High-Quality Statistical Parametric Speech Synthesis Using Generative Adversarial Networks

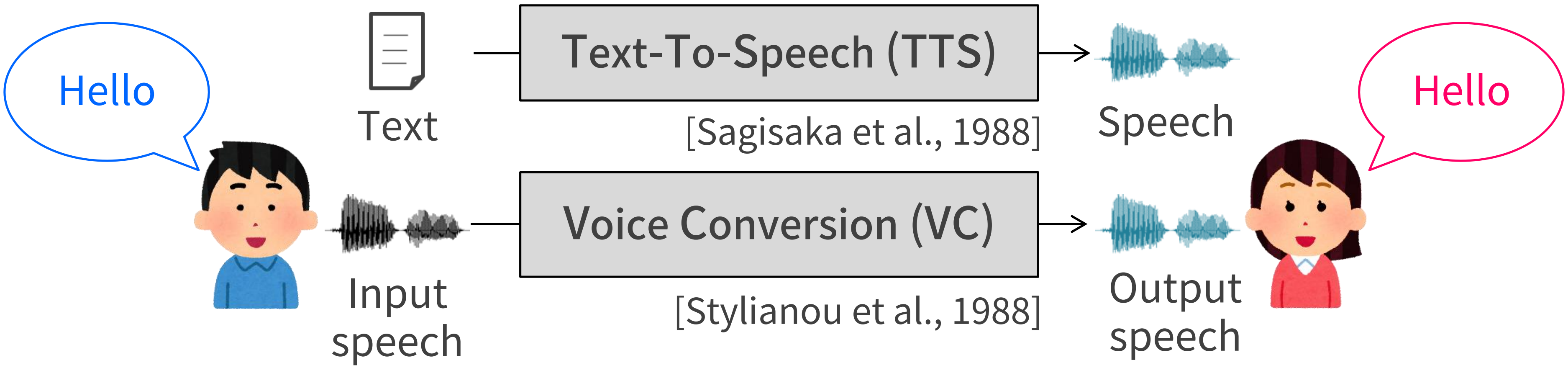
48-166605 Yuki Saito

Supervisor: Prof. Hiroshi Saruwatari

Research Field: Speech Synthesis

Speech Synthesis

Technique for synthesizing speech using computer



Applications

Speech communication assistance (e.g., speech translation)

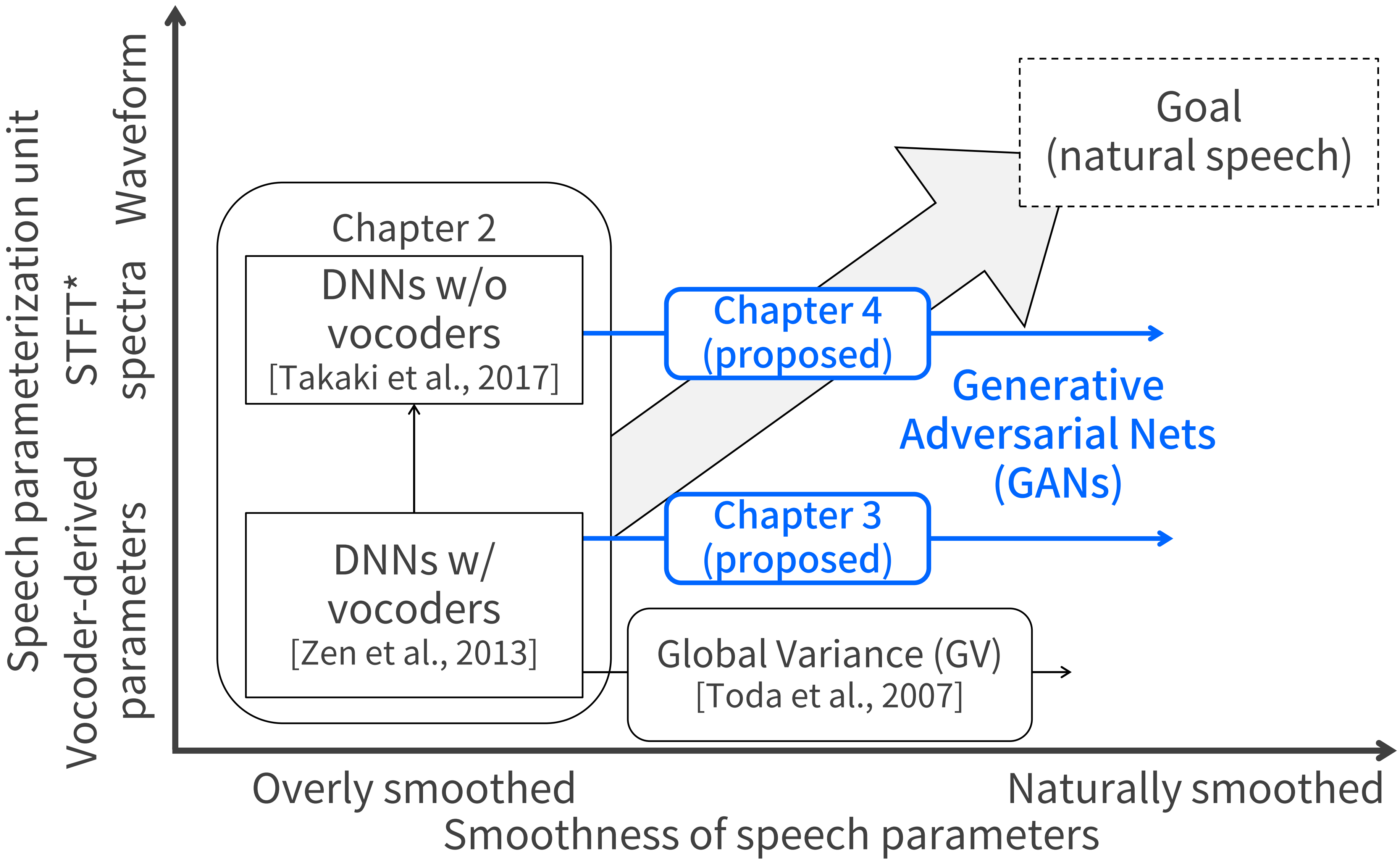
Entertainments (e.g., singing voice conversion)

DNN-based speech synthesis* [Zen et al. 2013]

