Voice Conversion and Anti-spoofing of Speaker Verification

Haizhou Li

Acknowledgement:
Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Xiaohai Tian
• Spoofing Attacks
• Voice Conversion
• Artifacts
• ASVspoof 2015
• Spoofing Attacks
• Voice Conversion
• Artifacts
• ASVspoof 2015
Speaker Verification

This is John!

Reject!

Speaker Verification

Yes, John!
Spoofing Attacks

Impersonation

Replay

Speech Synthesis

Voice Conversion

This is John!

Reject!

Speaker Verification

Yes, John!
## Spoofing Attacks

<table>
<thead>
<tr>
<th>Spoofing attack</th>
<th>Accessibility</th>
<th>Effectiveness (risk)</th>
<th>Countermeasure availability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Text-independent</td>
<td>Text-dependent</td>
</tr>
<tr>
<td>Impersonation</td>
<td>Low</td>
<td>Low/unknown</td>
<td>Low/unknown</td>
</tr>
<tr>
<td>Replay</td>
<td>High</td>
<td>Low</td>
<td>Low to high</td>
</tr>
<tr>
<td>Speech synthesis</td>
<td>Medium to high</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Voice conversion</td>
<td>Medium to high</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

---

### Impersonation

<table>
<thead>
<tr>
<th>Spoofing attack</th>
<th>Accessibility</th>
<th>Effectiveness (risk)</th>
<th>Countermeasure availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impersonation</td>
<td>Low</td>
<td>Low/unknown</td>
<td>Low/unknown</td>
</tr>
<tr>
<td>Replay</td>
<td>High</td>
<td>Low</td>
<td>Low to high</td>
</tr>
<tr>
<td>Speech synthesis</td>
<td>Medium to high</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Voice conversion</td>
<td>Medium to high</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spoofing attack</th>
<th>Accessibility</th>
<th>Effectiveness (risk)</th>
<th>Countermeasure availability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Text-independent</td>
<td>Text-dependent</td>
<td></td>
</tr>
<tr>
<td>Impersonation</td>
<td>Low</td>
<td>Low/unknown</td>
<td>Low/unknown</td>
</tr>
<tr>
<td>Replay</td>
<td>High</td>
<td>Low</td>
<td>Low to high</td>
</tr>
<tr>
<td>Speech synthesis</td>
<td>Medium to high</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Voice conversion</td>
<td>Medium to high</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

 Traits of Replay

- J. Villalba and E. Lleida, “Preventing replay attack on speaker verification systems”, IEEE ICCST 2011
## Spoofing Attacks

<table>
<thead>
<tr>
<th>Spoofing attack</th>
<th>Accessibility</th>
<th>Effectiveness (risk)</th>
<th>Countermeasure availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impersonation</td>
<td>Low</td>
<td>Low/unknown</td>
<td>N.A.</td>
</tr>
<tr>
<td>Replay</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Speech synthesis</td>
<td>Medium to high</td>
<td>High</td>
<td>Medium</td>
</tr>
<tr>
<td>Voice conversion</td>
<td>Medium to high</td>
<td>High</td>
<td>Medium</td>
</tr>
</tbody>
</table>
More Robust = More Vulnerable

This is John!

Synthetic Speech Detection

Reject!

No

Speaker Verification

Yes, John!
• Spoofing Attacks
• Voice Conversion
• Artifacts
• ASVspoof 2015
Voice Conversion: Vocoder

Source Speaker → Analysis → Feature conversion → Synthesis → Target Speaker
Vocoder: Analysis - Synthesis

Source Speaker → Analysis → Feature conversion → Synthesis → Target Speaker
Sinusoidal vocoders

- Harmonic plus noise model (HNM) vocoder
- Harmonic and stochastic vocoder
- Adaptive harmonic vocoder

Source-filter model

- Linear predictive vocoder
- Mel – generalised cepstral vocoder
- STRAIGHT
- Glottal vocoder
Vocoder: Copy Synthesis

Source → Analysis → Synthesis → Target

<table>
<thead>
<tr>
<th>Feature</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>10.98</td>
</tr>
<tr>
<td>MGDCC</td>
<td>1.25</td>
</tr>
<tr>
<td>MGDCC+PM</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Voice Conversion: Feature Conversion

Source Speaker → Analysis → Feature conversion → Synthesis → Target Speaker
Differences between Speakers

