TEXT-TO-SPEECH SYNTHESIS USING STFT SPECTRA BASED ON LOW-/MULTI-RESOLUTION GENERATIVE ADVERSARIAL NETWORKS

Yuki Saito, **Shinnosuke Takamichi**, and Hiroshi Saruwatari

The University of Tokyo, Japan

1. SYNOPSIS

In Text-To-Speech (TTS) synthesis w/ Short-Term Fourier Transform (STFT) spectra:

(1) Proposes Generative Adversarial Networks (GANs)[1]-based training algorithm using low-/multi-resolution spectra.

(2) Demonstrates that the proposed algorithm using GANs w/ low-resolution spectra improves synthetic speech quality.

2. CONVENTIONAL ALGORITHMS

DNN-based TTS w/ STFT spectra[2]

(1) Generate raw STFT amplitude spectra.

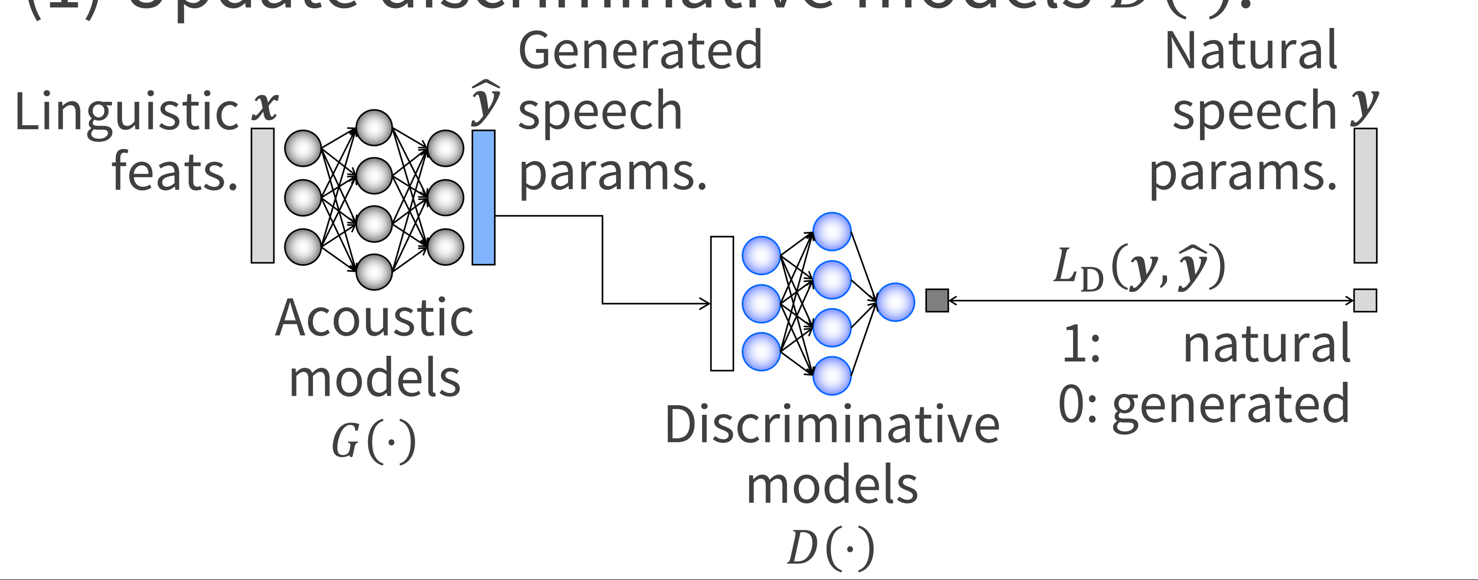
(2) Reconstruct phase spectra by using Griffin and Lim's method[3].

Pros. avoiding a vocoding process

Cons. over-smoothing of amplitude spectra

GAN-based TTS w/ vocoders[4]