High flexibility but low speech quality

DNN: Deep Neural Network

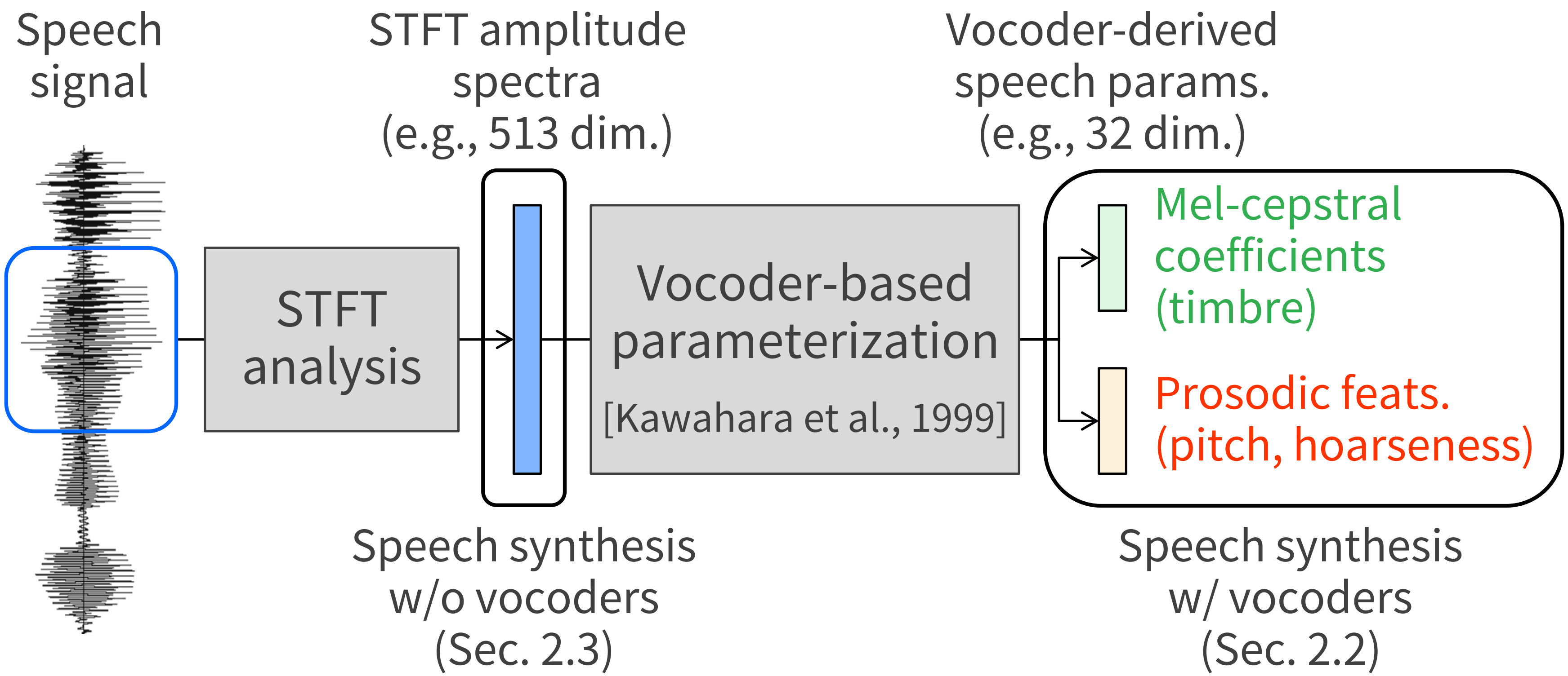
Thesis Overview



STFT: Short-Term Fourier Transform

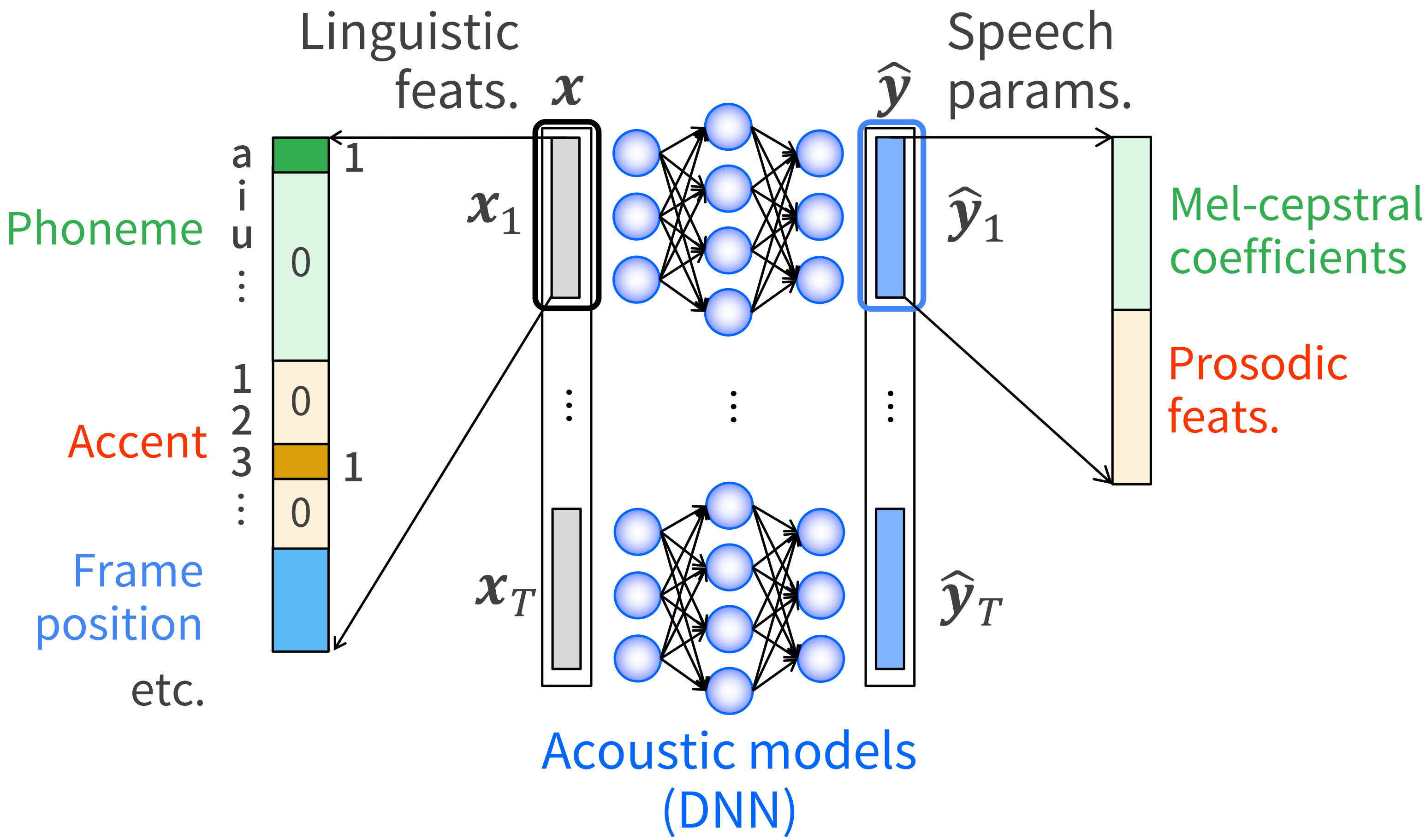
Table of Contents

