Investigating on Incorporating Pretrained and Learnable Speaker Representations for Multi-Speaker Multi-Style Text-to-Speech

Chung-Ming Chien, Jheng-Hao Lin*, Chien-yu Huang*, Po-chun Hsu* and Hung-yi Lee

College of Electrical Engineering and Computer Science, National Taiwan University

* These authors contributed equally.
Outline

- Task Description
- Background & Motivation
- Methodology
- Experiments
- Conclusion
Task Description
Multi-Speaker Multi-Style Voice Cloning

Task Description

Computer talks.

Few-Shot References

Multi-Speaker TTS System

Text

Synthesized Speech
Multi-Speaker Multi-Style Voice Cloning

Task Description

Computer talks.

Few-Shot References

Multi-Speaker TTS System

Text

Synthesized Speech

Similar in speaker & style
Multi-Speaker Multi-Style Voice Cloning

Task Description

Challenge

- Extract speaker and style information from limited references
- Enable the TTS system to generalize to different speakers/styles
Background & Motivation
General Framework of Multi-Speaker TTS

Text → Speaker Representation → TTS System → Synthesized Speech

Computer talks.
General Framework of Multi-Speaker TTS

**Learnable**
- Embedding Table
- Trainable Speaker Encoder

**Pretrained**
- Pretrained Speaker Encoder

---

Computer talks. \(\xrightarrow{\text{Text}}\) TTS System \(\xrightarrow{\text{Speaker Representation}}\) Synthesized Speech
General Framework of Multi-Speaker TTS

Learnable Speaker Representation

Speaker ID

Embedding Table

Speaker Representation

Computer talks.

Text

TTS System

Synthesized Speech

“Deep voice 3: Scaling text-to-speech with convolutional sequence learning”, Ping, et. al, ICLR’18
General Framework of Multi-Speaker TTS

Learnable Speaker Representation

Speaker ID

Embedding Table

Cannot generalize to unseen speakers

Speaker Representation

Computer talks.

TTS System

Text

Synthesized Speech

“Deep voice 3: Scaling text-to-speech with convolutional sequence learning”, Ping, et. al, ICLR’18
General Framework of Multi-Speaker TTS

Learnable Speaker Representation

Reference

Trainable Speaker Encoder

Speaker Representation

Computer talks.

TTS System

Synthesized Speech

“Neural voice cloning with a few samples”, Arik, et. al, NeurIPS’18

“Sample efficient adaptive text-to-speech”, Chen, et. al, ICLR’19
General Framework of Multi-Speaker TTS

Learnable Speaker Representation

- End-to-end optimized
- Arbitrary speaker

“Neural voice cloning with a few samples”, Arik, et. al, NeurIPS’18
“Sample efficient adaptive text-to-speech”, Chen, et. al, ICLR’19
General Framework of Multi-Speaker TTS

Pretrained Speaker Representation

Pretraining Task

Input → Speaker Encoder → Output
General Framework of Multi-Speaker TTS

Pretrained Speaker Representation

"Transfer learning from speaker verification to multi-speaker text-to-speech synthesis", Jia, et. al, NeurIPS’18
"Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings", Cooper, et. al, ICASSP’20
General Framework of Multi-Speaker TTS

Pretrained Speaker Representation

```
<table>
<thead>
<tr>
<th>Reference</th>
<th>Pretrained Speaker Encoder</th>
<th>Speaker Representation</th>
</tr>
</thead>
</table>

Fixed during training the TTS system
```

- D-vector
- X-vector
- ....

Text → TTS System → Synthesized Speech

“Transfer learning from speaker verification to multi-speaker text-to-speech synthesis”, Jia, et. al, NeurIPS’18

“Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings”, Cooper, et. al, ICASSP’20
Background & Motivation

General Framework of Multi-Speaker TTS

Pretrained Speaker Representation

- D-vector
- X-vector
- ....

Pretrained on speaker classification!

Reference

Pretrained Speaker Encoder

Fixed during training the TTS system

Speaker Representation

Text

Computer talks.

TTS System

Synthesized Speech

“Transfer learning from speaker verification to multi-speaker text-to-speech synthesis”, Jia, et. al, NeurIPS’18

“Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings”, Cooper, et. al, ICASSP’20
Motivation: Combining Different Representations

**Learnable**
- Embedding Table
- Trainable Speaker Encoder

**Pretrained**
- Pretrained Speaker Encoder

---

Computer talks. \[\rightarrow\] TTS System \[\rightarrow\] Synthesized Speech

Text

Speaker Representation
Motivation: Combining Different Representations

- **Learnable**
  - Embedding Table
  - Trainable Speaker Encoder

- **Pretrained**
  - Pretrained Speaker Encoder

Computer talks. → Speaker Representations → TTS System → Synthesized Speech

Text
Motivation: Different Pretraining Tasks

- D-vector
- X-vector
- ....

*Discriminative Pretraining Tasks*

*e.g. speaker classification*
Motivation: Different Pretraining Tasks

- D-vector
- X-vector
- ....

