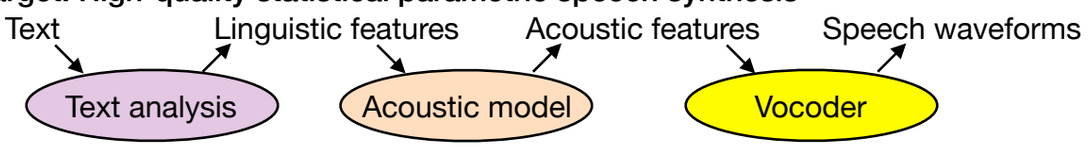
Subband WaveNet with overlapped single-sideband filterbanks

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**



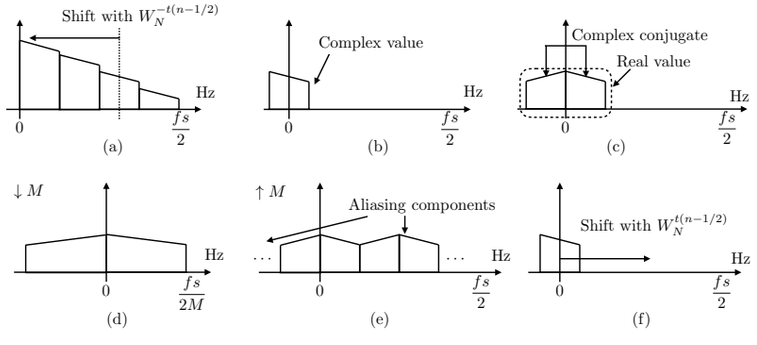
  - Conventional: HMM to DNN STRAIGHT, WORLD, GlottDNN etc..
  - State-of-the-art: 1. WaveNet
    2. End-to-end: Char2wav, Deep voice, Tacotron
- **WaveNet: Dilated causal convolutional NN-based raw audio generative model**
  - Autoregressive model to directly predict raw audio samples

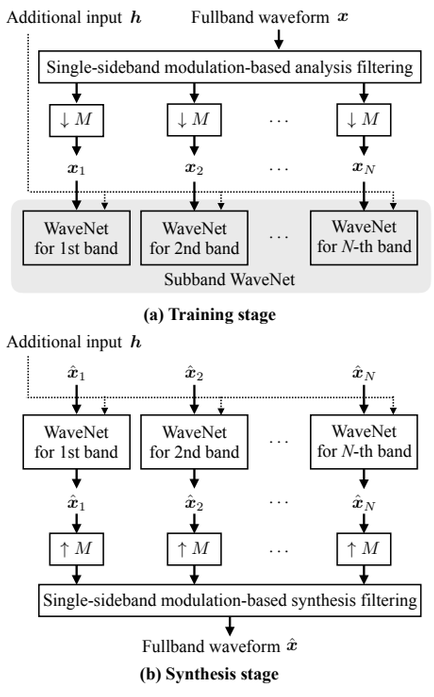
$$p(\boldsymbol{x}|\boldsymbol{h}) = \prod_{t=1}^{T} p(x(t)|x(1), \cdots, x(t-1), \boldsymbol{h})$$

    - ✵ Categorical problem (8bit μ-law encoding) rather than regression
  - Synthesis speed problem since past samples are required
    - ✵ **Parallel WaveNet**: Fast generation but complicated training architecture
- **Proposal: Rapid and high quality synthesis for WaveNet with simple approach**
  - Parallel training and synthesis based on Multirate signal processing
  - Realizing high-quality synthesis with square-root Hann window-based filterbanks

2. Subband WaveNet

- **Single-sideband (SSB) filterbank**
  1. Dividing fullband signal into N subband signals
  2. Decimating them with a factor M
     - ✵ Signal length and sampling frequency: 1/M
  3. Upsampling and inverse processing
     - ✵ Reconstructing fullband signal

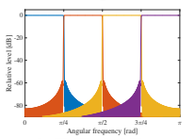


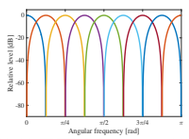


- **Subband WaveNet**
  - WaveNet training and synthesis for each subband waveform
    Enabling parallel synthesis: M times synthesis speed