Speech Analysis and Parameter Extraction



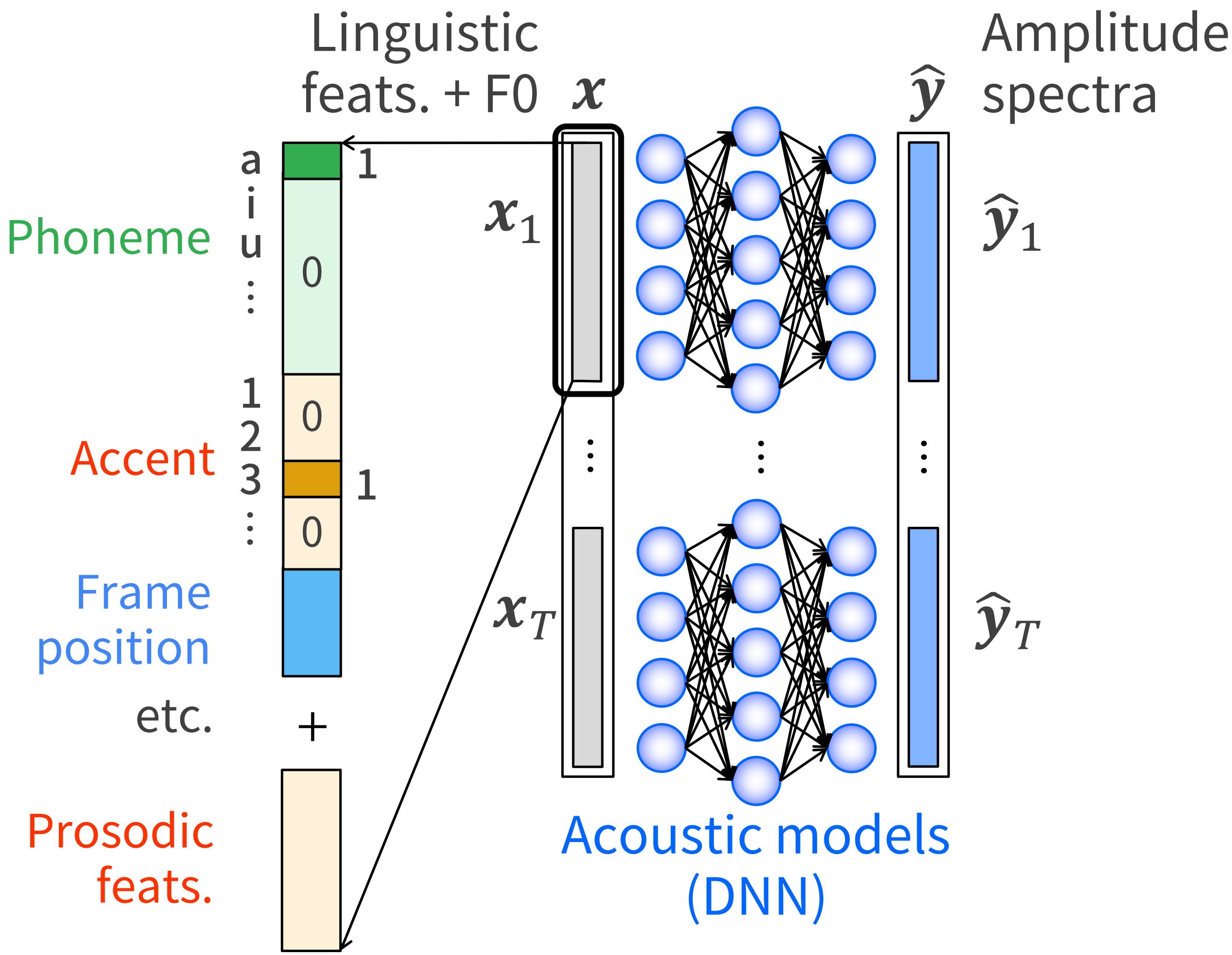
DNN-based Speech Synthesis w/ Vocoders

[Zen et al., 2013]



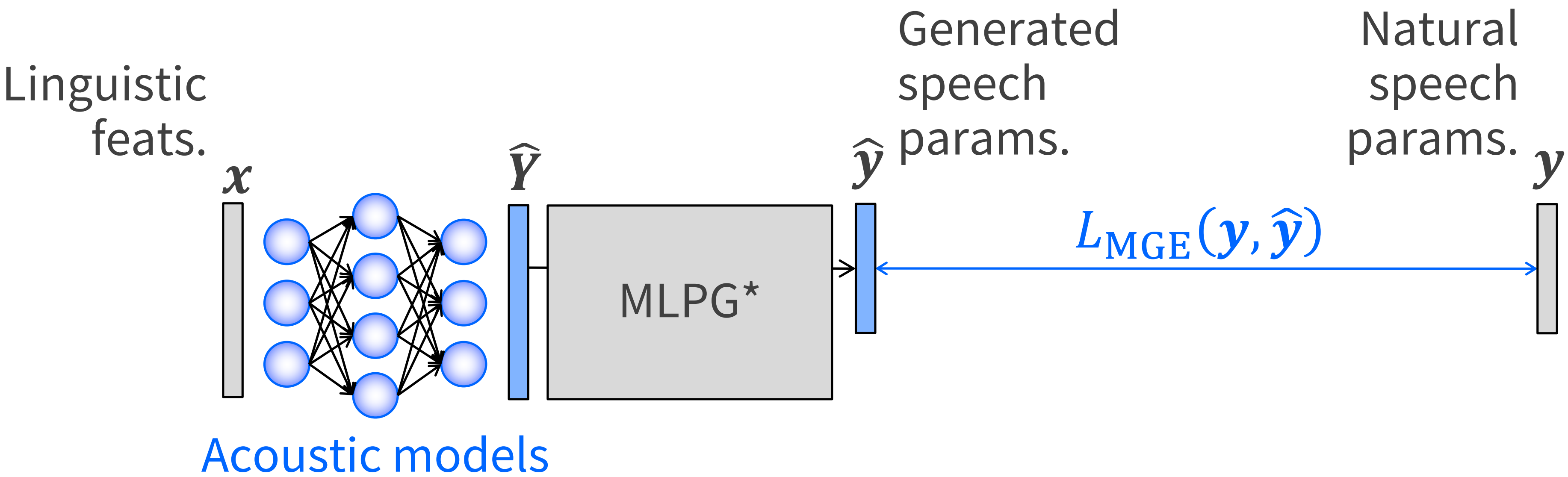
DNN-based Speech Synthesis w/o Vocoders

[Takaki et al., 2017]