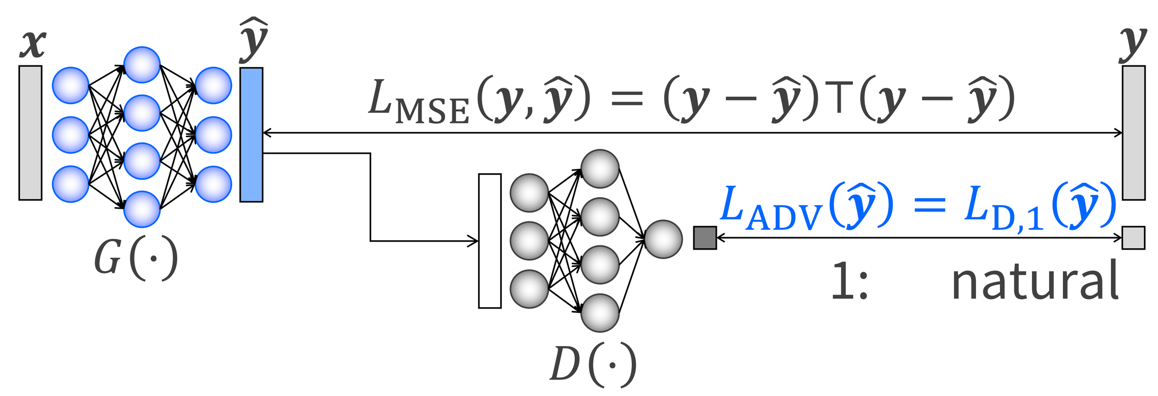
(1) Update discriminative models $D(\cdot)$.



$$L_{\mathrm{D}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \underbrace{-\textstyle\sum_t \log D(\boldsymbol{y}_t)}_{L_{\mathrm{D},1}(\boldsymbol{y})} \underbrace{- \textstyle\sum_t \log(1 - D(\hat{\boldsymbol{y}}_t))}_{L_{\mathrm{D},0}(\hat{\boldsymbol{y}})}$$

$L_{\mathrm{D},1}(\boldsymbol{y})$ (loss for **natural** params.) $L_{\mathrm{D},0}(\hat{\boldsymbol{y}})$ (loss for **generated** params.)

(2) Update acoustic models $G(\cdot)$.



$$L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\mathrm{MSE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \underbrace{\omega_{\mathrm{D}} \frac{E_{\hat{\boldsymbol{y}}}[L_{\mathrm{MSE}}]}{E_{\hat{\boldsymbol{y}}}[L_{\mathrm{ADV}}]} L_{\mathrm{ADV}}(\hat{\boldsymbol{y}})}_{\text{Minimizing approx. JS-divergence}}$$

3. PROPOSED ALGORITHMS

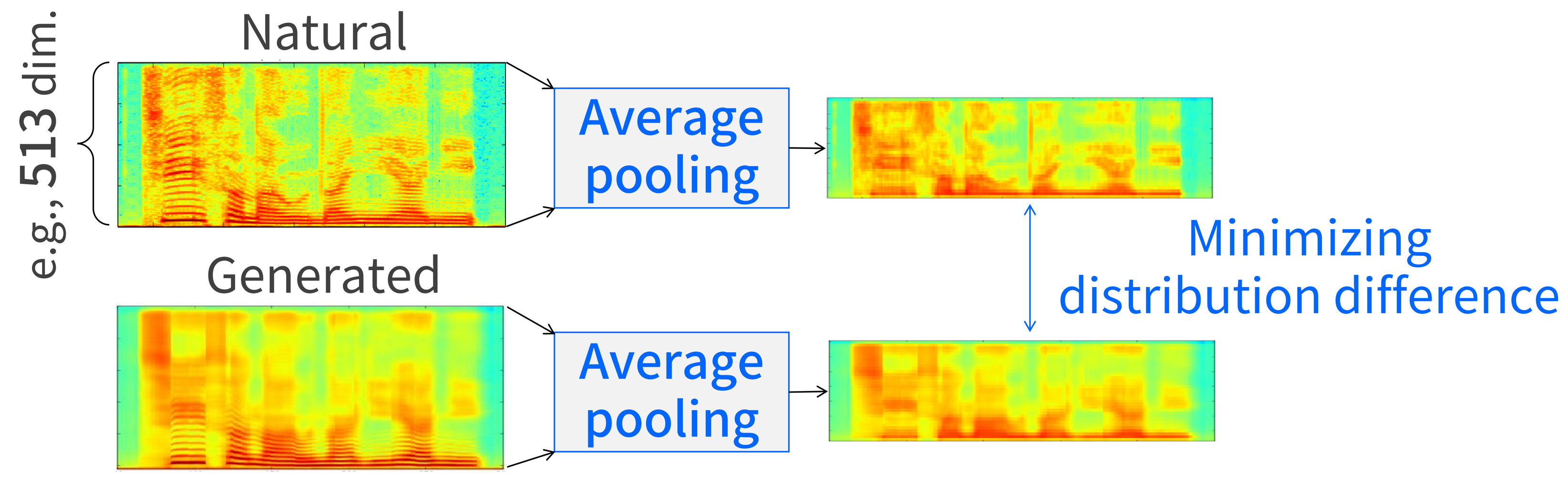
Motivation

Amplitude spectra: high-dimensional features including spectral & excitation characteristics.