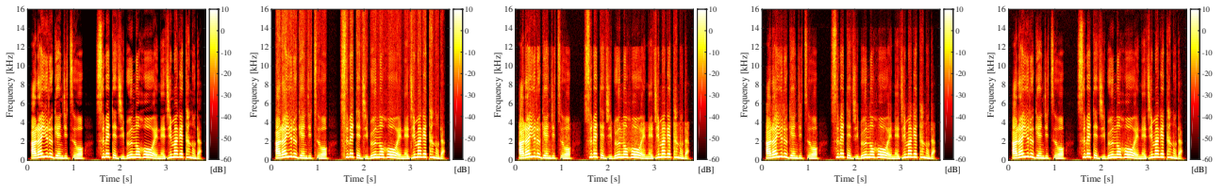
3. Experiments

- **Japanese speech corpora with a sampling frequency of 32 kHz**
  - Female and male corpora: 4.7 and 3.7 hours for training sets, 100 utts for test sets
- **Filterbank setting (Decimation factor: M=4)**
  - LPF-MD: N=4 with simple lowpass filter
  - LPF-OL: N=9 with simple lowpass filter
  - SQRT-Hann-OL: N=9
    with overlapped square-root Hann window-based analysis/synthesis filter

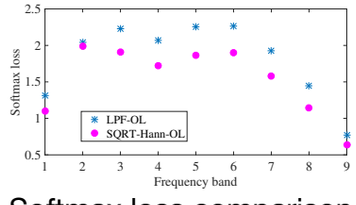

LPF-MD


SQRT-Hann-OL

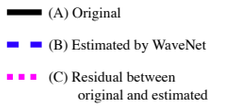
- **WaveNet models**
  - Unconditional WaveNet training and synthesis without additional input
    - ✵ Predicting $\hat{x}(t)$ from correct input $[x(1), \cdots, x(t-1)]$ ⇨ $\hat{\boldsymbol{x}} = [\hat{x}(1), \cdots, \hat{x}(T)]$
  - Receptive field: 0.192 s, mini-batch size: 2.5 s with Adam optimizer
  - Dilation channel: 32, Residual channel: 32, Skip channel: 512
  - Number of parameter update: 20 k (Fullband), 10 k (Subband)
- **Results of synthesis speed with CPUs: Female (3.85 s) and male (3.87 s)**
  - Fullband: 11.40 and 11.21 mins, Subband: 2.68 and 2.65 mins ⇨ about 4 times
- **Results of objective evaluations**
  - Proposed SQRT-Hann-OL colors each subband waveform
    - ✵ Improving prediction accuracy: Realizing higher-quality synthesis than fullband

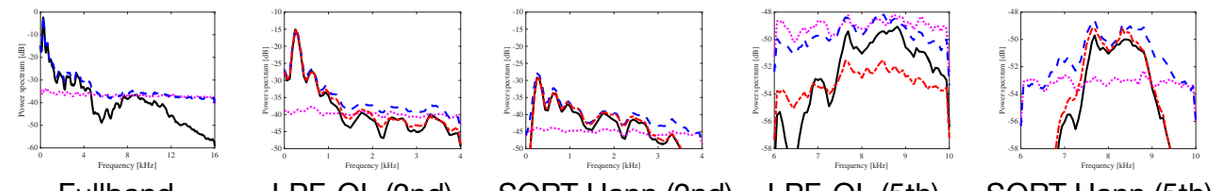

Original Fullband LPF-MD LPF-OL SQRT-Hann-OL

	Method	SNR [dB] (WaveNet)	SD [dB] (WaveNet)	MCD [dB] (WaveNet)
Female $fs = 32$ kHz	Fullband	**21.5 ± 0.35**	8.72 ± 0.07	2.45 ± 0.03
	LPF-MD	13.6 ± 0.28	7.68 ± 0.07	2.80 ± 0.04
	LPF-OL	13.1 ± 0.22	7.16 ± 0.05	1.95 ± 0.04
	SQRT-Hann-OL	13.0 ± 0.27	**6.45 ± 0.09**	**1.72 ± 0.04**
Male $fs = 32$ kHz	Fullband	**23.2 ± 0.26**	9.51 ± 0.08	2.75 ± 0.03
	LPF-MD	14.0 ± 0.25	8.17 ± 0.08	2.75 ± 0.05
	LPF-OL	13.4 ± 0.26	7.48 ± 0.06	2.10 ± 0.04
	SQRT-Hann-OL	13.6 ± 0.31	**7.26 ± 0.10**	**2.07 ± 0.06**



Softmax loss comparison




Fullband LPF-OL (2nd) SQRT-Hann (2nd) LPF-OL (5th) SQRT-Hann (5th)

- **Results of subjective evaluations: Paired comparison with 21 listening subjects**
  - Significant quality improvement by proposed SQRT-Hann-OL

	Fullband	SQRT-Hann-OL	*Neutral*	*p*-value	*Z*-score
Female (%)	33 (6.3)	**413 (78.7)**	79 (15.0)	$\ll 10^{-10}$	−18.0

	Fullband	SQRT-Hann-OL	*Neutral*	*p*-value	*Z*-score
Male (%)	31 (5.9)	**439 (83.6)**	55 (10.5)	$\ll 10^{-10}$	−18.8