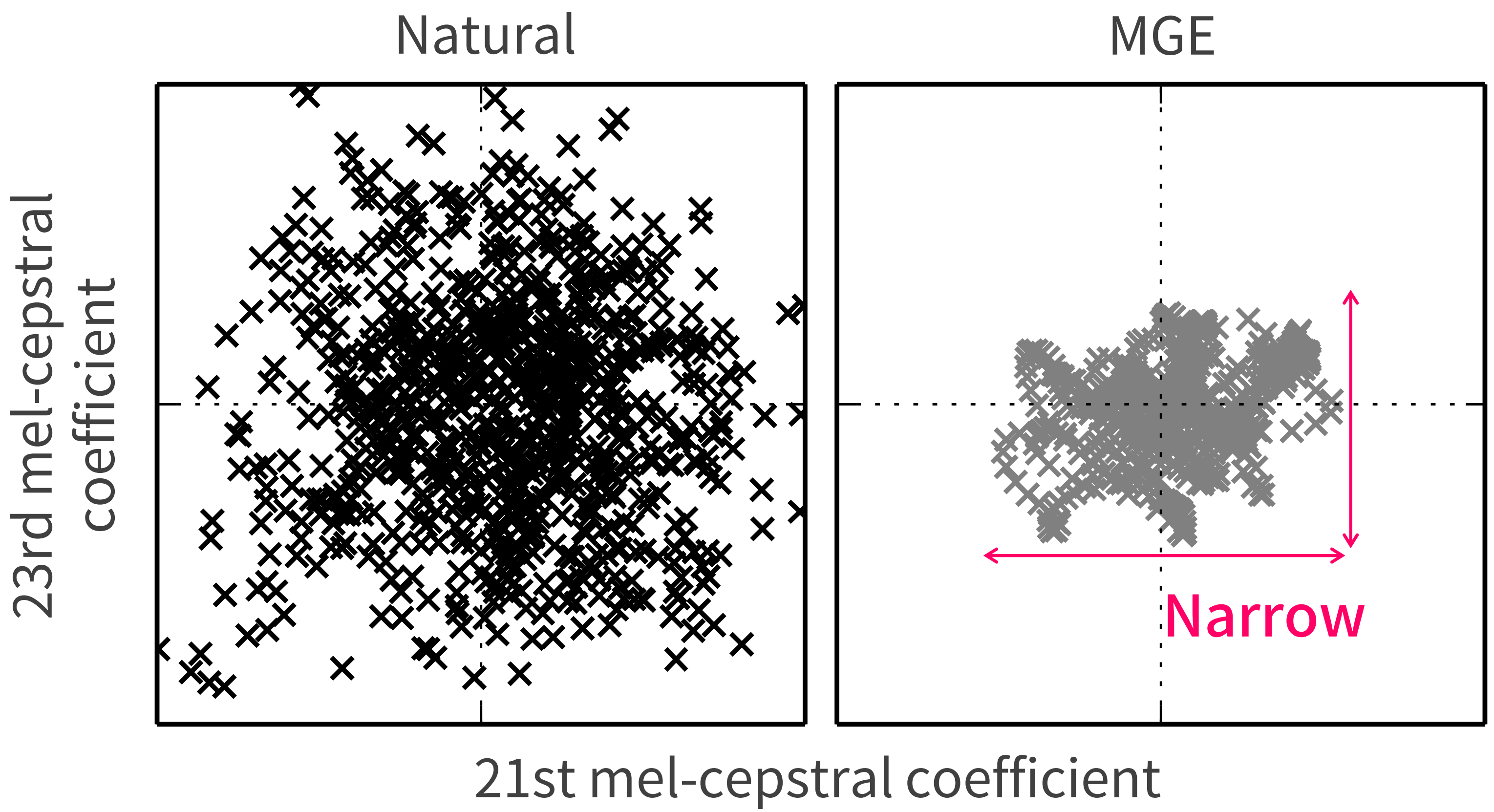
Minimum Generation Error (MGE) Training Algorithm

[Wu et al., 2016]



$$L_{\mathrm{MGE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{T}(\hat{\boldsymbol{y}} - \boldsymbol{y})^{\top}(\hat{\boldsymbol{y}} - \boldsymbol{y}) \rightarrow \text{Minimize}$$

Issue of DNN-based Speech Synthesis: Over-smoothing of Generated Speech Parameters

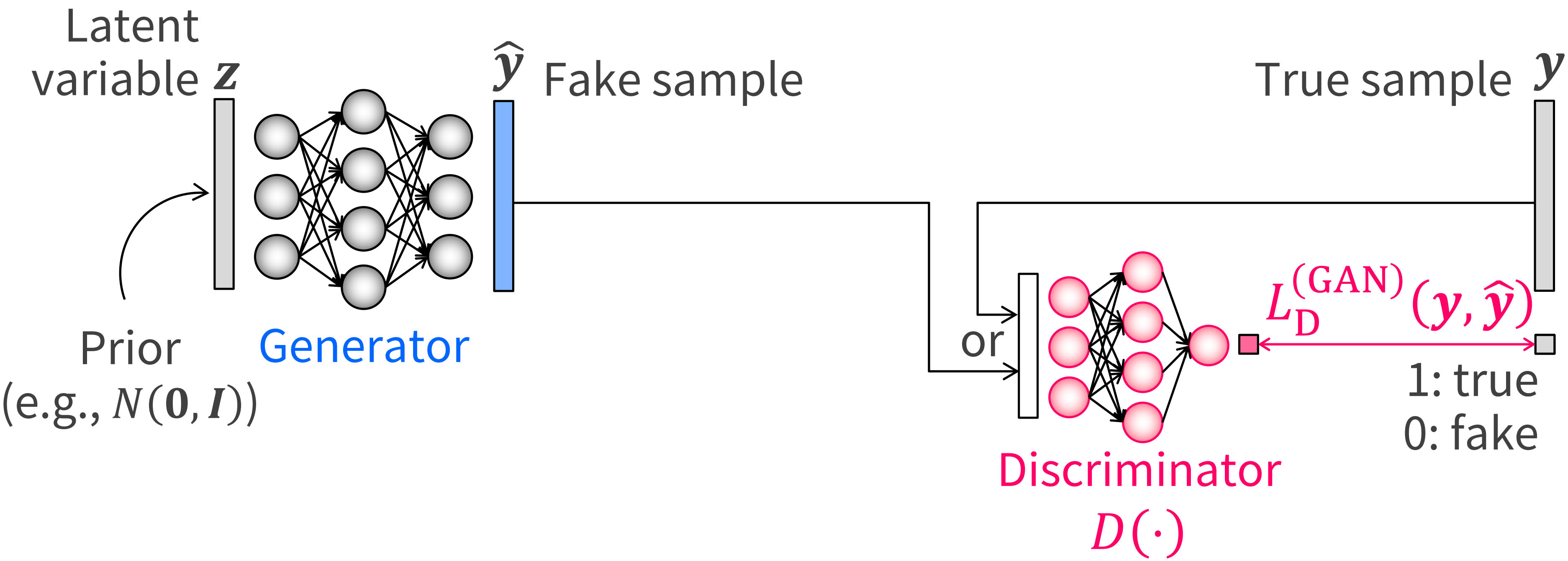


These distributions are significantly different...

(GV [Toda et al., 2007] explicitly compensates the 2nd moment.)

Table of Contents

Generative Adversarial Nets (GANs) [Goodfellow et al., 2014]



