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
Speech parameterization unit
Vocoder-derived parameters
STFT* spectra
Waveform

Overly smoothed
Smoothness of speech parameters

Naturally smoothed
Smoothness of speech parameters

Global Variance (GV)
[Toda et al., 2007]

DNNs w/ vocoders
[Zen et al., 2013]

DNNs w/o vocoders
[Takaki et al., 2017]

Chapter 2

Chapter 3
(proposed)

Chapter 4
(proposed)

Generative Adversarial Nets (GANs)

Goal (natural speech)

Vocoder-derived parameters

STFT*: Short-Term Fourier Transform

Thesis Overview
Table of Contents

Chapter 1. Introduction

Chapter 2. Speech Synthesis Using DNNs

Chapter 3. Speech Synthesis Using GANs w/ Vocoders

Chapter 4. Speech Synthesis Using GANs w/o Vocoders

Chapter 5. Conclusion
Speech Analysis and Parameter Extraction

Speech signal

STFT analysis

STFT amplitude spectra (e.g., 513 dim.)

Vocoder-based parameterization

Vocoder-derived speech params. (e.g., 32 dim.)

Speech synthesis w/o vocoders (Sec. 2.3)

Speech synthesis w/ vocoders (Sec. 2.2)

Mel-cepstral coefficients (timbre)

Prosodic feats. (pitch, hoarseness)

[Kawahara et al., 1999]
DNN-based Speech Synthesis w/ Vocoders

Acoustic models (DNN)

Linguistic feats. \( x \)

Speech params.

Mel-cepstral coefficients

Prosodic feats.

\[ \begin{align*}
    y &\rightarrow \hat{y}_1 \\
    x_1 &\rightarrow \hat{y}_1 \\
    x_T &\rightarrow \hat{y}_T \\
\end{align*} \]

[Zen et al., 2013]

Phoneme

Accent

Frame position etc.
DNN-based Speech Synthesis w/o Vocoders

[677x11]/23
[667x13]6
[62x495]DNN
[124x495]−
[134x495]based Speech Synthesis w/o Vocoders

\[ x \cdots \]

Acoustic models (DNN)

\[ y \]

Prosodic feats.

Linguistic feats. + F0

Phoneme

Accent

Frame position etc.

Amplitude spectra

\[ \hat{y}, \hat{y}_1, \hat{y}_T \]

[Takeki et al., 2017]
Minimum Generation Error (MGE) Training Algorithm

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

MLPG: Maximum Likelihood Parameter Generation [Tokuda et al., 2000]
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

Chapter 1. Introduction
Chapter 2. Speech Synthesis Using DNNs
Chapter 3. Speech Synthesis Using GANs w/ Vocoders
Chapter 4. Speech Synthesis Using GANs w/o Vocoders
Chapter 5. Conclusion
Generative Adversarial Nets (GANS) [Goodfellow et al., 2014]

Latent variable $z$

Prior (e.g., $N(0, I)$)

Generator

$\hat{y}$ Fake sample

True sample $y$

Discriminator $D(\cdot)$

Loss to recognize true sample as true

Loss to recognize fake sample as fake

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

$L_D^{(GAN)}(y, \hat{y})$ is equivalent to the cross-entropy function.
Generative Adversarial Nets (GANS) [Goodfellow et al., 2014]

Latent variable $\mathbf{z}$

Prior (e.g., $\mathcal{N}(\mathbf{0}, I)$)

Generator

Fake sample

Loss to recognize fake sample as true

$L^{(\text{GAN})}_{\text{ADV}}(\hat{\mathbf{y}}) = L^{(\text{GAN})}_{D,1}(\hat{\mathbf{y}}) \rightarrow \text{Minimize}$

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

JS: Jensen-Shannon
Proposed Method: Acoustic Model Training Using GANs

**Linguistic feats.** $x$ → **Acoustic models (generator)** → **MLPG** → **Generated** $\hat{y}$ → **Discriminator** $D(\cdot)$ → **Natural** $y$

$\omega_D$: weight, $E_{L_*}$: expectation values of $L_*$

$$L_G(y, \hat{y}) = L_{MGE}(y, \hat{y}) + \omega_D \frac{E_{L_{MGE}}}{E_{L_{ADV}}} L_{(GAN)}^{(ADV)}(\hat{y}) \rightarrow \text{Minimize}$$
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

<table>
<thead>
<tr>
<th>Train / evaluate data</th>
<th>450 sentences / 53 sentences (16 kHz sampling)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic feats.</td>
<td>442-dimensional vector</td>
</tr>
<tr>
<td>Speech params.</td>
<td>Mel-cepstral coefficients and prosodic features</td>
</tr>
<tr>
<td>Optimizer</td>
<td>AdaGrad [Duchi et al., 2011]</td>
</tr>
<tr>
<td>Acoustic models</td>
<td>Feed-Forward 442 – 3x512 (ReLU) – 94 (linear)</td>
</tr>
<tr>
<td>Discriminator</td>
<td>Feed-Forward 26 – 3x256 (ReLU) – 1 (sigmoid)</td>
</tr>
<tr>
<td>Weight $\omega_D$</td>
<td>1.0 (Secs. 3.4.2 and 3.4.4)</td>
</tr>
<tr>
<td>Methods</td>
<td>MGE [Wu et al., 2016], GV [Toda et al., 2007], Proposed</td>
</tr>
</tbody>
</table>
Subjective Evaluations in Terms of Speech Quality

Preference AB test (select better sounded speech)

(a) MGE vs. Proposed

(b) GV vs. Proposed

Proposed method improves synthetic speech quality!

Error bars denote 95% confidence intervals.
Table of Contents

Chapter 1. Introduction
Chapter 2. Speech Synthesis Using DNNs
Chapter 3. Speech Synthesis Using GANs w/ Vocoders
Chapter 4. Speech Synthesis Using GANs w/o Vocoders
Chapter 5. Conclusion
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

Natural amplitudes

Frequency (e.g., 513 bins)

Time

?
Acoustic Model Training Using
Low-resolution GANs

Linguistic feats. + F0 $\mathbf{x}$

Acoustic models

$\mathbf{y}$ Generated

Natural $\mathbf{y}$

$L_{MGE}(\mathbf{y}, \hat{\mathbf{y}})$

Low-resolution discriminator $D^{(L)}(\cdot)$

$\hat{\mathbf{y}}^{(L)}$

$\hat{\mathbf{y}}$

Average pooling $\phi(\cdot)$

$\hat{\mathbf{y}}^{(L)}$

$\phi(\cdot)$

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

(a) MGE vs. Proposed (Chap. 3)

Proposed (Chap. 3)  MGE

Degraded

(b) MGE vs. Proposed (Chap. 4)

Proposed (Chap. 4)  MGE

Improved

Low-resolution GAN improves synthetic speech quality!

Error bars denote 95% confidence intervals.
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Introduction</td>
</tr>
<tr>
<td>2.</td>
<td>Speech Synthesis Using DNNs</td>
</tr>
<tr>
<td>3.</td>
<td>Speech Synthesis Using GANs w/ Vocoders</td>
</tr>
<tr>
<td>4.</td>
<td>Speech Synthesis Using GANs w/o Vocoders</td>
</tr>
<tr>
<td>5.</td>
<td>Conclusion</td>
</tr>
</tbody>
</table>
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