Neural Voice Cloning with a Few Samples

Sercan O. Arik, Jitong Chen, Kainan Peng*, Wei Ping, Yanqi Zhou
Motivations

- Text-to-speech (TTS) models can be conditioned on text and speaker identity.
  - Text: linguistic information, content of the generated speech.
  - Speaker identity: speaker information (accent, pitch, speech rate...).
Motivations

• Text-to-speech (TTS) models can be conditioned on text and speaker identity.
  • Text: linguistic information, content of the generated speech.
  • Speaker identity: speaker information (accent, pitch, speech rate…).

Limitations:
• Can only generate speech for observed speakers during training.
• Require lots of speech samples per speaker (e.g., Deep Voice 2).
Voice Cloning

• Voice cloning: synthesize the voices of new speakers from a few speech samples (few-shot generative model).

• Applications: personalized speech interfaces, content creation, assistive technology...
Voice Cloning

• Voice cloning: synthesize the voices of new speakers from a few speech samples (few-shot generative model).

• Applications: personalized speech interfaces, content creation, assistive technology...

• Challenges:
  • Generalization: learn the voice of a new speaker.
  • Efficiency: extract the speaker characteristics from a few speech samples.
  • Computational cost: cloning with low latency and small footprint.

• Two approaches:
  • Speaker adaptation.
  • Speaker encoding.
Speaker Adaptation

• Fine-tune a pre-trained multi-speaker model for a new speaker.
• Training data: a few text and audio pairs.
Speaker Adaptation

- Fine-tune a pre-trained multi-speaker model for a new speaker.
- Training data: a few text and audio pairs.
- Two options for speaker adaptation:

  ![Diagram]
  - Fine-tune the whole model
  - Fine-tune the speaker embedding only
# Speaker Adaptation Analysis

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Speaker Adaptation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Embedding-only</td>
<td>Whole-model</td>
</tr>
<tr>
<td>Cloning time</td>
<td>8 h</td>
<td>5 min</td>
</tr>
<tr>
<td># of parameters per speaker</td>
<td>128</td>
<td>25 million</td>
</tr>
</tbody>
</table>
Speaker Encoding

- Directly predict a new speaker embedding for a multi-speaker model.
- Train a speaker encoder with audio and speaker embedding pairs.
Speaker Encoding

- Directly predict a new speaker embedding for a multi-speaker model.
- Train a speaker encoder with audio and speaker embedding pairs.
- Cloning time: a few seconds, more favorable for low-resource deployment.
## Results

- Vocoder: classical Griffin-Lim algorithm.
- Demo website: [http://audiodemos.github.io](http://audiodemos.github.io)

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Speaker Adaptation</th>
<th>Speaker Encoding</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Embedding-only</td>
<td>Whole-model</td>
</tr>
<tr>
<td>Mean Opinion Score (MOS)</td>
<td>Naturalness (5-scale)</td>
<td>2.67</td>
</tr>
<tr>
<td></td>
<td>Similarity (4-scale)</td>
<td>2.95</td>
</tr>
</tbody>
</table>
Voice Morphing via Embedding Manipulation

- BritishMale + AveragedFemale - AveragedMale = BritishFemale
- BritishMale + AveragedAmerican - AveragedBritish = AmericanMale
Thank you!

Welcome to our poster, and listen to samples!

Today, Session B, #91