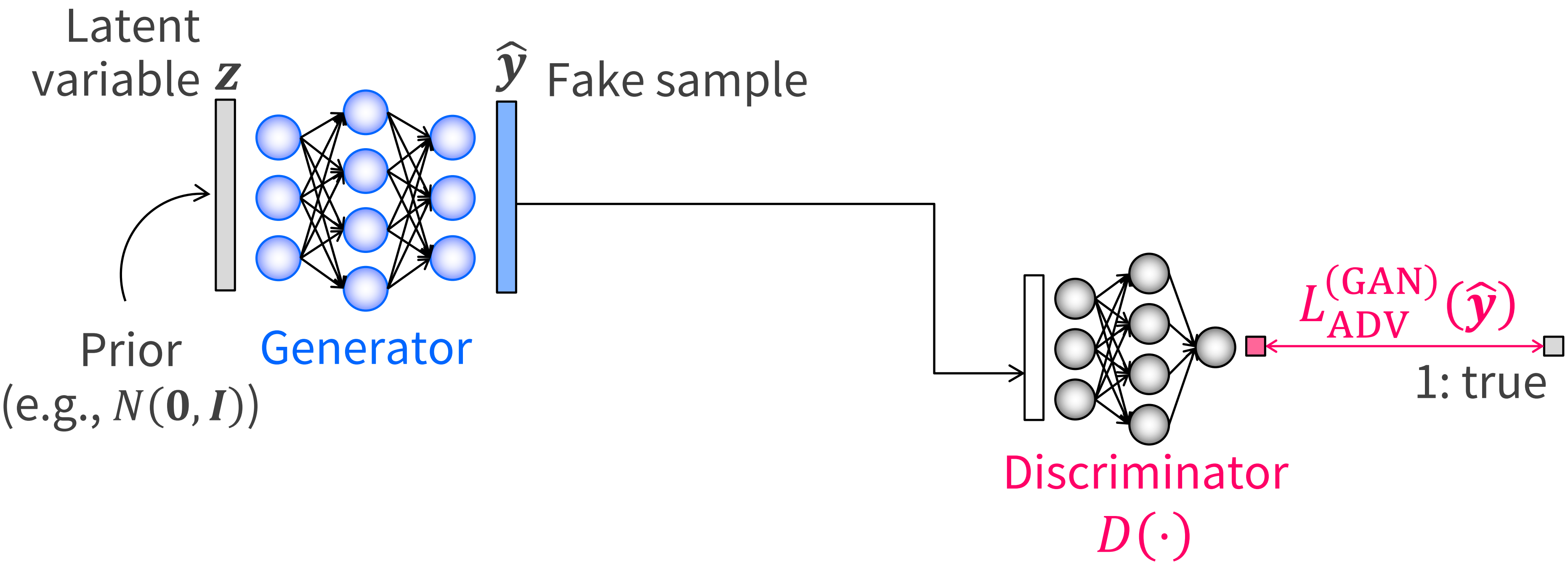
Loss to recognize true sample as true

Loss to recognize fake sample as fake

$$L_{\mathrm{D}}^{(\mathrm{GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\mathrm{D},1}^{(\mathrm{GAN})}(\boldsymbol{y}) + L_{\mathrm{D},0}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}}) \quad \rightarrow \text{Minimize}$$

$L_{\mathrm{D}}^{(\mathrm{GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ is equivalent to the cross-entropy function.

Generative Adversarial Nets (GANs) [Goodfellow et al., 2014]



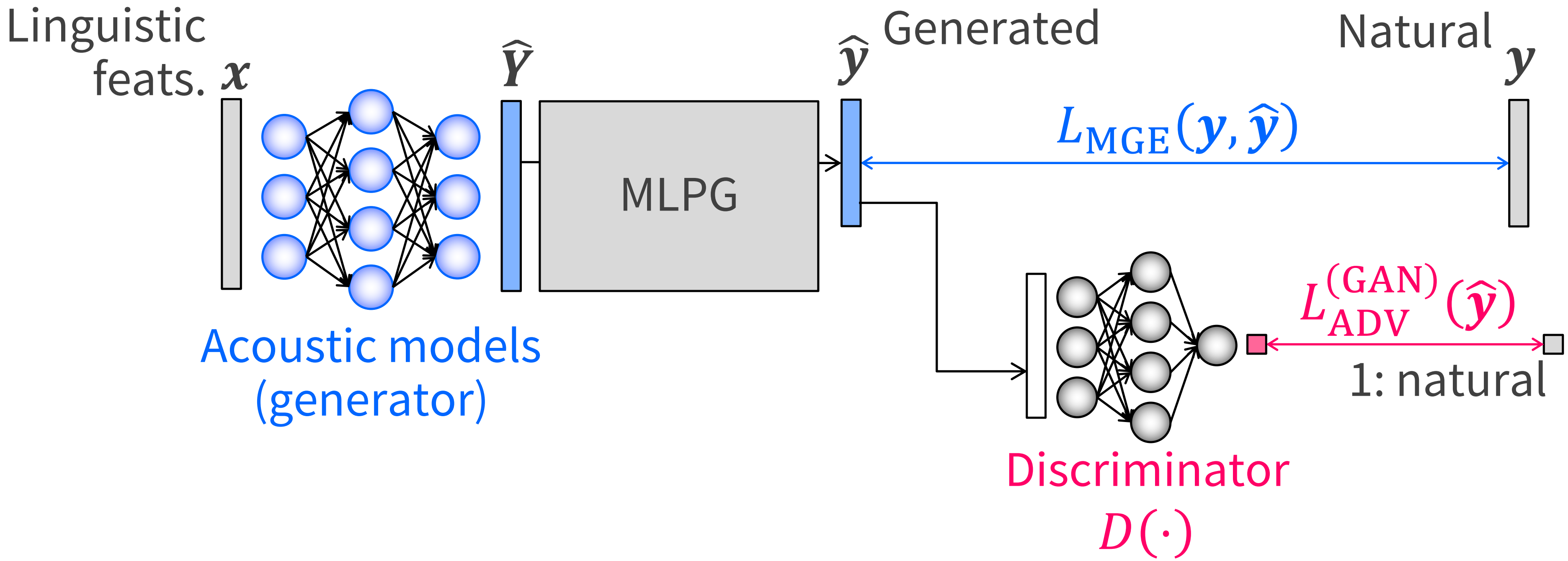
Loss to recognize fake sample as true

$$L_{\mathrm{ADV}}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}}) = L_{\mathrm{D},1}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}}) \rightarrow \text{Minimize}$$

Minimize approx. JS* divergence betw. dists. of $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$.

JS: Jensen-Shannon

Proposed Method: Acoustic Model Training Using GANs



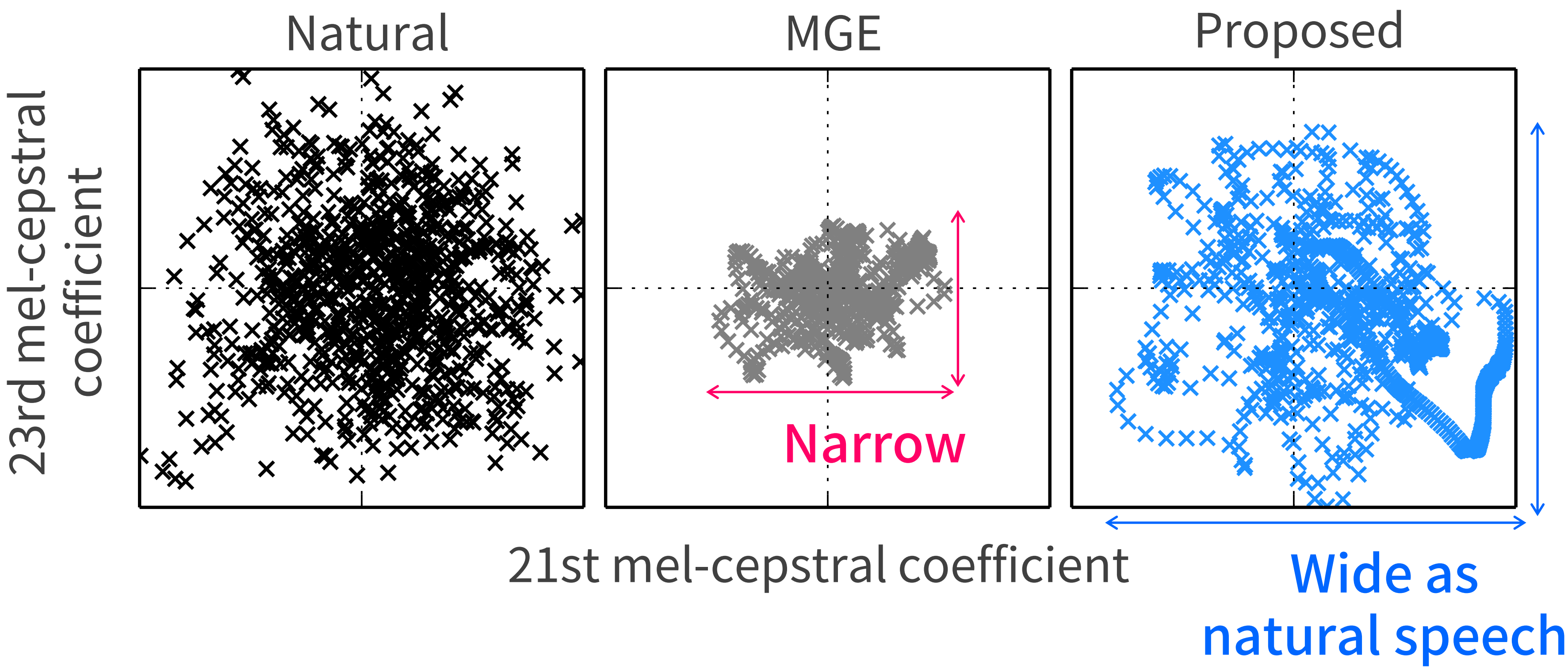
Loss to recognize generated params. as natural

