NWPU-ASLP System for the VoicePrivacy 2022 Challenge

Jixun Yao, Qing Wang, Li Zhang, Pengcheng Guo, Yuhao Liang, Lei Xie
Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, China
{yaojx, liangyuhao}@mail.nwpu.edu.cn, {qingwang, pcguo, lxie}@nwpu-aslp.org, lizhang.aslp.npu@gmail.com

Abstract
This paper presents the NWPU-ASLP speaker anonymization system for VoicePrivacy 2022 Challenge. Our submission does not involve additional Automatic Speaker Verification (ASV) model or x-vector pool. Our system consists of four modules, including feature extractor, acoustic model, anonymization module, and neural vocoder. First, the feature extractor extracts the Phonetic Posteriorgram (PPG) and pitch from the input speech signal. Then, we reserve a pseudo speaker ID from a speaker look-up table (LUT), which is subsequently fed into a speaker encoder to generate the pseudo speaker embedding that is not corresponding to any real speaker. To ensure the pseudo speaker is distinguishable, we further average the randomly selected speaker embedding and weighted concatenate it with the pseudo speaker embedding to generate the anonymized speaker embedding. Finally, the acoustic model outputs the anonymized mel-spectrogram from the anonymized speaker embedding and a modified version of HiFiGAN transforms the mel-spectrogram into the anonymized speech waveform. Experimental results demonstrate the effectiveness of our proposed anonymization system.

Index Terms: voice privacy, speaker anonymization, voice conversion, speech synthesis

1. Introduction
Speech data on the Internet are proliferating exponentially because of the emergence of virtual assistants, telecommunication and voice pay. Such applications send recorded personal data to centralized servers where the speech data get processed and stored [1]. However, speech data contain rich personal sensitive information that can be disclosed by automated systems. This includes, e.g., age, gender, health state and religious beliefs. Thus, with new regulations such as the General Data Protection Regulation (GDPR) in the EU, which strengthen privacy protection [2], how to suppress the original distribution of cosine distances between x-vectors is retained. Note that these approaches rely on additional ASV models or x-vectors pools and increase the within-class distance of the anonymized speaker.

Despite a lot of research in this area, speaker anonymization remains in its infancy. Consequently, the first VoicePrivacy 2020 Challenge was launched which focused on developing anonymization solutions for speech technology [9,11]. In a broad view, anonymization requires altering not only the speaker’s voice, but also linguistic content, extralinguistic traits, and background sounds which might reveal the speaker’s identity [12]. As a step towards this goal, the second edition – VoicePrivacy 2022 challenge – focuses on voice anonymization.

This paper aims at contributing to such anonymization systems and serves as a submission to the VoicePrivacy 2022 challenge. Different from the baseline system, our system does not involve additional ASV models or an x-vectors pool, which also reduces the risk of insufficient generalization of the ASV model and the complexity of anonymization computation. We propose an effective speaker anonymization method using two types of speaker embedding generated by a speaker encoder. Specifically, we reserve a pseudo speaker ID to generate pseudo speaker embedding ensuring the anonymized result is an artificial voice which is not corresponding to any real speaker. Furthermore, we average the randomly selected speaker embeddings, while the anonymized embedding is the weighted concatenation of averaged embedding and pseudo speaker embedding. Concatenation weight is a hyper-parameter chosen empirically. Given the linguistic feature and anonymized embedding, our method uses a neural acoustic model and a neural vocoder to generate an anonymized speech waveform. Experiments show that our proposed anonymization method leads to a notable increase in both primary metrics compared with baseline systems. Additionally, results show that the proposed anonymization system can suppress personally identifiable information and provide superior intelligibility to baseline systems.

2. System Overview
In this section, we will briefly introduce the framework of our proposed anonymization system which is presented in Fig.1. The anonymization system consists of four modules: (a) feature extraction, (b) acoustic model (AM), (c) anonymization module and (d) vocoder.

As shown in Fig.1, feature extraction module extracts fun-
damental frequency (F0), PPG and speaker ID from input speech. During training, speaker encoder is utilized to control the speaker identity by a one-hot speaker ID. Mel-spectrum is predicted by F0, PPG and speaker embedding using the acoustic model. Finally, vocoder reconstructs the generated mel-spectrum to waveform.