Discriminative Pretraining Tasks
  e.g. speaker classification

VS

Generative Pretraining Tasks?
Methodology
Workflow

Speaker Representation Pretraining ➔ TTS Training ➔ TTS Inference

Methodology
Speaker Representation Pretraining

Discriminative Tasks: D-vec & X-vec

Real Speech → Neural Network → Speaker Classification Loss

“Generalized end-to-end loss for speaker verification”, Wan, et. al, ICASSP’18

“X-vectors: Robust dnn embeddings for speaker recognition”, Snyder, et. al, ICASSP’18
Speaker Representation Pretraining

**Discriminative Tasks: D-vec & X-vec**

“Generalized end-to-end loss for speaker verification”, Wan, et. al, ICASSP’18

“X-vectors: Robust dnn embeddings for speaker recognition”, Snyder, et. al, ICASSP’18
Speaker Representation Pretraining

**Generative Tasks: AdaIN-VC (One-Shot)**

“One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization”,

Chou, et. al, InterSpeech’19
Speaker Representation Pretraining

**Generative Tasks: AdaIN-VC (One-Shot)**

“One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization”,

Chou, et. al, InterSpeech’19
TTS Training

- Computer talks.
- Text

Speaker Representations

- Tacotron 2
- FastSpeech 2

TTS System

Synthesized Speech

Optional Path
Necessary Path
TTS Training

Ground-Truth

Pretrained Speaker Encoder

Fixed during this stage

Speaker Representations

• D-vector
• X-vector
• AdaIN-VC

TTS System

Computer talks.

Text

Synthesized Speech

Methodology

Optional Path

Necessary Path
TTS Training

Ground-Truth

Pretrained Speaker Encoder

Fixed during this stage

Trainable Speaker Encoder

Speaker Representations

Global-Style Token (GST)

Optional Path

Necessary Path

Text

Computer talks.

TTS System

Synthesized Speech

"Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis", Wang, et. al, ICML’18
TTS Training

Ground-Truth

Pretrained Speaker Encoder

Trainable Speaker Encoder

Speaker Representations

Speaker ID

Embedding Table

Computer talks.

Text

TTS System

Synthesized Speech

Methodology

Optional Path

Necessary Path

Fixed during this stage
TTS Inference

Target Speaker Speech

Pretrained Speaker Encoder

Trainable Speaker Encoder

Speaker Representations

Target Speaker ID

Embedding Table

Text

Computer talks.

TTS System

Synthesized Speech
Experiments
Dataset

• Training: 96 hours of Mandarin speech by 230 speakers with transcriptions
  • AIShell-3
  • M2VoC dataset
Dataset

• Training: 96 hours of Mandarin speech by 230 speakers with transcriptions
  • AIShell-3
  • M2VoC dataset

• 6 few-shot target speakers
  • Track 1: 3 speakers with 100 recordings
  • Track 2: 3 speakers with 5 recordings
Dataset

• Training: 96 hours of Mandarin speech by 230 speakers with transcriptions
  • AIShell-3
  • M2VoC dataset
• 6 few-shot target speakers
  • Track 1: 3 speakers with 100 recordings
  • Track 2: 3 speakers with 5 recordings
• The few shot speakers are also used to train the speaker representation models and the TTS models
TTS Model Setup

- Tacotron 2 & FastSpeech 2
  - Speaker representations are added to encoder outputs

- WaveNet vocoder

![Diagram of TTS Model Setup](image-url)
Automatic Speaker Similarity Evaluation

Experiments

- Synthesized Speech
- Real Target Speaker Speech

Speaker Verification

Yes or No
Automatic Speaker Similarity Evaluation

**Metrics**

- **Speaker Verification Accuracy**
  - Scale: 0 ~ 1, the larger the better

Experiments

- Synthesized Speech
- Real Target Speaker Speech
- Speaker Verification
- Yes or No
# Automatic Speaker Similarity Evaluation

## Metrics

**Speaker Verification Accuracy**

*Scale: 0 ~ 1, the larger the better*

<table>
<thead>
<tr>
<th>Model</th>
<th>Speaker Representation</th>
<th>Results</th>
<th>Generative Pretraining &gt; Others</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pretrained</td>
<td>Learnable</td>
<td>SV Accuracy</td>
</tr>
<tr>
<td></td>
<td>d-vec</td>
<td>x-vec</td>
<td>VC</td>
</tr>
<tr>
<td>(a) Tacotron 2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b) FastSpeech2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Automatic Speaker Similarity Evaluation

## Metrics

**Speaker Verification Accuracy**

*Scale: 0 ~ 1, the larger the better*

<table>
<thead>
<tr>
<th>Model</th>
<th>Speaker Representation</th>
<th>Results SV Accuracy</th>
<th>Audio samples (Track 2, 5 references)</th>
<th>Target Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pretrained</td>
<td>Learnable</td>
<td>Track 1</td>
<td>Track 2</td>
</tr>
<tr>
<td>(a) Tacotron 2</td>
<td>✓</td>
<td>✓</td>
<td>.772</td>
<td>.367</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>.785</td>
<td>.377</td>
</tr>
<tr>
<td>(b) FastSpeech2</td>
<td>✓</td>
<td>✓</td>
<td>.630</td>
<td>.703</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>.102</td>
<td>.050</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>.977</td>
<td>.323</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>.980</td>
<td>.837</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>.978</td>
<td>.340</td>
</tr>
</tbody>
</table>
Automatic Speaker Similarity Evaluation

