Voice Conversion and Spoofing Countermeasure in Speaker Verification

Haizhou Li

SIDAS 2017, Beijing

Acknowledgement: Zhizheng Wu, Xiaohai Tian, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi
Citi is going to start using voice patterns to authenticate customers over the phone in Asia

HSBC has been left red-faced after a BBC reporter and his non-identical twin tricked its voice ID authentication service.

The BBC says its “Click” (a weekly TV show) reporter Dan Simmons created an HSBC account and signed up to the bank’s service. HSBC states that the system is secure because each person’s voice is “unique”.

As Banking Technology reported last year, HSBC launched voice recognition and touch security services in the UK, available to 15 million banking customers. At that time, HSBC said the system “works by cross-checking against over 100 unique identifiers including both behavioural features such as speed, cadence and pronunciation, and physical aspects including the shape of larynx, vocal tract and nasal passages”.

According to the BBC, the “bank let Dan Simmons’ non-identical twin, Joe, access the account via the telephone after he mimicked his brother’s voice.

“Customers simply give their account details and date of birth and then say: ‘My voice is my password.’”

Despite this biometric bamboozle, Joe Simmons couldn’t withdraw money, but he was able to access balances and recent transactions, and was offered the chance to transfer money between accounts.

Joe Simmons says: “What’s really alarming is that the bank allowed me seven attempts to mimic my brothers’ voiceprint and get it wrong, before I got in at the eighth time of trying.”

Separately, the BBC says a Click researcher “found HSBC Voice ID kept letting them try to access their account after they deliberately failed on 20 separate occasions spread over 12 minutes”.

The BBC says Click’s successful thwarting of the system is believed to be “the first time the voice security measure has been breached”.

HSBC declined to comment to the BBC on “how secure the system had been until now”.

An HSBC spokesman says: “The security and safety of our customers’ accounts is of the utmost importance to us. Voice ID is a very secure method of authenticating customers.

“Twins do have a similar voiceprint, but the introduction of this technology has seen a significant reduction in fraud, and has proven to be more secure than PINS, passwords and memorable phrases.”

Not a great response is it? But very typical of the kind of bland statements that have taken hold in the UK. There is a problem and HSBC needs to get it fixed.

The rest of the BBC report just contains security experts saying the same things – like “I’m shocked”. Whatever. No point in sharing such dull insight.

You can see the full BBC Click investigation into biometric security in a special edition of the show on BBC News and on the iPlayer from 20 May.
With just a few minutes of audio samples, attackers can imitate your voice well enough to trick humans and state-of-the-art digital security systems, according to new UAB research.

Here’s how it’s done:

1. Collect samples in person or online.
2. Build a model of the victim’s speech patterns using “voice-morphing” software.
3. Use the model to say virtually anything in the victim’s voice, from passwords to entire conversations.

The UAB team is developing smarter verification systems and other defense strategies to defeat voice imitation attacks.
Terrifying AI learns to mimic your voice in under 60 seconds

By Mike Wehner, BGR

May 2, 2017 | 10:52am

When it comes to personal privacy and overall security, we often think of passwords, fingerprints, and even our own faces as being the keys that unlock our world, but what about your voice? If someone could perfectly mimic your voice, what kind of damage could they do? If they contacted people you know, could they lie their way into gaining private information about you?
Biometrics firm Nuance, which has focused on voice recognition, has announced a new multi-modal suite of biometric security solutions, driven by artificial intelligence (AI).

The new suite features facial and behavioural biometrics, as well as voice, with the company saying that these combine to provide advanced protection against fraud.

"By combining a range of physical, behavioural, and digital characteristics to provide secure authentication and more accurately detect fraud across multiple channels - from the phone to the Web, mobile apps and more - Nuance’s new Security Suite allows enterprises to attack fraud head-on, while at the same time offering an improved customer experience”, wrote the firm.

In particular, the firm notes improved synthetic speech detection.

Nuance has said that deep neural networks (DNN) are being used in the new solution alongside advanced algorithms to detect “synthetic speech attacks”.

Other site news

- IARPA awards $12.5 million contract to SRI
- IriTech to showcase mobile iris-barcode solution
- NEC tests face recognition with CBP at Dulles International Airport
• Spoofing Attacks
• Voice Conversion
• Artifacts
• ASVspoof 2015
Speaker Verification

This is John!

Reject!

Speaker Verification

Yes, John!
Spoofing Attacks

This is John!

Reject!

Speaker Verification

Yes, John!

