An investigation of subband WaveNet vocoder covering entire audible frequency range with limited acoustic features

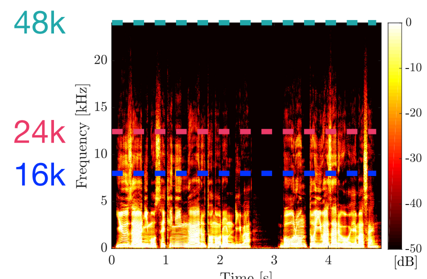
19th Apr. 2018.

Takuma Okamoto[1], Kentaro *Tachibana*[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

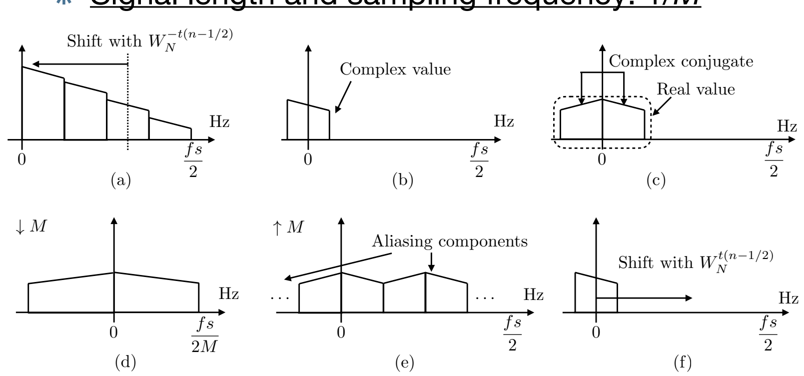
1. Introduction

- **Target: High-quality statistical parametric speech synthesis**
  - Conventional: DNN-based acoustic model with source-filter model-based vocoder
  - State-of-the-art: Raw waveform generation-based speech synthesis
    - Parallel WaveNet and WaveRNN: Linguistic features to raw waveforms (24k)
    - End-to-end text-to-speech synthesis with neural vocoders
      Char2wav (16k), Deep voice 3 (48k), Tacotron 2 (24k)
- **Purpose: Raw waveform generation-based high-quality speech synthesis covering entire human audible frequency range with subband WaveNet architecture**
  - Source-filter model-based vocoders with a sampling frequency (fs) of 48 kHz
    - Marlin toolkit and GlottDNN
  - Only Deep voice 3 introduces $fs = 48$ kHz
    - Unknown network structure
    - Huge GPU memory required for training
  - Introducing subband WaveNet architecture into neural vocoder for $fs = 48$ kHz
    - Smaller network size trainable by consumer GPUs with small memory
    - Only investigated "unconditional" training and synthesis with $fs = 32$ kHz
  - Investigating bandwidth extension effect with bandlimited acoustic features

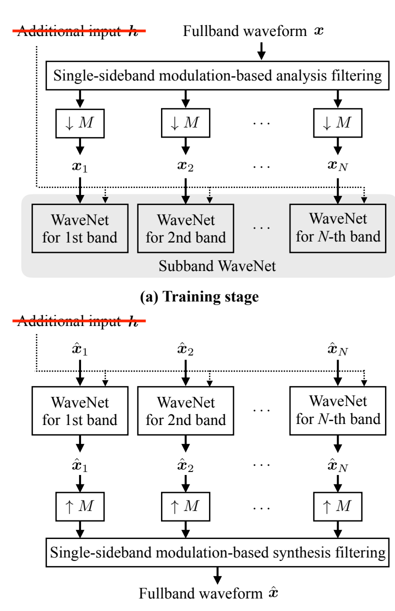


2. Subband WaveNet

- **Multirate signal processing**
  - Dividing fullband signal into N subband signals and decimating them with a factor M
    - Signal length and sampling frequency: 1/M



- **Square-root Hann window-based overlapped filterbank**
  - Easier training with colored subband signals
    - Realizing higher quality synthesis than fullband WaveNet in "unconditional" training and synthesis