**Metrics**

- Speaker Verification Accuracy
  
  *Scale: 0 ~ 1, the larger the better*

![Experiments]

(a) d-vector  
(b) x-vector  
(c) VC
Automatic Speaker Similarity Evaluation

**Metrics**

- **Speaker Verification Accuracy**
  
  *Scale: 0 ~ 1, the larger the better*

Experiments

(a) d-vector

(b) x-vector

(c) VC
**Automatic Speaker Similarity Evaluation**

### Metrics

<table>
<thead>
<tr>
<th>Model</th>
<th>Speaker Representation</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pretrained</td>
<td>Learnable</td>
</tr>
<tr>
<td></td>
<td>d-vec</td>
<td>x-vec</td>
</tr>
<tr>
<td>(b) FastSpeech2</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(c) FastSpeech2</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

*The colored row is the model used for the final submission to the ICASSP 2021 M2VoC challenge. Due to the time limitation, we did not submit our best model.*

**Experiments**

- **Track 1 (100 references):**
  - No obvious difference
## Automatic Speaker Similarity Evaluation

### Metrics

<table>
<thead>
<tr>
<th>Model</th>
<th>Speaker Representation</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pretrained</td>
<td>Learnable</td>
</tr>
<tr>
<td></td>
<td>d-vec</td>
<td>x-vec</td>
</tr>
<tr>
<td>(\text{(b) FastSpeech2})</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(\text{(c) FastSpeech2})</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

*The colored row is the model used for the final submission to the ICASSP 2021 M2VoC challenge. Due to the time limitation, we did not submit our best model.*

### Experiments

**Track 2 (5 references):**

- **Multiple Representations > Single Representation**

**Track 1 (100 references):**

- **No obvious difference**
Subjective Evaluation (FastSpeech 2, Track 2)

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Quality MOS</th>
<th>Speaker Similarity MOS</th>
</tr>
</thead>
</table>

*Scale: 1 ~ 5, the larger the better*
# Subjective Evaluation (FastSpeech 2, Track 2)

## Metrics

<table>
<thead>
<tr>
<th></th>
<th>Quality MOS</th>
<th>Speaker Similarity MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scale:</strong></td>
<td>1 ~ 5, the larger the better</td>
<td></td>
</tr>
</tbody>
</table>

## Experiments

<table>
<thead>
<tr>
<th>Model</th>
<th>x-vec</th>
<th>VC</th>
<th>Embed</th>
<th>VC+Embed</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOSquality</td>
<td>3.47 ± 0.13</td>
<td>3.61 ± 0.13</td>
<td><strong>3.65 ± 0.13</strong></td>
<td>3.55 ± 0.12</td>
</tr>
<tr>
<td>MOSsimilarity</td>
<td><strong>3.25 ± 0.13</strong></td>
<td>3.19 ± 0.14</td>
<td>3.27 ± 0.13</td>
<td><strong>3.38 ± 0.14</strong></td>
</tr>
</tbody>
</table>

*Speaker Similarity: Multiple Representations > Single Representation*
Subjective Evaluation (FastSpeech 2, Track 2)

**Metrics**

<table>
<thead>
<tr>
<th>Quality MOS</th>
<th>Speaker Similarity MOS</th>
</tr>
</thead>
</table>

*Scale: 1 ~ 5, the larger the better*

<table>
<thead>
<tr>
<th>Model</th>
<th>Speaker Representation</th>
<th>x-vec</th>
<th>VC</th>
<th>Embed</th>
<th>VC+Embed</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOSquality</td>
<td>3.47 ± .13</td>
<td>3.61 ± .13</td>
<td><strong>3.65 ± .13</strong></td>
<td>3.55 ± .12</td>
<td></td>
</tr>
<tr>
<td>MOSsimilarity</td>
<td>3.25 ± .13</td>
<td>3.19 ± .14</td>
<td>3.27 ± .13</td>
<td><strong>3.38 ± .14</strong></td>
<td></td>
</tr>
</tbody>
</table>

Audio samples (Track 2, 5 references)

Target Speaker 🎧 VC 🎧 VC+Embed 🎧
Official Evaluation Results

Experiments

No External Data

Track 2A

External Data Allowed
(but we did not use)

Track 2B

MOS similarity

MOS quality

MOS similarity

MOS quality

Fig. 3: The official subjective evaluation results of Track 2.
Conclusion
Conclusion

• Pretrained speaker representation + learnable speaker representations > single representation
Conclusion

• Pretrained speaker representation + learnable speaker representations > single representation
• Generative pretraining > discriminative pretraining
Resources

• Audio Samples: https://ming024.github.io/M2VoC/
• Code: https://github.com/ming024/FastSpeech2/tree/M2VoC
• Paper: https://arxiv.org/abs/2103.04088