Basics of Voice Conversion

Training

Conversion
• Masanobu Abe, Satoshi Nakamura, Kiyohiro Shikano, and Hisao Kuwabara. "Voice conversion through vector quantization." ICASSP 1988
Use partially the source spectrum information

Voice Conversion: Frame/Unit Selection

- Zhizheng Wu, Tuomas Virtanen, Tomi Kinnunen, Eng Siong Chng, Haizhou Li, "Exemplar-based unit selection for voice conversion utilizing temporal information", Interspeech 2013
Unit Selection Synthesis

- Source symbol – Target segment costs: suitability of unit for target
- Target segment - Target segment costs: acoustic continuity of two adjacent units
Subjective Analysis

Objective Analysis

“Spoofing Analysis”

1. *Spectral distortion*
2. *Temporal (magnitude/phase) discontinuity*
3. *Spectro-temporal artifacts*
4. *Pitch pattern*
5. *ASVspoof 2015 ?*
• Spoofing Attacks
• Voice Conversion
• Artifacts
• ASVspoof 2015
Magnitude

• Short-time Fourier transform (STFT)
• Smoothing effect (local vs global optimization)
• Temporal magnitude discontinuity

Phase

• Minimum phase vocoding
• Phase distortion
• Temporal phase discontinuity

… that are common to synthetic speech
… that are different from natural speech
• Time-Frequency resolution
• Spectral leakage
• Windowing tradeoffs
Fig. 1 Estimated spectra of Japanese vowel /a/ spoken by a male. Left wall of each panel also shows waveform and window shape. Three-dimensional plots have frequency axis (left to right in Hz), time axis (front to back in ms), and relative level axis (vertical in dB). Top panel shows spectrogram calculated using isometric Gaussian window. The center panel shows spectrogram with reduced temporal variation using a complementary set of windows. Bottom panel shows STRAIGHT spectrogram.

Magnitude: Smoothing in synthesized/converted speech

\[
\hat{y} = F(x) = \sum_{k=1}^{K} p_k(x)(\mu_k^{(y)} + \sum_{x}(\sum_{x})^{-1}(x - \mu_k^{(x)})),
\]

\[
p_k(x) = \frac{w_k \cdot \mathcal{N}(x | \mu_k^x, \Sigma_k^x)}{\sum_{k=1}^{K} w_k \cdot \mathcal{N}(x | \mu_k^x, \Sigma_k^x)},
\]


Fig. 5. Example of statistics and generated parameters from a sentence-level HMM composed of phoneme-level HMMs for /a/ and /i/. The dashed line and shading show the mean and standard deviation, respectively, of a Gaussian pdf at each state.

Magnitude: Pitch patterns in HMM-based synthesized speech

(a) Time stability of synthetic speech.

(b) Time stability of natural speech.

Fig. 6 An example of detection of time stability.

- Akio OGIHARA, Hitoshi UNNO, and Akira SHIOZAKI, Discrimination Method of Synthetic Speech Using Pitch Frequency against Synthetic Speech Falsification, IEICE TRANS. FUNDAMENTALS, VOL.E88–A, NO.1 JANUARY 2005


• Inter-Frame Difference of Log-Likelihood (IFDLL)

\[ \Delta l_t = |\log p(x_n | \lambda_C) - \log p(x_{n-1} | \lambda_C)| \]

• \(\Delta\)-Cepstrum and \(\Delta^2\)-Cepstrum

Figure 2: Inter-frame difference of log likelihood for natural and synthetic speech.

• Takayuki Satoh, Takashi Masuko, Takao Kobayashi, Keiichi Tokuda, “A Robust Speaker Verification System against Imposture Using an HMM-based Speech Synthesis System”, EUROspeech 2001
Fig. 2. Illustration of modulation feature extraction from power spectrogram.

Phase

- Wrapping
- Discontinuity
- Distortion

**Instantaneous Frequency**

**Time-derivative of phase for signal:**

\[ s_a(t) = a(t)e^{j\varphi(t)} \]

**Instantaneous Frequency:**

\[ f(t) = \frac{1}{2\pi} \frac{d\varphi(t)}{dt} \]

Frequency-derivative of phase

\[ \tau(\omega) = \frac{d\theta(\omega)}{d\omega} = -\left( \frac{d\left(\log X(\omega)\right)}{d\omega} \right)_I \]

\[ = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \]