$$L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\mathrm{MGE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \omega_{\mathrm{D}} \frac{E_{L_{\mathrm{MGE}}}}{E_{L_{\mathrm{ADV}}}} L_{\mathrm{ADV}}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}}) \rightarrow \text{Minimize}$$

ω_{D}: weight, E_{L_*}: expectation values of L_*

Distributions of Speech Parameters

GANs = minimizing divergence betw. two distributions



The proposed algorithm alleviates the over-smoothing effect!

Discussions

Compensating for distribution differences

The proposed method generalizes the conventional methods such as the GV.

Integrating voice anti-spoofing techniques

Features that are effective for detecting synthetic speech can be used (Sec. 3.4.8).

Changing a divergence to be minimized

Earth mover's distance (Wasserstein GAN [Arjovsky et al., 2017]) was the best for improving synthetic speech quality (Sec. 3.4.10).

Applying various speech synthesis

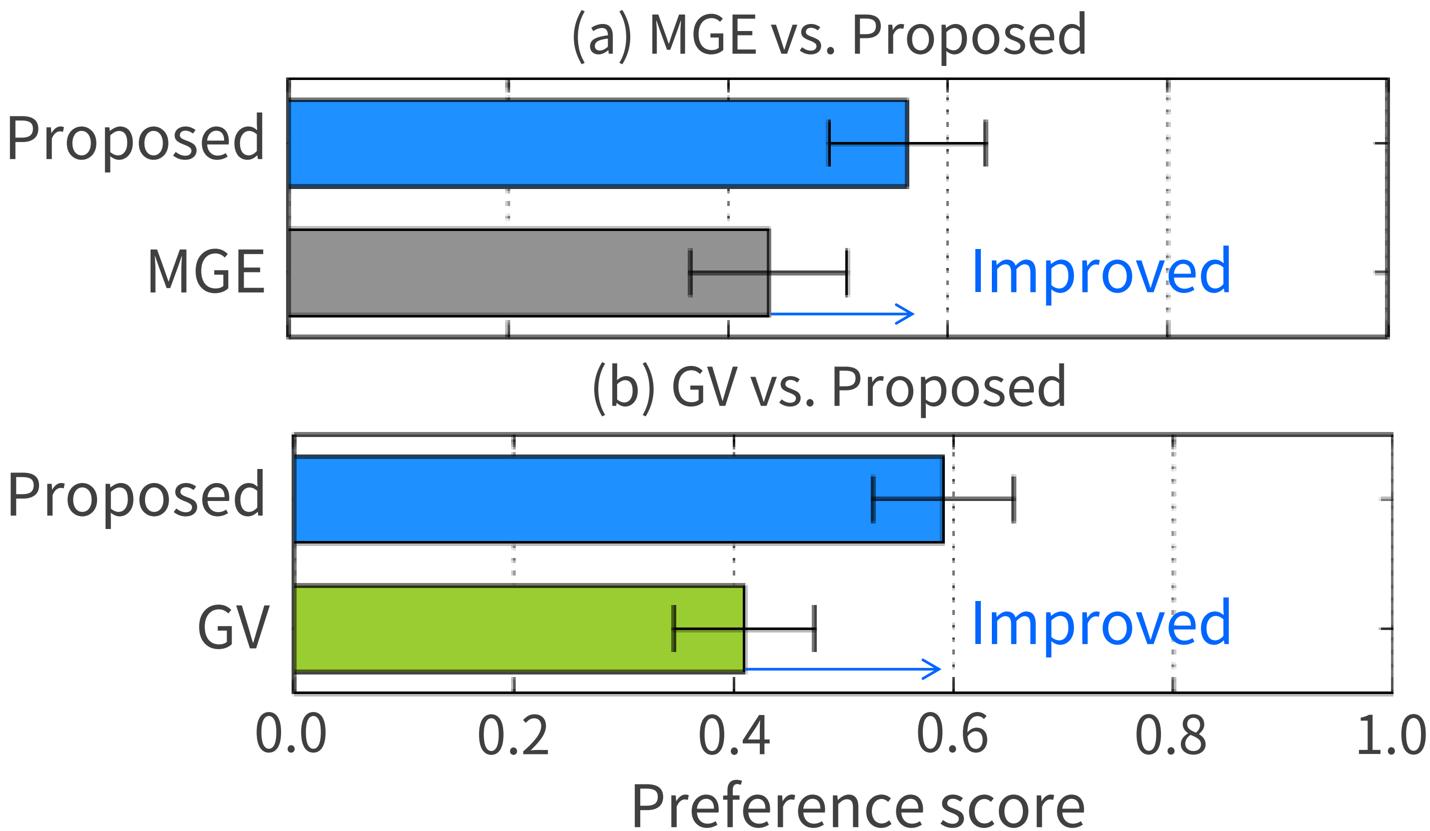
Not only TTS (next slides), but also VC (Sec. 3.5).

Experimental Conditions

Train / evaluate data	450 sentences / 53 sentences (16 kHz sampling)
Linguistic feats.	442-dimensional vector
Speech params.	Mel-cepstral coefficients and prosodic features
Optimizer	AdaGrad [Duchi et al., 2011]
Acoustic models	Feed-Forward 442 – 3x512 (ReLU) – 94 (linear)
Discriminator	Feed-Forward 26 – 3x256 (ReLU) – 1 (sigmoid)
Weight ω_{D}	1.0 (Secs. 3.4.2 and 3.4.4)
Methods	**MGE** [Wu et al., 2016], **GV** [Toda et al., 2007], **Proposed**

Subjective Evaluations in Terms of Speech Quality

Preference AB test (select better sounded speech)



MGE

GV

Proposed

Proposed method improves synthetic speech quality!

Error bars denote 95% confidence intervals.

