The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection

Tomi Kinnunen, University of Eastern Finland, FINLAND
Md Sahidullah, University of Eastern Finland, FINLAND
Héctor Delgado, EURECOM, FRANCE
Massimiliano Todisco, EURECOM, FRANCE
Nicholas Evans, EURECOM, FRANCE
Junichi Yamagishi, Univ. of Edinburgh, UK & National Institute of Informatics, JAPAN
Kong Aik Lee, Institute for Infocomm Research, SINGAPORE
Organizers

Tomi H. Kinnunen
UEF, Finland

Md Sahidullah
UEF, Finland

Héctor Delgado
EURECOM, France

Massimiliano Todisco
EURECOM, France

Nicholas Evans
EURECOM, France

Junichi Yamagishi
Univ. of Edinburgh, UK

Kong Aik Lee
I²R, Singapore

NII, Japan
Structure of the session

First slot 11:00 – 13:00

INTRODUCTION, 30 mins

6 ORAL PRESENTATIONS, each 12 + 3 min

Second slot 14:30 – 16:30

6 ORAL PRESENTATIONS, each 12 + 3 min

GENERAL DISCUSSION @ 16:00---

CHAIRS: Tomi Kinnunen, Junichi Yamagishi

CHAIRS: Nicholas Evans, Kong Aik Lee
Spoofing attacks
a.k.a. presentation attacks [ISO/IEC 30107-1:2016]

Sources: unknown
Replay attack

replay spoofing – Sneakers (1992)
History of ASVspoof

- 1999: small, purpose collected datasets
- 2006: adapted, standard datasets
- 2013: OCTAVE project starts
- 2014: 2013 Interspeech special session
- 2016: common datasets, metrics, protocols
- 2017: common datasets, replay, generalisation, channel variation
- ASVspoof 2015
- ASVspoof 2017
Replay attack countermeasures

1. Phrase prompting with utterance verification
   Did the user speak the prompted text?

2. Audio fingerprinting
   Do I know this recording?

3. Speaker-independent replay detection
   Is this recording authentic or replayed one?

Can be circumvented using voice conversion
Dynamically increasing database size
Most general - but can it be done?

ASVspoof 2017

Replayed or nonreplayed?

Authentic (non-replayed)

Replayed

Replayed
ASVspoof challenge task
Standalone, speaker-independent detection of spoofing attacks

**ASVspoof 2015**

A speech sample → Synthetic or converted voice detector → Score

High score ➔ more likely a live human being
Low score ➔ more likely a spoofed sample

**ASVspoof 2017**

A speech sample → Replay speech detector → Score
Evaluation metric:
Equal error rate (EER) of replay-nonreplay discrimination

- **ASVspoof 2015**: EERs averaged across attacks
- **ASVspoof 2017**: EERs from pooled scores

![Graph showing EERs for detector A and B]
Crowdsourced replay attacks

RedDots corpus
[https://sites.google.com/site/thereddotsproject/]

- Text-dependent automatic speaker verification
- Collected by volunteers (ASV researchers)
- Various Android devices, speakers, accents
Examples of replay configurations

**REPLAY CONFIGURATION =**
Playback device + Environment + Recording device

- **Smartphone → Smartphone**
- **Headphones → PC mic**
- **High-quality loudspeaker → smartphone, anechoic room**
- **High-quality loudspeaker → high-quality mic**
- **Laptop line-out → PC line-in using a cable**

- **TRAINING SET**
  - 10 speakers
  - 3 replay configs
- **DEVELOPMENT SET**
  - 8 speakers
  - 10 replay configs
- **EVAL SET**
  - 24 speakers
  - 110 replay configs

- Ground truth provided
- Re-partitioning allowed
Impact of replay samples to ASV

Genuine vs. zero-effort impostors
EER = 1.8 %

Genuine vs. replay impostors
EER = 31.5 %
Participant statistics

• Registration: 113 teams or individuals
• Submitted results: 49 (43%)
Challenge results and further analyses

