducing spectral distortion due to prediction error
  - Implemented by MLSA filter with averaged mel-cepstrums
    Efficient in WaveNet vocoder

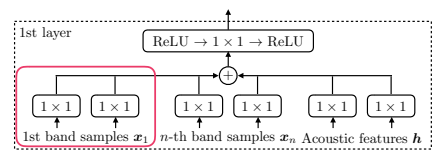


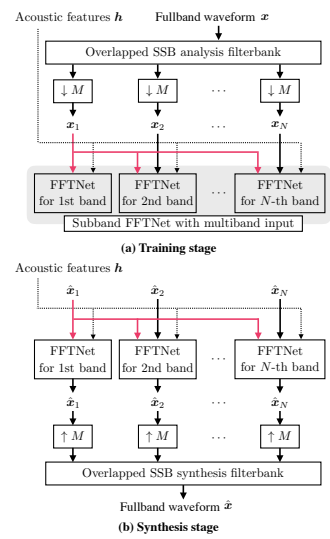

## 4. Subband FFTNet

- **Parallel training and synthesis in raw audio generative models with single-sideband filterbanks**
  - Squared-root Hann window-based overlapped filterbank
    - Improving prediction accuracy of WaveNet by coloring each band signal (T. Okamoto et al., ASRU 2017)




  - Subband WaveNet vocoder (T. Okamoto et al., ICASSP 2018)
    - Problem: Phase shift between subbands
- **Subband FFTNet with multiband input**
  - Simultaneously input multiband signals
    - For considering phase shift between subbands
    - Input only 1st band signals for reducing network parameters




## 5.Experiments

- **Speech corpus (Sampling frequency: 16 kHz)**
  - Japanese male voice (Training set: 5697 utterances [3.7 h])
- **Input acoustic features (27 dimensions)**
  - (Log) Fundamental frequency + v/uv: 2 dimensions
  - STFT-based simple mel-cepstrums: 25 dimensions
- **Network model size comparison**
  - 1/20 compared with WaveNet

| Model | Num of params |
|---|---|
| WaveNet (g) and (h) | 44,592,721 |
| FFTNet (a) to (c) | 2,251,857 |
| Subband FFTNet (d) and (e) | 1,857,105 (each subband) |
| Subband FFTNet with multiband input (f) | 1,988,117 (each subband) |

- **Objective evaluation results (Test set: 20 utts)**

| | Training softmax loss score | SNR [dB] | SD [dB] | MCD [dB] |
|---|---|---|---|---|
| (a):vanilla FFTNet (baseline) | 1.89 | 5.20 ± 0.26 | 10.29 ± 0.15 | 3.66 ± 0.11 |
| (b):FFTNet with residual connections | 1.81 | 5.50 ± 0.25 | 9.68 ± 0.12 | 3.33 ± 0.08 |
| (c):FFTNet with noise shaping | 2.19 | 4.00 ± 0.47 | **8.19 ± 0.05** | **2.84 ± 0.06** |
| (d):subband FFTNet | 1.39 | 4.00 ± 0.27 | 10.76 ± 0.30 | 2.96 ± 0.04 |
| (e):subband FFTNet with noise shaping | 1.55 | 2.90 ± 0.39 | 9.62 ± 0.22 | **2.84 ± 0.06** |
| (f):subband FFTNet with multiband input | ***1.35*** | **5.80 ± 0.36** | 10.84 ± 0.36 | 3.13 ± 0.39 |
| (g):vanilla WaveNet | 1.50 | *6.60 ± 0.36* | 9.16 ± 0.12 | 2.50 ± 0.08 |
| (h):WaveNet with noise shaping | 1.80 | 5.50 ± 0.60 | 7.58 ± 0.06 | *2.00 ± 0.07* |
| (i):STRAIGHT | – | 0.10 ± 0.47 | *7.09 ± 0.07* | 2.78 ± 0.08 |

- **MOS evaluation condition**
  - Test set: 20 utterances
  - 10 listening subjects
  - Using headphones
- **MOS results**
  - Improvement by proposal
  - But lower than WaveNet with noise shaping