In look-up table (LUT), we reserve a pseudo speaker ID to generate pseudo speaker embedding. Based on the proposed framework, the speaker encoder described before, is used as an anonymized embedding generator. That is, the randomly selected speaker embedding as the condition for the pseudo speaker embedding that generates as output the anonymized embedding.

3. Implementation Details

3.1. Feature extraction

Due to the irregular periodicity of the glottal pulse, we often hear creaky voice in speech, which is usually manifested as jitter or sub-harmonics in signals [13]. This situation makes hard for F0 extractor to extract correct F0. To solve this problem, we take a hint from [14], which is an improvement with YIN algorithm [15]. For details, we set 80 bins of Yingram to represent a semitone range and Yingram to represent the frequency between 10.77 Hz and 768.40 Hz by setting window size to 1024.

It is crucial to extract rich linguistic information from the input speech to reconstruct an intelligible waveform and achieve a lower word error rate (WER). To this end, we resort to WeNet toolkit[1]. WeNet is an open-source speech recognition toolkit in which a new two-pass approach named U2++ implemented to end-to-end (E2E) speech recognition in a single model [19]. U2++ consists of three parts, shared encoder, connectionist temporal classification (CTC) decoder [17] and attention decoder. We train U2++ on the LibriSpeech-600 [13] and LibriTTS-600 [19], consisting of clean-100 and other-500. At the same time, we select the top 10 models with the best WER metrics for fusion. The extracted features from the shared encoder have shown superior performance on downstream tasks such as voice conversion. Therefore, we use the shared encoder output of 512-dimensional as PPGs that can provide rich linguistic information. We generated the LUT using the LibriTTS-600 dataset, which contains 1407 real speakers and 1 pseudo speaker. The speaker embedding is generated from the original speaker ID during training and the average embedding and pseudo embedding in the LUT during inference.

3.2. Acoustic model

We follow the implementation of Tacotron [20] and delete the attention layer in the CBHG module because the PPGs and the acoustic features are initially aligned. The CBHG module consists of three parts: a set of 1D convolution layers, a multi-layer highway network [21] and a bidirectional gated recurrent unit (GRU) layer. Highway network and bidirectional GRU are used to extract higher-level features and contextual features of the sequence, respectively. Speaker embedding is concatenated on the final output of CBHG module as input to decoder. The decoder utilizes an autoregressive model to improve the quality and intelligibility of the reconstructed mel-spectrogram.
embedding. In our strategy, the averaged embedding is similar to
the mean x-vector of a set of randomly selected x-vectors
from an x-vector pool. But we do not need an x-vector pool,
and the LUT is equivalent to our x-vector pool.

**Pseudo speaker embedding.** A pseudo speaker ID is re-
served in the LUT, which does not correspond to any of
the real speakers in the training data. This means that the pseudo
speaker ID is not involved in optimizing the training process.
The real speaker ID and the pseudo speaker ID produce differ-
ent distributions of speaker embedding during inference.

**Anonymized embedding.** Finally, the averaged speaker em-
bedding and the pseudo speaker embedding weighted concate-
nate to generate the anonymized embedding as output. The av-
eraged speaker embedding is in place to guarantee that each
trial utterance from a particular speaker is delivered by a single
pseudo-speaker, while the trial utterances from various speakers
are delivered by different pseudo-speakers.

### 3.4. Vocoder

The vocoder is our modified version of HiFiGAN gener-
a
tor \[22\], which can effectively reconstruct high-quality wave-
form. Specifically, we improve the original HiFiGAN by the fol-
lowing aspects. First, we increase the receptive field of the gen-
erator. Second, we use the multi-resolution STFT loss to bet-
ter measure the difference between fake and real speech. Last,
following the setting of multi-band MelGAN \[23\], we extend
HiFiGAN with multi-band processing as follows. The generator
takes mel-spectrograms as input and produces sub-band signals
instead of full-band signals. The modified HiFiGAN is proven
to be beneficial to speech generation. As for the upsampling
module in our HiFiGAN, 200x upsampling is conducted through
3 upsampling layers with 2x, 5x and 5x factors respectively
because of predicting 4 sub-bands simultaneously, the output
channels of the 3 upsampling networks are 256, 128 and 64,
respectively, and each residual dilated convolution stack has 3
layers with dilation 1, 3, 5 with kernel-size 3, 7, 11.