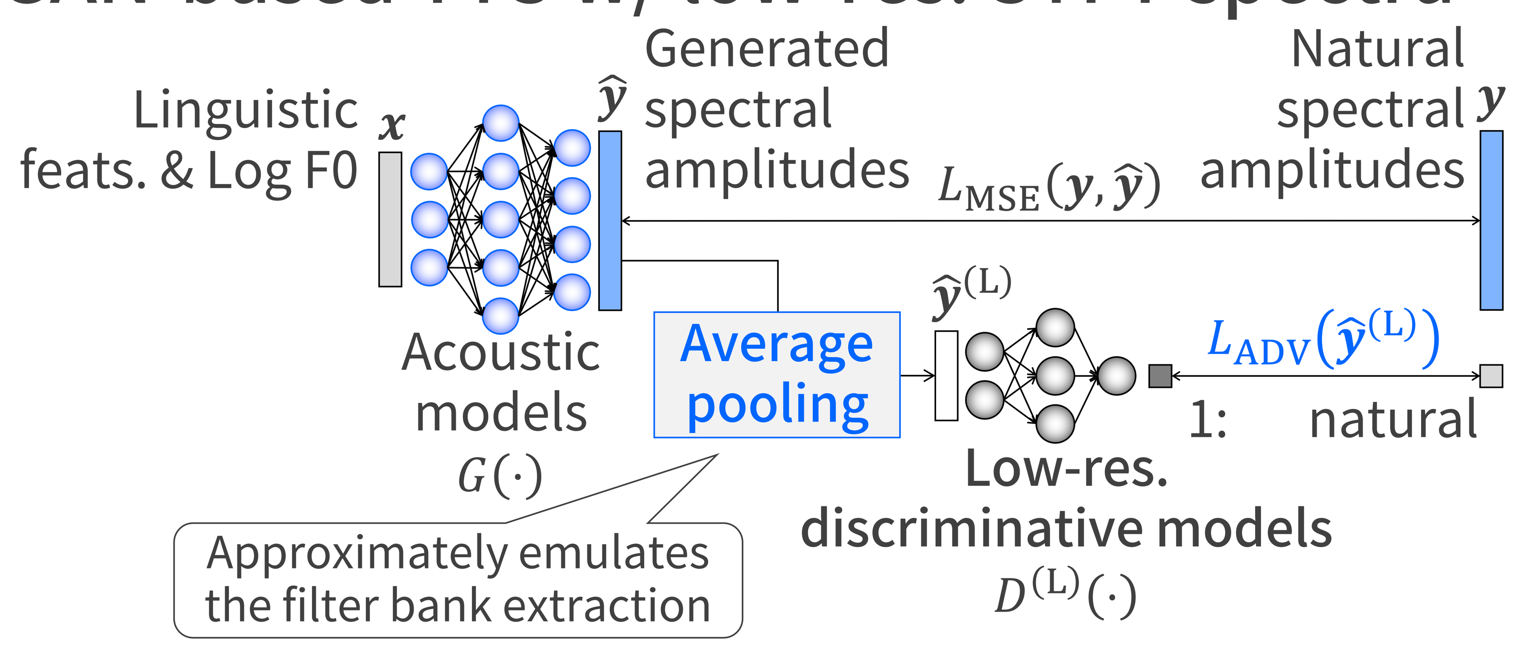
GAN-based TTS[3]: effective for spectral envelopes (i.e., mel-cepstral coefficients).

Assumption:

low-res. spectra ≒ spectral envelopes



GAN-based TTS w/ low-res. STFT spectra



$$L_{\mathrm{G}}^{(\mathrm{Low})}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\mathrm{MSE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \omega_{\mathrm{D}}^{(\mathrm{L})} \frac{E_{\hat{\boldsymbol{y}}}[L_{\mathrm{MSE}}]}{E_{\hat{\boldsymbol{y}}^{(\mathrm{L})}}[L_{\mathrm{ADV}}]} L_{\mathrm{ADV}}(\hat{\boldsymbol{y}}^{(\mathrm{L})})$$

GAN-based TTS w/ multi-res. STFT spectra

$$L_{\mathrm{G}}^{(\mathrm{Multi})}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\mathrm{MSE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \omega_{\mathrm{D}}^{(\mathrm{L})} \frac{E_{\hat{\boldsymbol{y}}}[L_{\mathrm{MSE}}]}{E_{\hat{\boldsymbol{y}}^{(\mathrm{L})}}[L_{\mathrm{ADV}}]} L_{\mathrm{ADV}}(\hat{\boldsymbol{y}}^{(\mathrm{L})}) + \omega_{\mathrm{D}} \frac{E_{\hat{\boldsymbol{y}}}[L_{\mathrm{MSE}}]}{E_{\hat{\boldsymbol{y}}}[L_{\mathrm{ADV}}]} L_{\mathrm{ADV}}(\hat{\boldsymbol{y}})$$