Impersonation

Replay

Speech Synthesis

Voice Conversion
## 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>

• Zhizheng Wu, Sheng Gao, Eng Siong Chng, Haizhou Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification", Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 2014

• Luca Cuccovillo, Patrick Aichroth, "OPEN-SET MICROPHONE CLASSIFICATION VIA BLIND CHANNEL ANALYSIS," Luca Cuccovillo, Patrick Aichroth, ICASSP 2016
Ideas Against Replay Attack

App uses smartphone compass to stop voice hacking

The app uses the magnetometer in a phone, which is there for the phone’s compass, to detect a magnetic field. Credit: University at Buffalo
Tandem double-channel pop noise detection

Figure 2: Recording process in double-channel algorithm

Sayaka Shiota et al, VOICE LIVENESS DETECTION FOR SPEAKER VERIFICATION BASED ON A TANDEM SINGLE/DUOUBLE-CHANNEL POP NOISE DETECTOR, Speaker Odyssey 2016
There are 3 subfields of Phonetics, i.e. Articulatory Phonetics, Acoustic Phonetics, and Auditory Phonetics. *Denes & Pinson* (1993)
Speaker Verification


- Modeling the human voice production system
- Modeling the peripheral auditory system
Speaker Verification
transducer, channel
state of health, mood, aging
session variability

Challenges and Opportunities
Systems assume natural speech inputs
More robust = More vulnerable
Machines and Humans listen in different ways**
Better speech perceptual quality ≠ less artifacts*

This is Ming!

Spoofing Attacks

Speaker Verification

Yes, Ming!
• 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

source → Analysis → Synthesis → target
Vocoder: STRAIGHT

Source

![Waveform](image)

Analysis

![Feature](image)

Synthesis

Target

![Waveform](image)

<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
Chronological Map of Voice Conversion

Neural network-based methods
- ANN[17]

Exemplar-based methods

Frequency warping methods
- DFW[8]
- VTLN[9]
- Formant Mapping[10]

Parametric methods
- JD-GMM with MLPG and GV[5]
- PLS[6]
- DKPLS[7]

Codebook mapping methods
- VQ[1]
- Fuzzy VQ[2]

Other methods
- DNN[19]
- LSTM[20]
- AMA[22]
- Boltzmann machine[18]
- NMF[13]
- NMF-RC[14]
- CUTE[16]
- EEW-RC[15]
- PPG[21]

Timeline:
- 1988
- 1992
- 1995
- 1998
- 2003
- 2005
- 2007
- 2010
- 2012
- 2013
- 2014
- 2015
- 2016
- 2017
Subjective Analysis
1. Same/difference
2. Better/worse/same

Objective Analysis
1. Harmonics to Noise ratio
2. Jitter
3. Perceptual Evaluation of Speech Quality (PESQ)
4. …

“Spoofing Analysis”
1. Spectral distortion
2. Temporal discontinuity
3. Spectro-temporal artifacts
4. Pitch pattern
5. ASVspoof 2015?
• Spoofing Attacks
• Voice Conversion
• Artifacts
• ASVspoof 2015
Artifacts

Magnitude

• Short-time Fourier transform
• Smoothing effect (local vs global optimization)
  • HMM/GMM optimization
  • Vocoder effect
• Inaccuracy in statistical acoustic modeling
  • Glottal flow
  • Vocal tract modeling

Phase

• Minimum phase vocoding
• Phase distortion
• Phase discontinuity

\[
X(\omega) = |X(\omega)| e^{j\varphi(\omega)} \\
Y(\omega) = H(\omega) X(\omega)
\]

Gain & Phase shift
- **Time-Frequency resolution**
  - Fixed length windows
  - Individual frequency bands are distributed linearly, which is different from the distribution in the human cochlea

- **Windowing & spectral leakage**
• Inter-Frame Difference of Log-Likelihood (IFDPLL)
\[ \Delta l_t = |\log p(x_n | \lambda_c) - \log p(x_{n-1} | \lambda_c)| \]

• \( \Delta \)-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. 6  An example of detection of time stability.
Phase

- Wrapping
- Discontinuity
- Distortion

*Oppenheim, Schafer & Buck*, Discrete time digital signal processing, 2\textsuperscript{nd} Edition, Prentice Hall
**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

\[
\tau(\omega) = -\frac{d\theta(\omega)}{d\omega} = -\left(\frac{d(\log X(\omega))}{d\omega}\right)_I \\
= \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2},
\]