The modified HiFiGAN vocoder is trained with batch size
of 16 and AdamW optimizer with learning rate of 2e-4 on 1
NVIDIA GTX 2080 GPU. The discriminator uses the same op-
timization settings, with a fixed learning rate of 2e-4. As for
the generator, to reduce training time and GPU memory usage,
we employ a windowed generator for training, randomly sam-
pling segments of mel-spectrogram with a window size of 50
frames as input to the vocoder. We use the same training data
as the training data of the acoustic model and the ground truth

### 4. Evaluations and results

#### 4.1. VoicePrivacy 2022 Challenge

According to the VoicePrivacy 2022 Challenge, the anonymiza-
tion systems should: (a) preserve the linguistic content and par-
alinguistic attributes; (b) conceal the speaker identity; (c) en-
sure that all trial utterances from a given speaker are uttered by
the same pseudo-speaker and pseudo-speaker distinguishabil-
ity \[12\]. The training, development and evaluation corpora are
the same as for the VoicePrivacy 2020 Challenge. VoxCeleb-
1.2 \[24\] corpus and subsets of the LibriSpeech \[18\] and Lib-
riTTS \[19\] corpora are used for training voice anonymization
systems. The development set comprises LibriSpeech-dev-clean
and a subset of the VCTK \[25\] corpus, denoted VCTK-dev. For
VCTK-dev, there are two trial subsets: common and different,
common subset is intended to support subjective evaluation of
speaker verifiability in a text-dependent manner \[12\]. The eval-
uation set is similar to the development set.

#### 4.2. Evaluations

The VoicePrivacy 2022 Challenge focuses mainly on a pair of
metrics: the Equal Error Rate (EER) as the privacy met-
ric and the Word Error Rate (WER) as the primary utility
metric. The basic automatic speaker verification (ASV) and
automatic speech recognition (ASR) systems are trained on the
LibriSpeech-train-clean-360 and retrained on the utterance-
level anonymized data of LibriSpeech-train-clean-360 (denoted

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Gender</th>
<th>Weight</th>
<th>Orig</th>
<th>BL.a</th>
<th>BL.b</th>
<th>Condition1</th>
<th>Condition2</th>
<th>Condition3</th>
<th>Condition4</th>
</tr>
</thead>
<tbody>
<tr>
<td>LibriSpeech-dev</td>
<td>female</td>
<td>0.25</td>
<td>8.67</td>
<td>17.76</td>
<td>19.03</td>
<td>13.92</td>
<td>21.02</td>
<td>25.28</td>
<td>26.28</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>0.25</td>
<td>1.24</td>
<td>6.37</td>
<td>5.59</td>
<td>15.53</td>
<td>19.57</td>
<td>22.05</td>
<td>23.45</td>
</tr>
<tr>
<td>VCTK-dev (diff)</td>
<td>female</td>
<td>0.20</td>
<td>2.36</td>
<td>12.46</td>
<td>8.25</td>
<td>18.36</td>
<td>29.14</td>
<td>38.80</td>
<td>40.31</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>0.20</td>
<td>1.44</td>
<td>9.33</td>
<td>6.01</td>
<td>22.28</td>
<td>31.46</td>
<td>36.92</td>
<td>37.77</td>
</tr>
<tr>
<td>VCTK-dev (comm)</td>
<td>female</td>
<td>0.05</td>
<td>2.62</td>
<td>13.95</td>
<td>9.01</td>
<td>21.37</td>
<td>29.91</td>
<td>37.04</td>
<td>37.89</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>0.05</td>
<td>1.43</td>
<td>13.11</td>
<td>9.40</td>
<td>21.37</td>
<td>29.91</td>
<td>37.04</td>
<td>37.89</td>
</tr>
<tr>
<td>Weighted average</td>
<td></td>
<td></td>
<td>3.54</td>
<td>11.74</td>
<td>9.93</td>
<td>17.51</td>
<td>25.08</td>
<td>30.55</td>
<td>31.73</td>
</tr>
<tr>
<td>LibriSpeech-test</td>
<td>female</td>
<td>0.25</td>
<td>7.66</td>
<td>12.04</td>
<td>9.49</td>
<td>16.61</td>
<td>17.88</td>
<td>20.99</td>
<td>22.08</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>0.25</td>
<td>1.11</td>
<td>8.91</td>
<td>7.80</td>
<td>10.69</td>
<td>14.03</td>
<td>17.37</td>
<td>19.15</td>
</tr>
<tr>
<td>VCTK-test (diff)</td>
<td>female</td>
<td>0.20</td>
<td>4.89</td>
<td>16.00</td>
<td>10.91</td>
<td>23.10</td>
<td>34.83</td>
<td>40.84</td>
<td>40.64</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>0.20</td>
<td>2.07</td>
<td>10.05</td>
<td>7.52</td>
<td>23.19</td>
<td>30.20</td>
<td>37.54</td>
<td>38.81</td>
</tr>
<tr>
<td>VCTK-test (comm)</td>
<td>female</td>
<td>0.05</td>
<td>2.89</td>
<td>17.34</td>
<td>15.32</td>
<td>23.99</td>
<td>34.68</td>
<td>40.46</td>
<td>40.46</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>0.05</td>
<td>1.13</td>
<td>9.89</td>
<td>8.19</td>
<td>23.16</td>
<td>32.20</td>
<td>38.14</td>
<td>38.70</td>
</tr>
<tr>
<td>Weighted average</td>
<td></td>
<td></td>
<td>3.79</td>
<td>11.81</td>
<td>9.18</td>
<td>18.44</td>
<td>24.32</td>
<td>29.19</td>
<td>30.15</td>
</tr>
</tbody>
</table>

