1. Introduction: Verification-to-Synthesis (V2S) Attack

Automatic speaker verification (ASV):[1]
✓ identifies the speaker of the input voice.
→ If an attacker hacks the ASV, voices of enrolled speakers risk being reproduced.

Voice conversion (VC):
✓ predicts targeted speaker’s voice.
→ VC is a possible technique used in impersonation attack.

Deceiving the ASV has some possibility of reproducing the targeted speaker’s individuality by VC. We name this attack “verification-to-synthesis (V2S) attack”.

Our approach
✓ proposes VC training with the “white-boxed” ASV and pre-trained automatic speech recognition (ASR) models without the targeted speaker’s voice data.

Proposed VC performs comparably to the standard VC methods using a tiny amount of parallel voice data.

2. Standard VC Using Targeted Speaker’s Voices

(1) One-to-one parallel VC (ParaVC)[8]
✓ is trained to minimize mean squared error (MSE) between \( \hat{y} \) and \( y \) using a targeted speaker’s voice.

(2) One-to-many non-parallel VC (NonparaVC)[8]
✓ can convert an attacker voice to any arbitrary speaker’s voice.
✓ is often trained using multi-speaker corpora in advance.

Two standard VC models are trained using a targeted speaker’s voice. These are used as references to evaluate the performances of the proposed speaker V2S attack.

3. V2S Attack: VC without Using Targeted Speaker’s Voices

V2S attack model
✓ uses two DNN models for training the VC
→ Loss can be backpropagated to VC model.

White-boxed ASV model \( V(\cdot) \):
→ Targeted speaker’s label (\( l_y \)) is given.
→ \( V(\cdot) \) estimates speaker similarity between input voices (\( \hat{y} \)) & targeted speaker’s voices as softmax cross-entropy, \( L_{SCE}(l_y, V(\hat{y})) \).
→ It helps to reproduce the targeted speaker’s individuality, but does not keep the phonetic property of the input voice.

Pre-trained automatic speech recognition (ASR) model \( R(\cdot) \):
→ \( R(\cdot) \) estimates the discrepancy as MSE between \( R(x) \) and \( R(\hat{y}) \).
→ It helps to restore the phonetic property of the input voice.

4. Experimental Evaluation

Experimental conditions

<table>
<thead>
<tr>
<th>Compared model</th>
<th>(a) ParaVC: trained by ( 5, 30 ) utterances</th>
<th>(b) NonparaVC: trained by 260 pre-stored speakers</th>
<th>(c) V2S: trained by 200 utterances of attacker</th>
</tr>
</thead>
<tbody>
<tr>
<td>The number of enrolled speakers</td>
<td>260 Japanese speakers (130 males and 130 females)</td>
<td>39-dim. mel-cepstral coefficients, Log F0, 10-dim. bap</td>
<td></td>
</tr>
<tr>
<td>Speech params. (including ( \Delta ))</td>
<td>Feed-Forward (see our paper)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNN architectures</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attacker &amp; Targeted speakers</td>
<td>one attacker (one male) &amp; four targeted speakers (two males and two females)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation data</td>
<td>25 parallel voices</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Subjective evaluation
Naturalness (preference AB tests)

<table>
<thead>
<tr>
<th></th>
<th>male-to-male</th>
<th>male-to-female</th>
</tr>
</thead>
<tbody>
<tr>
<td>ParaVC</td>
<td>0.540 vs. 0.470</td>
<td>0.597 vs. 0.554</td>
</tr>
<tr>
<td>NonparaVC</td>
<td>0.560 vs. 0.520</td>
<td>0.610 vs. 0.602</td>
</tr>
<tr>
<td>V2S</td>
<td>0.572 vs. 0.538</td>
<td>0.610 vs. 0.610</td>
</tr>
</tbody>
</table>

Speaker individuality (preference XAB tests)

<table>
<thead>
<tr>
<th></th>
<th>male-to-male</th>
<th>male-to-female</th>
</tr>
</thead>
<tbody>
<tr>
<td>ParaVC</td>
<td>0.713 vs. 0.593</td>
<td>0.716 vs. 0.702</td>
</tr>
<tr>
<td>NonparaVC</td>
<td>0.593 vs. 0.593</td>
<td>0.638 vs. 0.593</td>
</tr>
<tr>
<td>V2S</td>
<td>0.680 vs. 0.569</td>
<td>0.770 vs. 0.710</td>
</tr>
</tbody>
</table>

5. Conclusion

V2S attack: voice impersonation attack using VC
✓ uses ASV, and ASR model for VC training.
✓ is trained without the targeted speaker’s voices.

Experimental result
→ V2S attack can synthesize voices that has naturalness and speaker individuality comparable to a standard parallel VC with a tiny amount of data.

We are planning to
✓ improve the performances of the V2S attack.
✓ investigate ways of preventing the V2S attack.

References