• Spoofing Attacks
• Voice Conversion
• Artifacts
• ASVspoof 2015

ASVspoof 2015: Speaker verification spoofing and countermeasures challenge

Organisers
Zhizheng Wu, University of Edinburgh, UK
Tomi Kinnunen, University of Eastern Finland, Finland
Nicholas Evans, EURECOM, France
Junichi Yamagishi, University of Edinburgh, UK
## Voice Conversion Algorithms

<table>
<thead>
<tr>
<th># utterances</th>
<th>Algorithm</th>
<th>Vocoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training (10 male/15 female)</td>
<td>Development (15 male/20 female)</td>
<td>Evaluation (20 male/26 female)</td>
</tr>
<tr>
<td>Genuine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>VC :Frame-selection</td>
<td>STRAIGHT</td>
</tr>
<tr>
<td>S2</td>
<td>VC: Slope-shifting</td>
<td>STRAIGHT</td>
</tr>
<tr>
<td>S3</td>
<td>SS: HMM</td>
<td>STRAIGHT</td>
</tr>
<tr>
<td>S4</td>
<td>SS: HMM</td>
<td>STRAIGHT</td>
</tr>
<tr>
<td>S5</td>
<td>VC: GMM</td>
<td>MLSA</td>
</tr>
<tr>
<td>S6</td>
<td>VC: GMM</td>
<td>STRAIGHT</td>
</tr>
<tr>
<td>S7</td>
<td>VC: GMM</td>
<td>STRAIGHT</td>
</tr>
<tr>
<td>S8</td>
<td>VC: Tensor</td>
<td>STRAIGHT</td>
</tr>
<tr>
<td>S9</td>
<td>VC: KPLS</td>
<td>STRAIGHT</td>
</tr>
<tr>
<td>S10</td>
<td>SS: unit-selection</td>
<td>None</td>
</tr>
</tbody>
</table>
Table 4: Summary of primary submission results in the ASVspoof 2015 challenge.

<table>
<thead>
<tr>
<th>System ID</th>
<th>Equal Error Rates (EERs)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Known attacks</td>
<td>Unknown attacks</td>
<td>Average</td>
</tr>
<tr>
<td>A</td>
<td>0.408</td>
<td>2.013</td>
<td>1.211</td>
</tr>
<tr>
<td>B</td>
<td>0.008</td>
<td>3.922</td>
<td>1.965</td>
</tr>
<tr>
<td>C</td>
<td>0.058</td>
<td>4.998</td>
<td>2.528</td>
</tr>
<tr>
<td>D</td>
<td><strong>0.003</strong></td>
<td>5.231</td>
<td>2.617</td>
</tr>
<tr>
<td>E</td>
<td>0.041</td>
<td>5.347</td>
<td>2.694</td>
</tr>
<tr>
<td>F</td>
<td>0.358</td>
<td>6.078</td>
<td>3.218</td>
</tr>
<tr>
<td>G</td>
<td>0.405</td>
<td>6.247</td>
<td>3.326</td>
</tr>
<tr>
<td>H</td>
<td>0.670</td>
<td>6.041</td>
<td>3.555</td>
</tr>
<tr>
<td>I</td>
<td>0.005</td>
<td>7.447</td>
<td>3.726</td>
</tr>
<tr>
<td>J</td>
<td>0.025</td>
<td>8.168</td>
<td>4.097</td>
</tr>
<tr>
<td>K</td>
<td>0.210</td>
<td>8.883</td>
<td>4.547</td>
</tr>
<tr>
<td>L</td>
<td>0.412</td>
<td>13.026</td>
<td>6.719</td>
</tr>
<tr>
<td>M</td>
<td>8.528</td>
<td>20.253</td>
<td>14.391</td>
</tr>
<tr>
<td>N</td>
<td>7.874</td>
<td>21.262</td>
<td>14.568</td>
</tr>
<tr>
<td>O</td>
<td>17.723</td>
<td>19.929</td>
<td>18.826</td>
</tr>
<tr>
<td>Average</td>
<td><strong>3.337</strong></td>
<td><strong>9.294</strong></td>
<td><strong>6.316</strong></td>
</tr>
</tbody>
</table>

(Standard Deviation: 6.782, 6.861, 6.558)