Table 2: Privacy results on different conditions. EER achieved by ASV on data processed by our anonymization method vs. EER achieved by baseline BL.a or BL.b and original (Orig).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Orig</th>
<th>BL.a</th>
<th>BL.b</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td>LibriSpeech-dev</td>
<td>3.82</td>
<td>4.34</td>
<td>4.19</td>
<td>3.91</td>
<td>3.71</td>
<td>3.65</td>
<td>3.65</td>
</tr>
<tr>
<td>VCTK-dev</td>
<td>10.79</td>
<td>11.54</td>
<td>10.98</td>
<td>8.10</td>
<td>7.73</td>
<td>7.68</td>
<td>7.62</td>
</tr>
<tr>
<td>Average</td>
<td>7.31</td>
<td>7.94</td>
<td>7.59</td>
<td>6.00</td>
<td>5.72</td>
<td>5.66</td>
<td>5.63</td>
</tr>
<tr>
<td>LibriSpeech-test</td>
<td>4.15</td>
<td>4.75</td>
<td>4.43</td>
<td>3.96</td>
<td>3.98</td>
<td>3.84</td>
<td>3.87</td>
</tr>
<tr>
<td>VCTK-test</td>
<td>12.82</td>
<td>11.82</td>
<td>10.69</td>
<td>8.37</td>
<td>7.85</td>
<td>7.81</td>
<td>7.85</td>
</tr>
<tr>
<td>Average</td>
<td>8.49</td>
<td>8.29</td>
<td>7.56</td>
<td>6.16</td>
<td>5.91</td>
<td>5.82</td>
<td>5.86</td>
</tr>
</tbody>
</table>

Table 3: Primary utility evaluation: WER achieved by ASV on data processed by our anonymization method (with the large LM). C* denotes different target EER conditions.
as \( AS_{\text{eval}}^{\text{man}} \) and \( AS_{\text{eval}}^{\text{sup}} \). To trade-off the privacy-utility, results were divided into four conditions based on the minimum target EER and the lower WER for a given EER condition, the better the rank of the anonymization system. The four minimum target EER condition is 15%, 20%, 25% and 30%.

In addition to the primary metrics, another two secondary utility metrics, namely pitch correlation \( \rho F \) and the gain of voice distinctiveness \( G_{VD} \) which measured how well anonymization preserves the intonation of the original utterance and evaluated the requirement to preserve voice distinctiveness, respectively. The gain of voice distinctiveness metric is defined as the ratio of diagonal dominance of the original matrices and the anonymized matrices. These two metrics are not used for final ranking, but the anonymized results should achieve a minimum average pitch correlation of \( \rho F > 0.3 \) for each dataset and each condition.

### 4.3. Results

The EER results are computed by averaging three EERs on LibriSpeech-test-clean, VCTK-test (common) and VCTK-test (different) datasets with weights of 0.5, 0.1 and 0.4, respectively. And the WER results are computed by averaging two WERs on LibriSpeech-test-clean and VCTK-test with equal weights. We meet the different EER conditions by adjusting the weight of pseudo speaker embedding and averaged embedding. The weights of the pseudo speaker embedding corresponding to different conditions are 0.6, 0.8, 0.9 and 0.95, respectively, and the averaged speaker embedding weights are 0.4, 0.2, 0.1 and 0.05, respectively.

