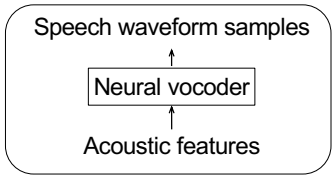
STFT spectral loss for training a neural speech waveform model

Shinji Takaki[1], Toru Nakashika[2], Xin Wang[1], Junichi Yamagishi[1,3]

[1] National Institute of Informatics, [2] The University of Electro-Communications, [3] The University of Edinburgh

Introduction

Neural vocoder
- Statistical vocoder
- High-performance
- Training difficulty
- Complex architecture

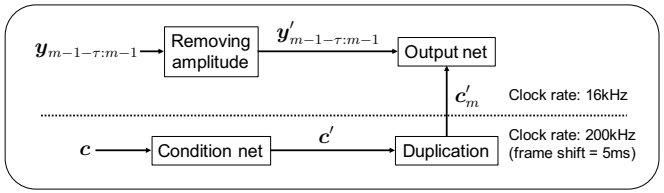


STFT spectral loss
- Loss in the short-time Fourier transform domain
  - Model training based on amplitude and phase info.
- The spectrum can contain multiple waveform samples

Simple Network architecture
- uni-LSTM + auto-regressive structure

Analysis-by-synthesis & TTS experiments using WORLD, WaveNet, and Proposed vocoders

Network architecture



Removing amplitude from feedback waveform
- Feedback + teacher-forcing training → over-fitting

$$\boldsymbol{y}'_{m-1-\tau:m-1} = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\boldsymbol{y}_{m-1-\tau:m-1})}{|\mathcal{F}(\boldsymbol{y}_{m-1-\tau:m-1})|}\right)$$

$\mathcal{F}$: DFT

$\mathcal{F}^{-1}$: inverse DFT

Network architecture
- Condition net
  - biLSTM (#hidden units: 80)
  - CNN (#filters: 80, filter size: 5 (time) x 80 (freq.))
- Duplication: 1 → 80
- Output net
  - Feedback size: $f_\tau = 400$
  - uni-LSTM x 3 (#hidden units: 256)

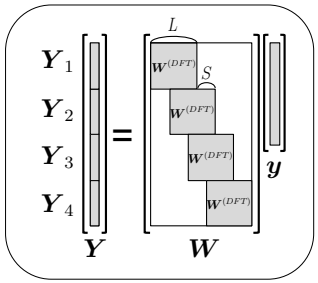
STFT spectrum

Complex spectrum

$\boldsymbol{W} \in \mathbb{C}$: STFT matrix

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$$

$$Y_{t,n} = \boldsymbol{W}_{t,n}\boldsymbol{y}$$



Amplitude spectrum

$$A_{t,n} = |Y_{t,n}| = (\boldsymbol{y}^\top \boldsymbol{W}_{t,n}^H \boldsymbol{W}_{t,n}\boldsymbol{y})^{\frac{1}{2}}$$

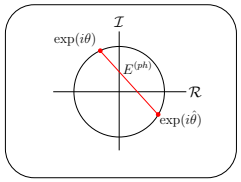
Phase spectrum (Euler's formula)

$$\exp(i\theta_{t,n}) = \exp(i\angle Y_{t,n}) = \frac{\boldsymbol{W}_{t,n}\boldsymbol{y}}{(\boldsymbol{y}^\top \boldsymbol{W}_{t,n}^H \boldsymbol{W}_{t,n}\boldsymbol{y})^{\frac{1}{2}}}$$

STFT spectral loss

Amplitude spectral loss

$$E_{t,n}^{(amp)} = \frac{1}{2}(\hat{A}_{t,n} - A_{t,n})^2$$
$$= \frac{1}{2}(\hat{A}_{t,n} - (\boldsymbol{y}^\top \boldsymbol{W}_{t,n}^H \boldsymbol{W}_{t,n}\boldsymbol{y})^{\frac{1}{2}})^2$$



Phase spectral loss

$$E_{t,n}^{(ph)} = \frac{1}{2}\left|1 - \exp(i(\hat{\theta}_{t,n} - \theta_{t,n}))\right|^2$$
$$= 1 - \frac{1}{2}\left(\frac{\hat{Y}_{t,n}}{\hat{A}_{t,n}}\frac{(\boldsymbol{y}^\top \boldsymbol{W}_{t,n}^H \boldsymbol{W}_{t,n}\boldsymbol{y})^{\frac{1}{2}}}{\boldsymbol{W}_{t,n}\boldsymbol{y}} + \frac{\overline{\hat{Y}}_{t,n}}{\hat{A}_{t,n}}\frac{(\boldsymbol{y}^\top \boldsymbol{W}_{t,n}^H \boldsymbol{W}_{t,n}\boldsymbol{y})^{\frac{1}{2}}}{\overline{\boldsymbol{W}}_{t,n}\boldsymbol{y}}\right)$$

Loss function for model training

$$E^{(sp)} = \sum_{t,n}(E_{t,n}^{(amp)} + \alpha_{t,n}E_{t,n}^{(ph)})$$

$\alpha_{t,n}$ is a weight parameter

- $\alpha_{t,n} = 0$: Using only the amplitude loss
- $\alpha_{t,n} = 1$: A simple combination
- $\alpha_{t,n} = v_t$: The phase loss computed in unvoiced parts is omitted

v_t is a voiced/unvoiced flag (1: voiced, 0: unvoiced)

Minimization of $E^{(sp)}$ is equivalent to that of L

$$L = -\log\prod_{t,n} P_g(\hat{A}_{t,n} \mid A_{t,n}, 1)P_{vm}^{\alpha_{t,n}}(\hat{\theta}_{t,n} \mid \theta_{t,n}, 1)$$

Gauss dist. von Mises dist.

Experiment

Database

Database	CMU-ARCTIC slt Train 1,032 / Test 50 utterances
Sampling rate	16000 Hz

Condition feature (Mel-spectrogram)

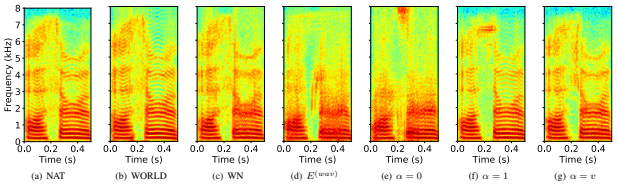
Dimension	80
Frame shift	80 points (5ms)
Frame length	400 points (25ms)

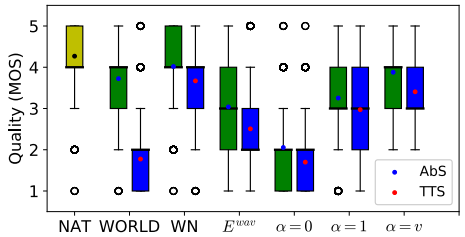
Output analysis

Frame shift	1 point
Frame length	400 points (25ms)

Training
- ADAM (Learning rate: 0.0005, #iteration: 100k)
- Mini-batch (0.125sec x 40 = 5sec)

Synthetic spectrogram & Subjective evaluation





- Omitting phase loss → noisy spectrograms
- Using v/u flags → Improving performance
- WORLD vs. Proposed model ($\alpha = v$)
  - The proposed model is robust
- WaveNet vs. Proposed model ($\alpha = v$)
  - The performance of proposed model is slightly worse

Further investigations

- Network w/o auto-regressive structure (SLP-L8.6)
- Other time-frequency analysis (CWT case: 1903.12392)