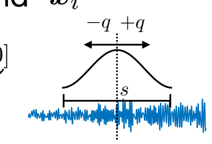


T. Okamoto *et al*. ASRU 2017

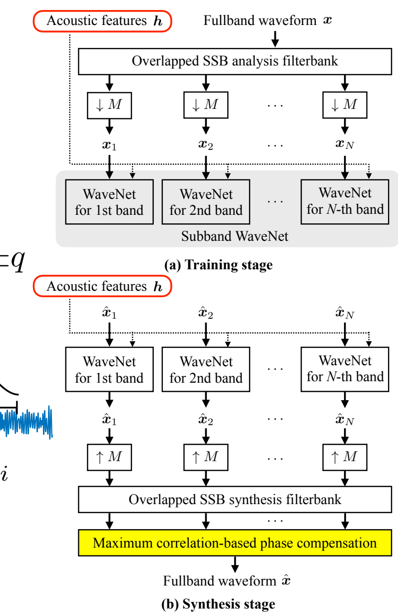
3. Subband WaveNet vocoder

- **Subband WaveNet conditioned on acoustic features**
- **Introducing maximam correlation-based phase compensation between subbands in synthesis stage**
  - Using common frequency component between adjacent subbands
    1. Finding a time shift for higher subband $x_{\mathrm{high},i}$ within $\pm q$ that maximize correlation between $x_{\mathrm{high},i}$ and x_i

$$x_i = x_{\mathrm{low},i} + [x(s/2+1)_{\mathrm{high},i-1}, \cdots x(s)_{\mathrm{high},i-1}, \underbrace{0, \cdots 0}_{s/2}]$$

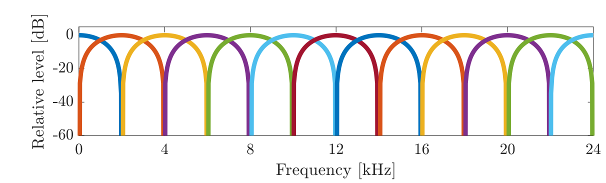


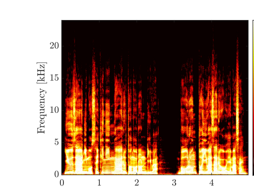
    2. $x_{\mathrm{high},i}$: Overlap-and-added
    3. Sequentially compensated from low subbands



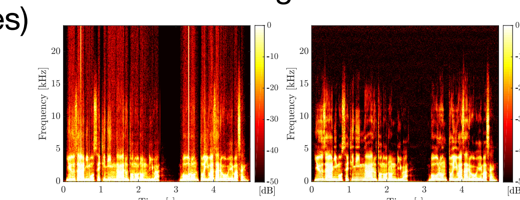
4. Experiments

- **Japanese male speech corpus with a sampling frequency of 48 kHz**
  - 3.7 hours for training set, 23 utterances for test set
- **Subband WaveNet vocoder setting**
  - Filterbank (M = 6 and N = 13) $fs = 8$ kHz
    - Prototype FIR filter (1535 samples)
  - Acoustic features: analyzed every 5 ms
    - Fundamental frequency (f_o): analyzed by NDF
    - STFT-based simple mel-cepstrums:
      35 dims (48 kHz), 25 dims (16 kHz), 17 dims (8 kHz)
  - Time resolution adjustment between h and x
    - Simple copy (No transposed convolution)
  - WaveNet model (Parameter update: 100,000 times)
    - Receptive field: 0.192 s (9 x 3 = 27 layers)
- **Baseline (Source-filter model-based vocoders)**
  - MLSA (f_o + STRAIGHT mel-cepstrums 50 dims)
  - STRAIGHT (f_o + STRAIGHT mel-cepstrums 60 dims + aperiodicity 25 dims)
- **MOS test with 15 listening subjects**
  - MNRU: $y(t) = x(t) + 10^{-Q/20}x(t)n(t)$
  - 11 types x 23 sentences = 253 evaluation utterances
- **Results**
  - Proposal with fullband features outperformed others
  - Higher frequency components of h are required





Original



Fullband Subband