Table 4: Secondary utility evaluation: pitch correlation \( \rho F \) achieved on data processed by B1.a, B1.b and our anonymized results.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Gender</th>
<th>B1.a</th>
<th>B1.b</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td>LibriSpeech-dev</td>
<td>female</td>
<td>0.77</td>
<td>0.84</td>
<td>0.70</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>0.73</td>
<td>0.76</td>
<td>0.69</td>
<td>0.69</td>
<td>0.69</td>
<td>0.69</td>
</tr>
<tr>
<td>VCTK-dev (dif)</td>
<td>female</td>
<td>0.84</td>
<td>0.87</td>
<td>0.76</td>
<td>0.76</td>
<td>0.77</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>0.78</td>
<td>0.76</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>VCTK-dev (com)</td>
<td>female</td>
<td>0.79</td>
<td>0.84</td>
<td>0.71</td>
<td>0.71</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>0.72</td>
<td>0.72</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>Weighted average dev</td>
<td>female</td>
<td>0.77</td>
<td>0.80</td>
<td>0.71</td>
<td>0.71</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>0.77</td>
<td>0.85</td>
<td>0.71</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>VCTK-test (dif)</td>
<td>female</td>
<td>0.87</td>
<td>0.87</td>
<td>0.77</td>
<td>0.76</td>
<td>0.77</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>0.79</td>
<td>0.77</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>VCTK-test (com)</td>
<td>female</td>
<td>0.79</td>
<td>0.85</td>
<td>0.72</td>
<td>0.71</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>0.70</td>
<td>0.71</td>
<td>0.64</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td>Weighted average test</td>
<td>female</td>
<td>0.77</td>
<td>0.80</td>
<td>0.70</td>
<td>0.70</td>
<td>0.71</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>0.77</td>
<td>0.84</td>
<td>0.70</td>
<td>0.71</td>
<td>0.72</td>
<td>0.72</td>
</tr>
</tbody>
</table>

#### 4.3.1. Primary privacy results

Privacy results on different conditions are shown in Table 2. The original speech and two baseline results are denoted as Orig, B1.a and B1.b, respectively. It is observed that our anonymization method meets all the conditions and achieves 18.44%, 24.32%, 29.19% and 30.15% average EERs. Our approach leads to a notable increase in average EER of up to 18.34% compared with B1.a and 20.97% compared with B1.b which demonstrates our anonymization method can protect speaker privacy effectively. We find that our results on LibriSpeech male corpus for different genders perform similarly instead of the significant difference in baseline systems.

#### 4.3.2. Primary utility results

As shown in Table 3, we observe that the WERs for all conditions are better than both baseline systems, with a slight improvement for the LibriSpeech-test, and a significant improvement for the VCTK-test. The lowest WERs are 3.84% and 7.781% on the Librispeech-test and the VCTK-test, respectively. The average WER is 2.47% and 1.74% lower than B1.a and B1.b which demonstrates our anonymization method can achieve better intelligibility than baseline systems.

#### 4.3.3. Secondary primary results

We report in Table 4 and Table 5 the results for the two secondary primary results. It can be clearly found that the pitch correlation of our anonymized data exceeds the minimum threshold and achieve similar results with baseline systems for each condition. The highest pitch correlation is achieved in condition 1 of 0.7. The results of \( G_{VD} \) show that voice distinctiveness worsens as the EER rises. This may be due to the increasing proportion of pseudo embedding in the anonymized embedding. Similarity matrices of pseudo speaker embedding do not have a clearly dominant diagonal, so the \( D_{\text{diag}}(M_{pp}) \) metric described in [26] will converge to 0.

#### 5. Conclusion

This paper presents our anonymization system for VoicePrivacy 2022 Challenge. The proposed system utilizes averaged embedding as the condition for pseudo speaker embedding to produce the final anonymized embedding, with the anonymized mel-spectrogram generated by the acoustic model. Finally, a ground truth alignment data retrained vocoder converts the anonymized mel-spectrograms to waveform. Under the VoicePrivacy 2022 Challenge evaluation plan, the experiment results show that our system achieves the best average EER and WER of 30.15% and 5.82%, respectively. The average EER and WER results demonstrate that our anonymization method can suppress personally identifiable information in the speech signal and outperform baseline systems in terms of intelligibility.
6. References