4. EXPERIMENTAL EVALUATION

Experimental conditions

Dataset	4,007 utterances taken fromJapanese female speaker (subset of JSUT[5] corpus, 3,808/199 for training/evaluation)
STFT analysis	Frame length: 400, shift length: 80, FFT length: 1024, analysis window: Hamming
Average pooling	Zero padding: 6, pooling width w: {14, 30, 70}, stride: $w/2$
ω_{D} and $\omega_{\mathrm{D}}^{(\mathrm{L})}$	1.0
Dims. of $\boldsymbol{x}/\boldsymbol{y}/\boldsymbol{y}^{(\mathrm{L})}$	444 (linguistic feats, durations, Log F0, UV)/513/{74, 34, 14} (dim. of $\boldsymbol{y}^{(\mathrm{L})}$ was changed in accordance with w)
DNN architectures	Feed-Forward (see our paper)

Subjective evaluation (preference AB tests)

MSE: minimizing $L_{\mathrm{MSE}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$[2]

ADV: minimizing $L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$[4]

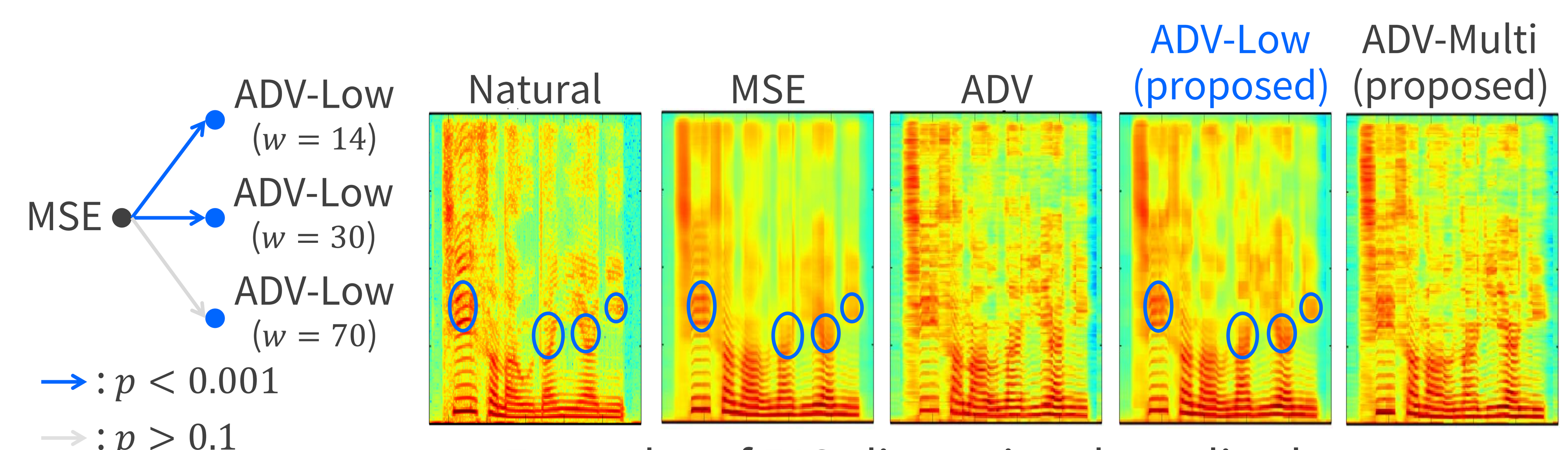
ADV-Low: minimizing $L_{\mathrm{G}}^{(\mathrm{Low})}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ (proposed)

ADV-Multi: minimizing $L_{\mathrm{G}}^{(\mathrm{Multi})}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ (proposed)

Results:

(1) ADV & ADV-Multi didn't improve speech quality.
⇒ Because of difficulty in minimizing distribution differences in original resolution

(2) ADV-Low improved speech quality.
⇒ Effect of the spectral envelopes compensation



Examples of 513-dimensional amplitude spectra

[References]
[1] Goodfellow et al., *Proc. NIPS*, 2014. [2] Takaki et al., *Proc. INTERSPEECH*, 2017. [3] Griffin et al., *IEEE Trans. on ASLP*, 1984. [4] Saito et al., *IEEE/ACM Trans. on ASLP*, 2018. [5] Sonobe et al., *arXiv:1711.00354*, 2017.