Evaluation of the Speech Resynthesis Capabilities of the VoicePrivacy Challenge Baseline B1

Únal Ege Gaznepoglu, Nils Peters (ege.gaznepoglu@audiolabs-erlangen.de)

1. Introduction

Problem: Speaker anonymization systems sound unnatural

- Many systems build upon the VPC Baseline B1 [1]
- Similar speech representation (F0, x-vector, bottleneck features)
- Same vocoder (AM-NSF, HiFiGAN etc.)

A non-intrusive estimate of PESQ by torchaudio-SQUIM [2]

Subjective evaluation with a MUSHRA-like listening test

2. Methodology

- Four objective, intrusive metrics
  - Mel-cepstral distortion (MCD):
    \[ \sum_{t=1}^{T} \sqrt{\sum_{i} (C_{t,i} - \hat{C}_{t,i})^2} \]
  - Scale-invariant signal-to-noise ratio (SI-SNR):
    \[ \frac{|\alpha_s|^2}{|\alpha_s - \hat{s}|^2} \text{, for } \alpha = \frac{\hat{s}^T s}{|s|^2} \]
  - Gross pitch error (GPE): \( \frac{\text{num. of frames whose error > 20\%}}{\text{num. of correctly identified voiced frames}} \)
  - Perceptual evaluation of speech quality (PESQ)

- A non-intrusive estimate of PESQ by torchaudio-SQUIM [2]
- Subjective evaluation with a MUSHRA-like listening test

3. Evaluated Systems

We bypass the anonymization block and resynthesize VoicePrivacy Challenge (VPC) test datasets [3] with the following B1 variants:

<table>
<thead>
<tr>
<th>ID</th>
<th>X-vector</th>
<th>Vocoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>mel-nsf-pt-spk</td>
<td>speaker-level</td>
<td>NSF</td>
</tr>
<tr>
<td>mel-nsf-spk</td>
<td>speaker-level</td>
<td>(C-based) NSF</td>
</tr>
<tr>
<td>mel-nsf-spk-4k</td>
<td>speaker-level</td>
<td>(C-based) NSF</td>
</tr>
<tr>
<td>am-nsf-spk</td>
<td>speaker-level</td>
<td>(C-based) AM + NSF</td>
</tr>
<tr>
<td>B1b-utt</td>
<td>utterance-level</td>
<td>joint NSF (+HiFiGAN-D)</td>
</tr>
<tr>
<td>B1b-spk</td>
<td>speaker-level</td>
<td>joint NSF (+HiFiGAN-D)</td>
</tr>
<tr>
<td>joint-hifigan-spk</td>
<td>speaker-level</td>
<td>joint HiFiGAN</td>
</tr>
</tbody>
</table>

4. Objective Evaluation Results (a)

GPE [%] (↓)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>35</td>
<td>40</td>
<td>45</td>
<td>50</td>
<td>55</td>
<td>60</td>
</tr>
</tbody>
</table>

MCD (↓)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>15</td>
<td>20</td>
<td>25</td>
<td>30</td>
<td>35</td>
<td>40</td>
</tr>
</tbody>
</table>

5. Subjective Evaluation (n ≥ 14)

Subjects commented joint-hifigan-nsf “Americanizes” the voices, unable to retain speaker characteristics

6. Conclusions

- 2020 systems (mel-nsf, am-nsf) perform better w.r.t. intrusive measures we considered (GPE, MCD, PESQ)
- Performance gap between copy synthesis (mel-nsf) and other systems → utilized representation causes additional information loss.
- torchSQUIM-PESQ estimate came the closest to predicting subjective preferences, whereas intrusive metrics could not predict them
- PyTorch and C implementations of am-nsf are not equivalent

Possible improvements:
- Additional incentive to help vocoders learn the speaker space
- Trying an improved speaker embedding (e.g., ECAPA as done by [4])

References