Table of Contents

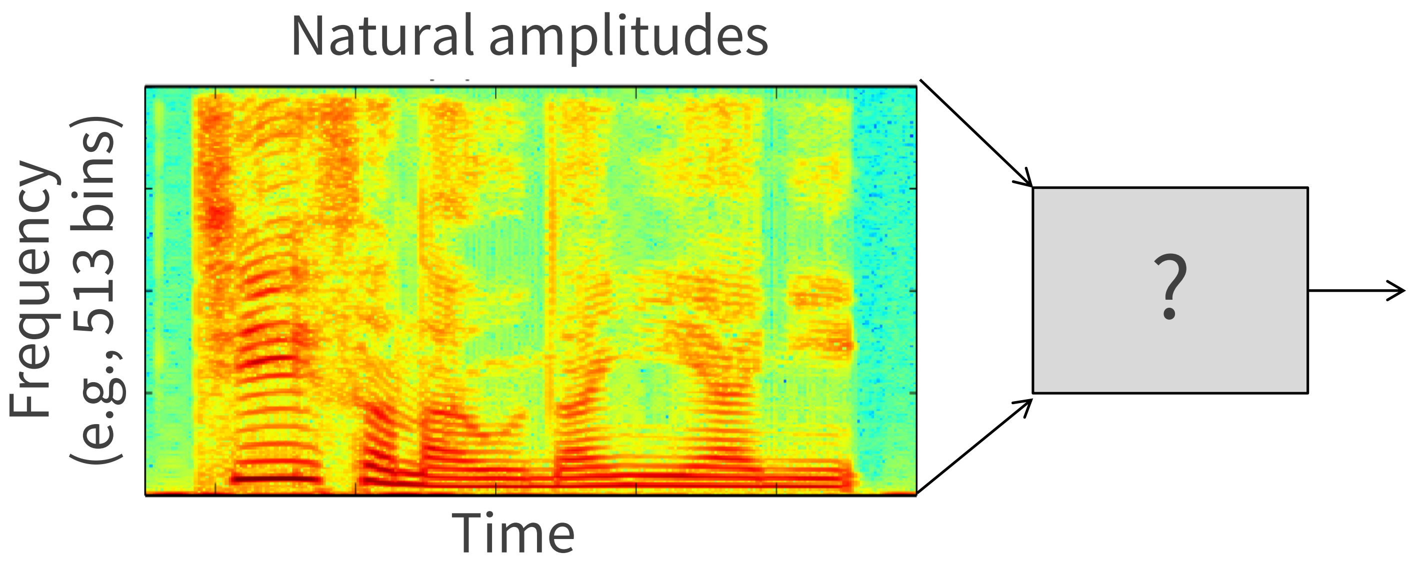
Issue in Speech Synthesis w/o Vocoders

Over-smoothing of generated STFT amplitude spectra

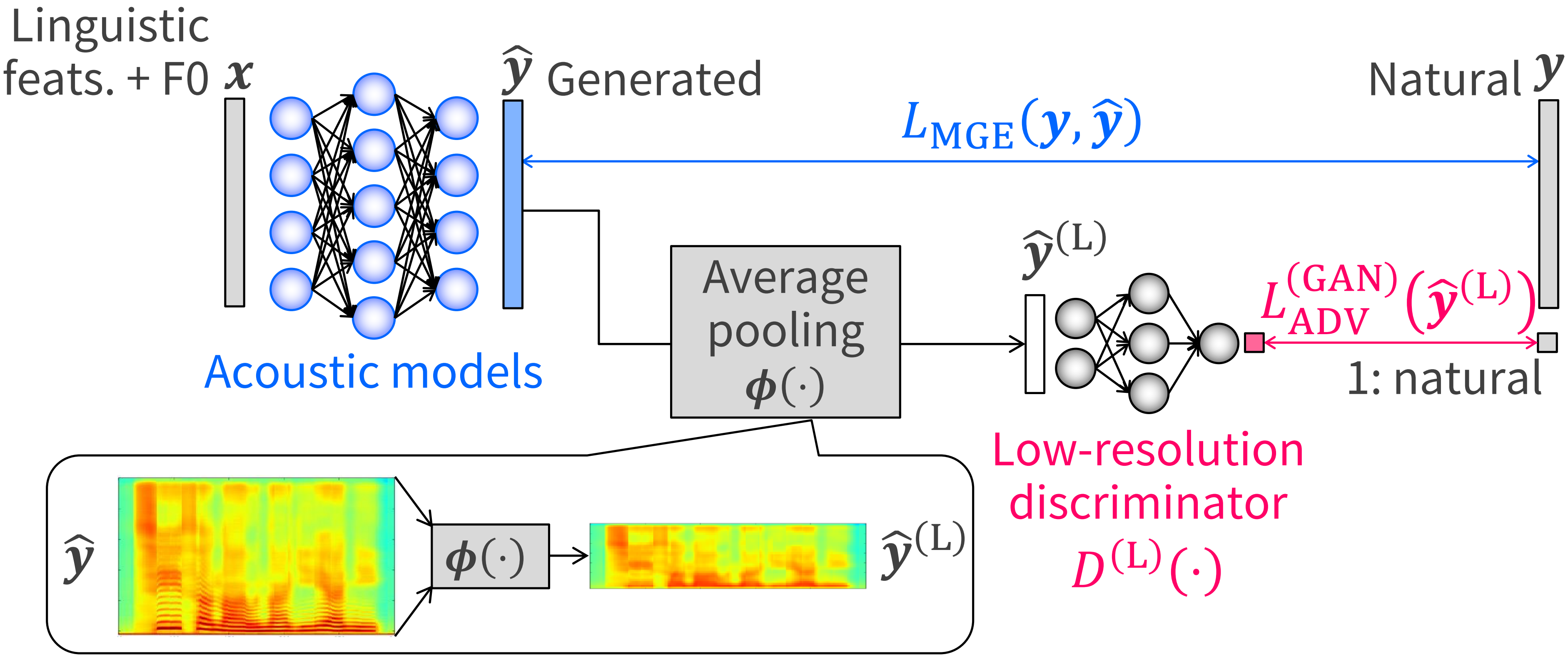
- Formants (spectral peaks) tend to be weakened.
- The method proposed in Chap. 3 cannot be applied directly.
  - **Difficulties in modeling highly complex distribution**

To deal with the issue...