• 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
Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilci, Md Sahidullah, Aleksandr Sizov, "ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge", INTERSPEECH 2015
Joint effort of speech synthesis, voice conversion, and speaker verification community

# Voice Conversion Algorithms

<table>
<thead>
<tr>
<th># utterances</th>
<th>Training (10 male/15 female)</th>
<th>Development (15 male/20 female)</th>
<th>Evaluation (20 male/26 female)</th>
<th>Algorithm</th>
<th>Vocoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genuine</td>
<td>3750</td>
<td>3497</td>
<td>9404</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>S1</td>
<td>2525</td>
<td>9975</td>
<td>18400</td>
<td>VC: Frame-selection</td>
<td>STRAIGHT</td>
</tr>
<tr>
<td>S2</td>
<td>2525</td>
<td>9975</td>
<td>18400</td>
<td>VC: Slope-shifting</td>
<td>STRAIGHT</td>
</tr>
<tr>
<td>S3</td>
<td>2525</td>
<td>9975</td>
<td>18400</td>
<td>SS: HMM</td>
<td>STRAIGHT</td>
</tr>
<tr>
<td>S4</td>
<td>2525</td>
<td>9975</td>
<td>18400</td>
<td>SS: HMM</td>
<td>STRAIGHT</td>
</tr>
<tr>
<td>S5</td>
<td>2525</td>
<td>9975</td>
<td>18400</td>
<td>VC: GMM</td>
<td>MLSA</td>
</tr>
<tr>
<td>S6</td>
<td>0</td>
<td>0</td>
<td>18400</td>
<td>VC: GMM</td>
<td>STRAIGHT</td>
</tr>
<tr>
<td>S7</td>
<td>0</td>
<td>0</td>
<td>18400</td>
<td>VC: GMM</td>
<td>STRAIGHT</td>
</tr>
<tr>
<td>S8</td>
<td>0</td>
<td>0</td>
<td>18400</td>
<td>VC: Tensor</td>
<td>STRAIGHT</td>
</tr>
<tr>
<td>S9</td>
<td>0</td>
<td>0</td>
<td>18400</td>
<td>VC: KPLS</td>
<td>STRAIGHT</td>
</tr>
<tr>
<td>S10</td>
<td>0</td>
<td>0</td>
<td>18400</td>
<td>SS: unit-selection</td>
<td>None</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/)
Table 4: Summary of primary submission results in the ASVspoof 2015 challenge.

<table>
<thead>
<tr>
<th>System ID</th>
<th>Known attacks</th>
<th>Unknown attacks</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<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>3.355</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>
</tbody>
</table>

Average values: Known attacks (STD: 6.782), Unknown attacks (STD: 6.861), Average (STD: 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>
The cochlear filter, as the most important part of the transform, is defined as

$$
\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \psi \left( \frac{t - b}{a} \right)
= \frac{1}{\sqrt{|a|}} \left( \frac{t - b}{a} \right)^{\alpha} \exp \left[ -2\pi f_L \beta \left( \frac{t - b}{a} \right) \right]
\times \cos \left[ 2\pi f_L \left( \frac{t - b}{a} \right) + \theta \right] u(t - b) \tag{5}
$$

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.

“. . ., we use a longer window for a lower frequency band to average out the background noise and a shorter window for a higher frequency band to protect high-frequency information.”

Recent Progress: Instantaneous Frequency of Auditory Transform (CFCCIF)

Figure 2: (a) A natural utterance (16 kHz) provided from the challenge [20], (b) CFCC: the output of 28 cochlear subband filters and (c) CFCCIF: the output of 28 cochlear subband filters with the IF information.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Known attacks (% EER)</th>
<th>Unknown attacks (% EER)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S1</td>
<td>S2</td>
</tr>
<tr>
<td>A: DA-IICT</td>
<td>0.1013</td>
<td>0.8629</td>
</tr>
<tr>
<td>Average (Proposed)</td>
<td>0.407899</td>
<td></td>
</tr>
<tr>
<td>Avg. of 16 submissions</td>
<td>3.337</td>
<td></td>
</tr>
</tbody>
</table>

Recent Progress: Fourier Transform vs Constant Q

- 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

$$Q = \frac{f_k}{\delta f} \quad N_k = \frac{f_s}{f_k} Q$$

- Massimiliano Todisco, H´ector Delgado and Nicholas Evans, “A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients”, Odyssey 2016
Recent Progress: Constant Q Cepstral Coefficients (CQCC)

Block diagram of CQCC feature extraction

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

Front-end: CQCC-A (19+0th 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