- Official challenge results

- Further analyses
Official challenge results
• Very difficult challenge!
• 21 submissions outperformed the baseline
• S01: >70% relative improvement w.r.t baseline B01
• B01 – B02: Important performance improvement when using pooled train+dev data for training
## Summary of top 10 systems

<table>
<thead>
<tr>
<th>ID</th>
<th>EER</th>
<th>Features</th>
<th>Post-proc.</th>
<th>Classifiers</th>
<th>Fusion</th>
<th>#Subs.</th>
<th>Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>S01</td>
<td>6.73</td>
<td>Log-power Spectrum, LPCC</td>
<td>MVN</td>
<td>CNN, GMM, TV, RNN</td>
<td>Score</td>
<td>3</td>
<td>T</td>
</tr>
<tr>
<td>S02</td>
<td>12.34</td>
<td><strong>CQCC</strong>, MFCC, PLP</td>
<td>WMVN</td>
<td>GMM-UBM, TV-PLDA, GSV-SVM, GSV-GBDT, GSV-RF</td>
<td>Score</td>
<td>-</td>
<td>T</td>
</tr>
<tr>
<td>S03</td>
<td>14.03</td>
<td>MFCC, IMFCC, RFCC, LFCC, PLP, <strong>CQCC</strong>, SCMC, SSFC</td>
<td>-</td>
<td>GMM, FF-ANN</td>
<td>Score</td>
<td>18</td>
<td>T+D</td>
</tr>
<tr>
<td>S04</td>
<td>14.66</td>
<td>RFCC, MFCC, IMFCC, LFCC, SSFC, SCMC</td>
<td>-</td>
<td>GMM</td>
<td>Score</td>
<td>12</td>
<td>T+D</td>
</tr>
<tr>
<td>S05</td>
<td>15.97</td>
<td>Linear filterbank feature</td>
<td>MN</td>
<td>GMM, CT-DNN</td>
<td>Score</td>
<td>2</td>
<td>T</td>
</tr>
<tr>
<td>S06</td>
<td>17.62</td>
<td><strong>CQCC</strong>, IMFCC, SCMC, Phrase one-hot encoding</td>
<td>MN</td>
<td>GMM</td>
<td>Score</td>
<td>4</td>
<td>T+D</td>
</tr>
<tr>
<td>S07</td>
<td>18.14</td>
<td>HPCC, <strong>CQCC</strong></td>
<td>MVN</td>
<td>GMM, CNN, SVM</td>
<td>Score</td>
<td>2</td>
<td>T+D</td>
</tr>
<tr>
<td>S08</td>
<td>18.32</td>
<td>IFCC, CFCCIF, Prosody</td>
<td>-</td>
<td>GMM</td>
<td>Score</td>
<td>3</td>
<td>T</td>
</tr>
<tr>
<td>S10</td>
<td>20.32</td>
<td><strong>CQCC</strong></td>
<td>-</td>
<td>ResNet</td>
<td>None</td>
<td>1</td>
<td>T</td>
</tr>
<tr>
<td>S09</td>
<td>20.57</td>
<td>SFFCC</td>
<td>-</td>
<td>GMM</td>
<td>None</td>
<td>1</td>
<td>T</td>
</tr>
<tr>
<td>D01</td>
<td>7.00</td>
<td>MFCC, <strong>CQCC</strong>, WT</td>
<td>MVN</td>
<td>GMM, TV-SVM</td>
<td>Score</td>
<td>26</td>
<td>T+D</td>
</tr>
</tbody>
</table>

**Using baseline CQCC features**

**DNN-based classifier**

**Other classifier**

T: training

T+D: training + development
Further analyses
Defining evaluation conditions

- **110 replay configurations** in evaluation set
- Characterize replay configurations through objective measurements
  - **Signal-to-noise ratio (SNR)**
  - **Cepstral distance (CSD)**: measures the degradation of a replayed recording w.r.t. its source recording
- Intuition:
  - More difficult attacks → **High SNR, low CSD**
  - Easier attacks → **Low SNR, high CSD**
Average quality measures per replay configuration