- Dimensionality reduction retaining spectral structures

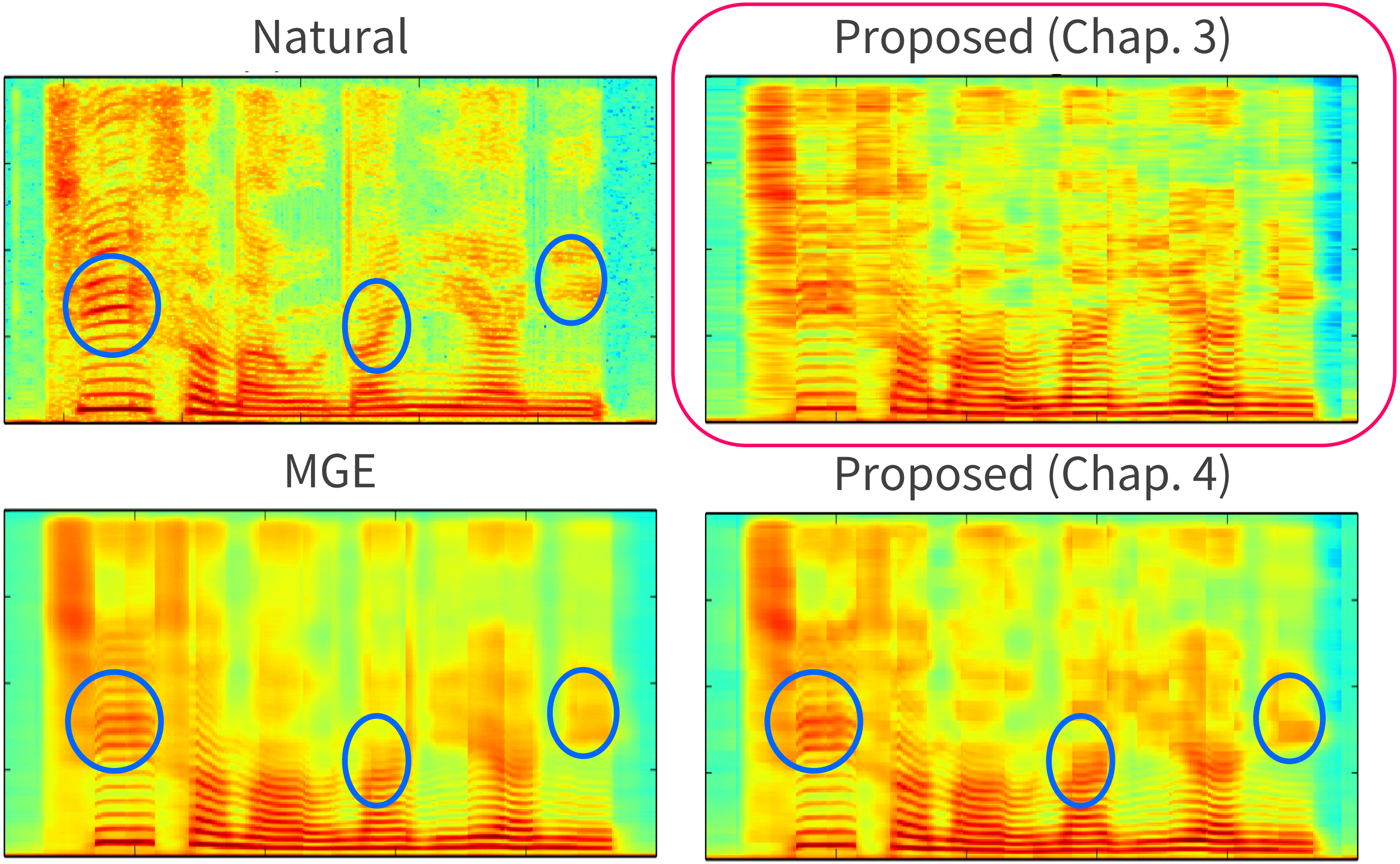


Acoustic Model Training Using Low-resolution GANs



$$L_{\mathrm{G}}^{(\mathrm{L})}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\mathrm{MGE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \omega_{\mathrm{D}}^{(\mathrm{L})} \frac{E_{L_{\mathrm{MGE}}}}{E_{L_{\mathrm{ADV}}}} L_{\mathrm{ADV}}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}}^{(\mathrm{L})}) \rightarrow \text{Minimize}$$

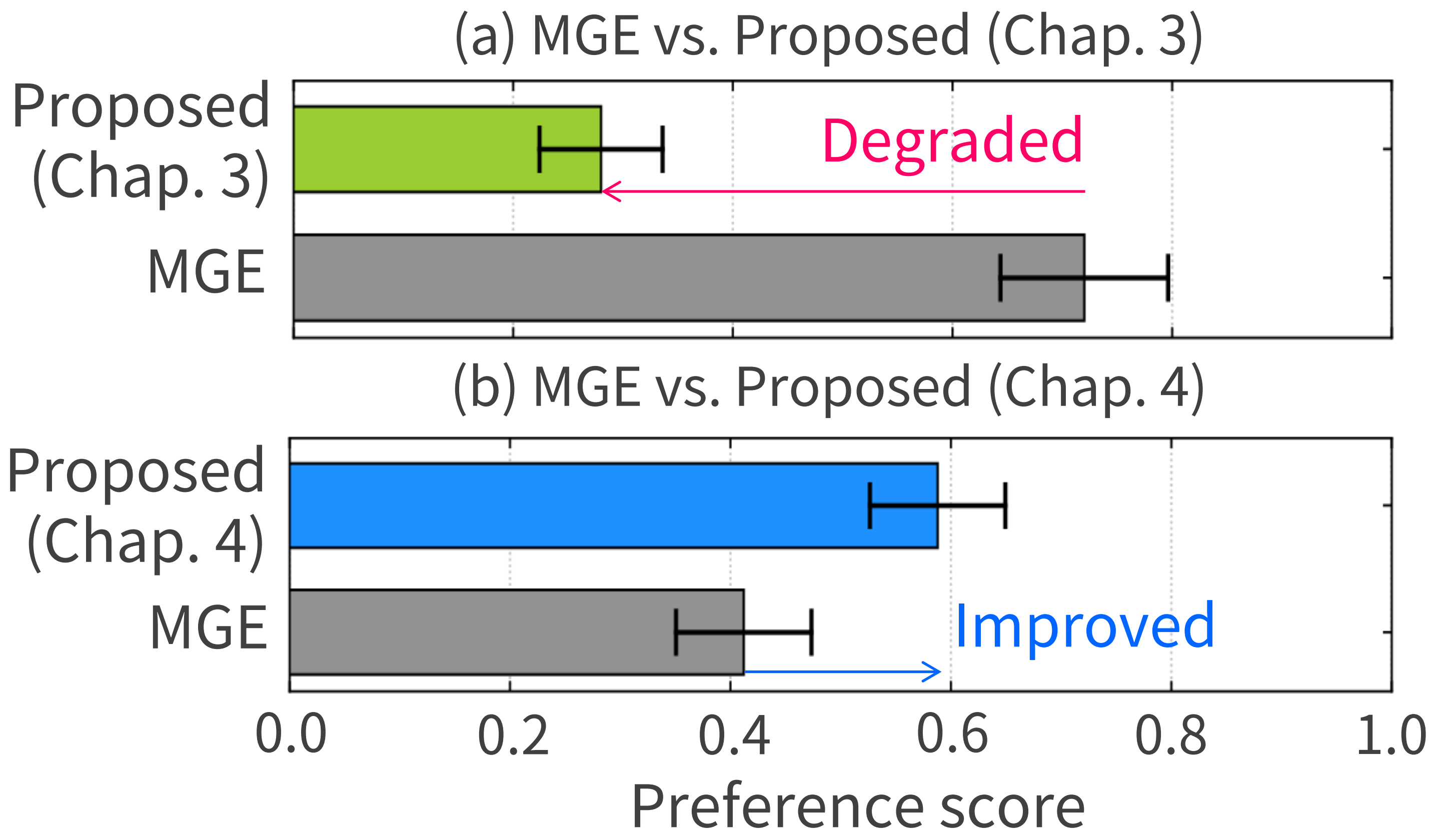
Examples of Natural and Generated Amplitude Spectra



Low-resolution GANs capture differences in formants!

Subjective Evaluations in Terms of Speech Quality

Preference AB test (select better sounded speech)



MGE

Proposed (Chap. 3)

Proposed (Chap. 4)

Low-resolution GAN improves synthetic speech quality!

Error bars denote 95% confidence intervals.

Table of Contents

Conclusion

Purpose: improving synthetic speech quality of SPSS

Proposed: acoustic model training algorithms using GANs

- They compensate for the distribution differences betw. natural / generated speech parameters.

Results

- The proposed algorithms improved synthetic speech quality compared to conventional methods.

Future works

- Investigating anti-spoofing techniques
- Further improving speech quality using STFT spectra