Four times higher than that of known attacks

<table>
<thead>
<tr>
<th>Team</th>
<th>Average (all)</th>
<th>Average (without S10)</th>
<th>S10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1.211</td>
<td>0.402</td>
<td>8.490</td>
</tr>
<tr>
<td>B</td>
<td>1.965</td>
<td>0.008</td>
<td>19.571</td>
</tr>
<tr>
<td>C</td>
<td>2.528</td>
<td>0.076</td>
<td>24.601</td>
</tr>
<tr>
<td>D</td>
<td>2.617</td>
<td>0.003</td>
<td>26.142</td>
</tr>
<tr>
<td>E</td>
<td>2.694</td>
<td>0.060</td>
<td>26.393</td>
</tr>
<tr>
<td>F</td>
<td>3.218</td>
<td>0.400</td>
<td>28.581</td>
</tr>
<tr>
<td>G</td>
<td>3.326</td>
<td>0.360</td>
<td>30.021</td>
</tr>
<tr>
<td>H</td>
<td>3.726</td>
<td>0.021</td>
<td>37.068</td>
</tr>
<tr>
<td>I</td>
<td>3.898</td>
<td>0.703</td>
<td>32.651</td>
</tr>
<tr>
<td>J</td>
<td>4.097</td>
<td>0.029</td>
<td>40.708</td>
</tr>
<tr>
<td>K</td>
<td>4.547</td>
<td>0.203</td>
<td>43.638</td>
</tr>
<tr>
<td>L</td>
<td>6.719</td>
<td>3.478</td>
<td>35.890</td>
</tr>
<tr>
<td>M</td>
<td>14.391</td>
<td>12.482</td>
<td>31.574</td>
</tr>
<tr>
<td>N</td>
<td>14.568</td>
<td>11.299</td>
<td>43.991</td>
</tr>
<tr>
<td>O</td>
<td>18.826</td>
<td>16.304</td>
<td>41.519</td>
</tr>
<tr>
<td>P</td>
<td>21.518</td>
<td>18.786</td>
<td>46.102</td>
</tr>
</tbody>
</table>
S1-S5: voice conversion in train/dev/eval sets

S1: VC - Frame selection
S2: VC - Slope shifting
S3: TTS – HTS with 20 adaptation sentences
S4: TTS – HTS with 40 adaptation sentences
S5: VC – Festvox (http://festvox.org/)

S6 – S10: Only appear in the eval set

S6: VC – ML-GMM with GV enhancement
S7: VC – Similar to S6 but using LSP features
S8: VC – Tensor (eigenvoice)-based approach
S9: VC – Nonlinear regression (KPLS)
S10: TTS – MARY TTS unit selection (http://mary.dfki.de/)
LMS, Temporal CNN over 100 frames

LMS (using S1-S5 for training)

Spoofed: 0  score  Natural: 1
Fourier Transform vs Auditory Transform

Fig. 2. Impulse responses of the BM in the auditory transform (AT) when $\alpha = 3$ and $\beta = 0.2$, plotted by (5). The labels on the far left of each subplot represent the central frequency of the plotted impulse response. They are very similar to psychological measurements, such as the figures in [11], [12], [36, Fig. 1.12], [13], etc.

- Massimiliano Todisco, Héctor Delgado and Nicholas Evans, “A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients”, Odyssey 2016

Fig. 3. Frequency responses of the cochlear filters when $\alpha = 3$: (a) $\beta = 0.2$; and (b) $\beta = 0.035$. The filter band width can be adjusted by $\beta$ for different applications.
In STFT, the time and frequency resolutions are constant.

CQT employs a variable time/frequency resolution:
- greater time resolution for higher frequencies
- greater frequency resolution for lower frequencies

\[ Q = \frac{f_k}{\delta f} \]
\[ N_k = \frac{f_s}{f_k} Q \]

- Massimiliano Todisco, Héctor Delgado and Nicholas Evans, “A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients”, Odyssey 2016

*Courtesy of Nick Evans
Constant Q Cepstral Coefficients (CQCC)

Block diagram of CQCC feature extraction

Comparison of results (EER [%]) on ASVspoof2015 Database

Front-end: CQCC-A (19+0\textsuperscript{th} second derivative coefficients)
Back-end: 2 GMMs (512 components, EM training), one for human speech and one for spoofed speech

Matlab implementation of CQCC extraction can be downloaded from http://audio.eurecom.fr/content/software

*Courtesy of Nick Evans
• Most systems assume natural speech inputs
• More robust = More vulnerable
• Better speech perceptual quality ≠ less artifacts*
• Machines (frame-by-frame) and Humans (spectro-temporal) listen in different ways**
• Features are more important than classifiers

Thank You!