Averaged CSD vs. SNR for each replay configuration

Average CSD vs. SNR scatter plot for the 110 replay configurations
Data-driven clustering process

**Alternative approach**: define evaluation conditions according to countermeasure performance

1. Top Countermeasures fusion
2. Trial score computation and Replay Configuration averaging
3. Clustering

Evaluation conditions
1. Countermeasure fusion

Oracle linear fusion\(^1\) of systems S01 to B01 to obtain a high performance countermeasure

\begin{table}
\begin{tabular}{|l|l|}
\hline
\textbf{System} & \textbf{EER (%)} \\
\hline
S01 & 6.73 \\
S02 & 12.34 \\
S03 & 14.03 \\
S04 & 14.66 \\
S05 & 15.97 \\
S06 & 17.62 \\
S07 & 18.14 \\
S08 & 18.32 \\
S09 & 20.32 \\
S10 & 20.57 \\
S11 & 21.11 \\
S12 & 21.51 \\
S13 & 21.98 \\
S14 & 22.17 \\
S15 & 22.39 \\
S16 & 23.16 \\
S17 & 23.24 \\
S18 & 23.29 \\
S19 & 23.78 \\
B01 & 24.77 \\
D01 & 7.00 \\
\textbf{Fused} & \textbf{2.76} \\
\hline
\end{tabular}
\end{table}

\(^1\)Using the Bosaris toolkit
2. Average Replay Configuration (RC) scores computation and sorting

Data-driven clustering process

- Replay segments
- Countermeasure scores
- Average CM scores per RC
- Sorted average CM scores per RC

Replay segments:
- RC-001: seg_1, seg_2, ..., seg_N_001
- RC-002: seg_1, seg_2, ..., seg_N_002
- RC-110: seg_1, seg_2, ..., seg_N_110

Countermeasure scores:
- Score_1, score_2, ..., score_N_001
- Score_1, score_2, ..., score_N_002
- Score_1, score_2, ..., score_N_110

Average CM scores per RC:
- Avg_score

Sorted average CM scores per RC:
- Avg_score
- Avg_score
- Avg_score
3. Average scores clustering with k-means

Data-driven clustering process

Clustering solution based on CM averaged fused scores per replay configuration

Replay configuration index (sorted by increasing fused score)

C1
C2
C3
C4
C5
C6

Loopcable

Smartphone / tablet / portable device / laptop

Netbook speaker + webcam mic

Loopcable, anechoic chamber, good quality speakers/mics...
Obtained evaluation conditions

Averaged fused score, cepstral distortion and signal-to-noise ratio of the resulting evaluation conditions
Performance of top-10 primary submissions per evaluation condition

Box plot of top-10 systems' performance for clusters C1-C6

Pooled EER vs. weighted EER for top-10 systems

(equal to average EER used in ASVspoof 2015)
Conclusions

• Successful crowdsourcing approach to replay data collection
• Probably the most ‘wild’ replay data for ASV
  – Difficult to characterize
• Top-ranked system
  – ~70% relative improvement w.r.t. the baseline system
  – Fusion of only 3 subsystems!
• Encouraging performance
  – Limits of replay detection
  – Excepting unrealistic attacks (loopcable), high detection performance for high quality attacks
The 2nd Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2017) Database

Citation
Kinnunen, Tomi; Sahdullah, Md. Delgado, Héctor; Todisco, Massimiliano; Evans, Nicholas; Yamagishi, Junichi; Lee, Kong Aik. (2017). The 2nd Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2017) Database. [sound]. The Centre for Speech Technology Research (CSTR), University of Edinburgh.

http://dx.doi.org/10.7488/ds/2105

Averaged CSD vs. SNR for each replay configuration

Equal error rate (EER, %)

1. Top Countermeasures fusion
2. Trial score computation and Replay Configuration averaging
3. Clustering

System ID