A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case

Zhizheng Wu$^{1,2}$, Tomi Kinnunen$^3$, Eng Siong Chng$^{1,2}$, Haizhou Li$^{1,2,4,5}$, Eliathamby Ambikairajah$^5$

1. School of Computer Engineering, Nanyang Technological University, Singapore
2. Temasek Laboratories@NTU, Nanyang Technological University, Singapore
3. School of Computing, University of Eastern Finland, Finland
4. Human Language Technology Department, Institute for Infocomm Research, Singapore
5. School of Electrical Engineering and Telecomm, University of New South Wales, Australia
Outline

- Introduction of speaker verification
- Spoofing attack
- Motivation
- Anti-spoofing techniques
- Experiments and results
- Conclusions and future work
Speaker verification

Speaker verification

I am Obama, verify me!

Not Obama!
Spoofing attack in speaker verification

I am Obama, verify me!

Voice conversion techniques can break a speaker verification system!
Motivation

- To prevent spoofing attack to speaker verification systems, we need to develop a technique to detect synthesized/converted speech.

- Phase information was not given much attention in speaker verification studies, but phase artifacts in synthesized speech is an informative cue.


Overview of voice conversion (1/4)

- GMM-based voice conversion

Source \[\rightarrow\] Analysis \[\rightarrow\] Transformation function \[\rightarrow\] Synthesis

Phase artifacts created between analysis and synthesis

Target
Overview of voice conversion (2/4)

- Unit-selection based voice conversion

Source frame sequence

Source frame sequence

Target frame sequence

Phase artifacts created between analysis and synthesis

Source

Analysis

Target Speech Inventory

Source

Synthesis

Target
Overview of voice conversion (3/4)

- An analysis-synthesis pass-through without transformation

Phase artifacts created between analysis and synthesis
The three voice conversion systems all adopt analysis-synthesis module.

- Use analysis module to extract features
- Use synthesis filter to reconstruct speech signal from features
- Hence, we can use the pass-through speech as training data for the synthetic speech detector
Anti-spoofing attack in speaker verification (1/3)

- Feature for detection: modified group delay phase

![Diagram showing natural and converted speech spectrograms with DCT coefficients highlighted.](image)

Apply DCT and keep 12 coefficients as the feature.
Anti-spoofing attack in speaker verification (2/3)

- **GMM-based detector**
  
  The decision is made based on the log-likelihood threshold

\[
\Lambda(C) = \log p(C | \lambda_{\text{converted}}) - \log p(C | \lambda_{\text{natural}})
\]

- \(C\) is the feature vector sequence of a testing speech
- \(\lambda_{\text{converted}}\) is the GMM model for converted speech
- \(\lambda_{\text{natural}}\) is the GMM model for natural speech
- 512 Gaussian components are employed in this study
Anti-spoofing attack in speaker verification (3/3)

Speaker verification

I am Obama, verify me!

Voice conversion

Obama?

Synthetic detector

Synthetic speech
Experimental setups (1/3)

- Original Dataset
  - subset of NIST 2006 SRE core task

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique speakers</td>
<td>298</td>
<td>206</td>
<td>504</td>
</tr>
<tr>
<td>Genuine trials</td>
<td>2,349</td>
<td>1,629</td>
<td>3,978</td>
</tr>
<tr>
<td>Impostor trials</td>
<td>1,636</td>
<td>1,146</td>
<td>2,782</td>
</tr>
</tbody>
</table>

- The duration of each conversation is about 5 minutes
• Spoofing dataset
  - Converted the 2,782 impostor samples to the target speakers (the claimed identities)
  - Use \texttt{3conv4w} and \texttt{8conv4w} training sections for voice conversion function training
  - SPTK: \url{http://sp-tk.sourceforge.net/} is used for analysis-synthesis
    ▸ Analysis: Mel-cepstral analysis
    ▸ Synthesis: MLSA filter
Experimental setups (3/3)

- Speaker verification systems
  - GMM Joint factor analysis (GMM-JFA) system
    - Models the intersession and speaker variability in the GMM supervector space
  - Probabilistic linear discriminant analysis system
    - PLDA system is similar as JFA system, but use i-vector as the basis for factor analysis
Experimental results

- Performance of speaker verification systems with/without anti-spoofing attack
  - Equal error rate (EER)
    - The lower of EER, the better performance

<table>
<thead>
<tr>
<th>Voice conversion</th>
<th>EER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without anti-spoofing</td>
</tr>
<tr>
<td></td>
<td>GMM-JFA</td>
</tr>
<tr>
<td>Baseline (No conversion)</td>
<td>3.24</td>
</tr>
<tr>
<td>GMM conversion</td>
<td>17.36</td>
</tr>
<tr>
<td>Unit-selection conversion</td>
<td>32.54</td>
</tr>
</tbody>
</table>
Conclusions and future work

- **Conclusions**
  - Voice conversion techniques present a threat to state-of-the-art speaker verification systems
  - Phase features are useful for detecting the synthetic speech from natural speech

- **Future work:**
  - Investigate vocoder independent features
  - Investigate temporal features for synthetic speech detection