IBERSPEECH 2022
GRANADA 14-16 NOV

# Secure and explainable voice biometrics


**Massimiliano Todisco**

**EURECOM, France**

GRADUATE SCHOOL AND RESEARCH CENTER IN DIGITAL SCIENCE

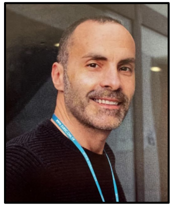
EURECOM
Sophia Antipolis

# Audio Security and Privacy group

EURECOM
Sophia Antipolis

**Digital Security Department**

**Nicholas Evans**
Full Professor
Head of the group



**Massimiliano Todisco**
Professor

**Hemlata Tak**
PhD student

Front-end optimisation & graph attention network for anti-spoofing

**Oubaïda Chouchane**
PhD student

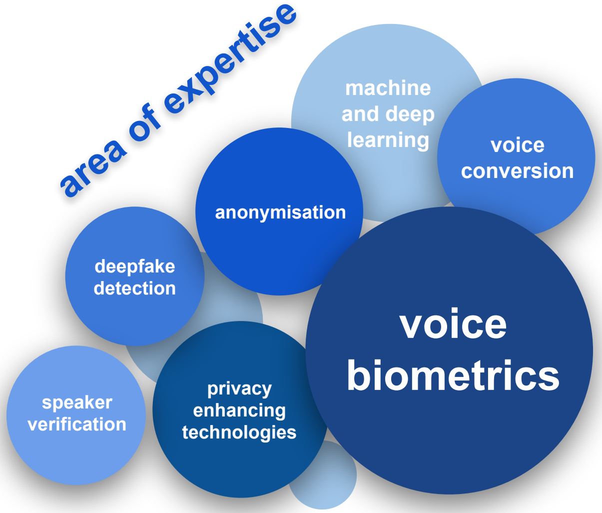
Cryptographic primitives for GDPR compliance

**Wanying Ge**
PhD student

Evolutive and end-to-end learning for ASV anti-spoofing

**Michele Panariello**
PhD student

Speaker anonymisation and protection




on-going projects and initiatives

RESPECT
ANR
DFG Deutsche Forschungsgemeinschaft

TReSPAsS-ETN
MARIE SKŁODOWSKA-CURIE ACTIONS

JST-ANR VoicePersonae Project (2018-2023)

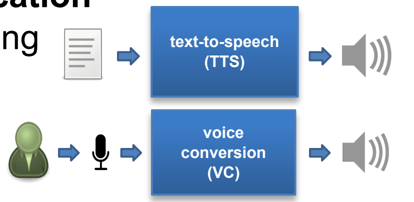
ASVspoof
Automatic Speaker Verification and Spoofing Countermeasures Challenge

Voice Privacy Challenge

# Intro

## Voice biometrics and anti-spoofing

- **Automatic Speaker Verification**
  - vulnerability to spoofing
- **Spoofing**
  - speech synthesis
  - voice conversion
  - replay attacks





- **Countermeasures**
  - CQCC features*
  - DNN based countermeasures
  - data augmentation techniques
  - single and integrated systems
- **ASVspoof initiative co-founders** [www.asvspoof.org](http://www.asvspoof.org)
  - ASVspoof 2015, 2017, 2019, 2021
  - ASVspoof5 is under development

## Voice privacy enhancing technologies

- **Voice biometrics and speech processing**
  - vulnerability to privacy
  - privacy threats
  - biometric template theft
  - biometric data theft


- **Privacy solutions**
  - anonymisation
  - homomorphic encryption
  - multi-party computation
  - DNN based encryption

- **VoicePrivacy initiative co-organisers** [www.voiceprivacychallenge.org](http://www.voiceprivacychallenge.org)
  - VoicePrivacy 2020, 2022

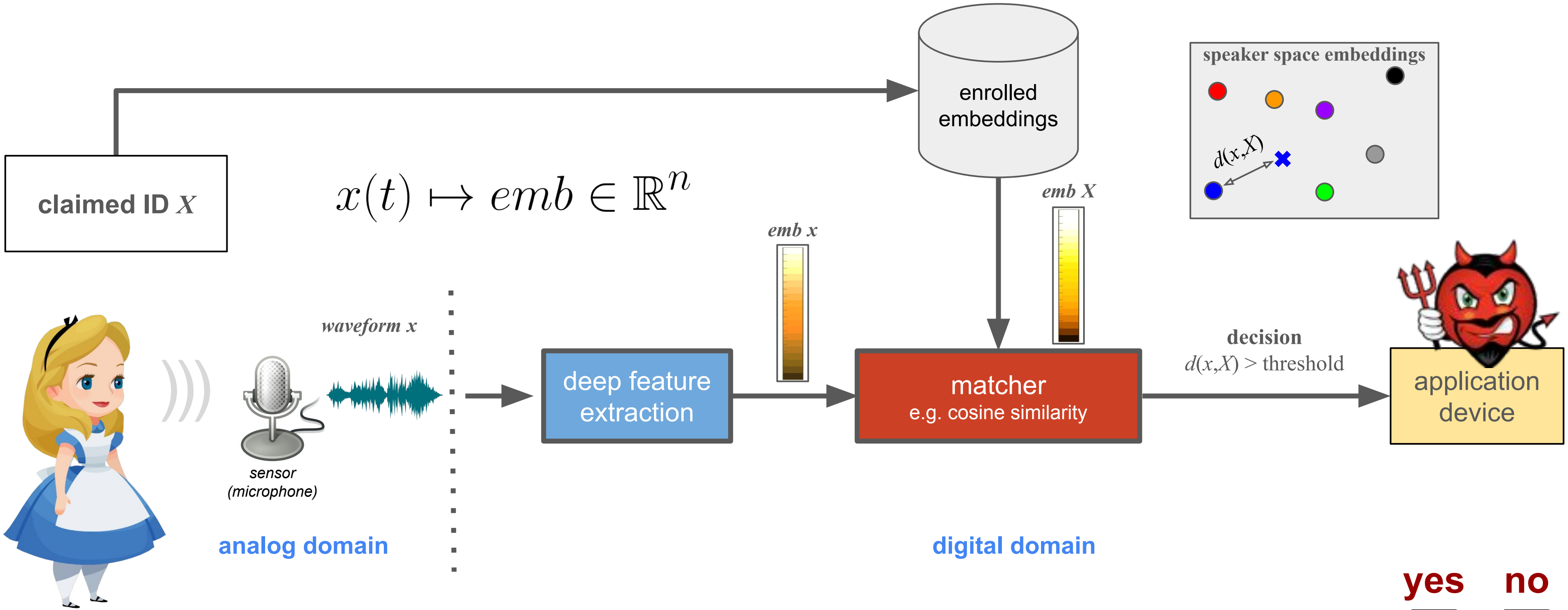

*__ISCA 2020 Award__ for the best article published in the journal *Computer Speech and Language* during the quinquennium 2015-2019

# Agenda

- **The era of voice cloning against voice biometrics**
  - automatic speaker verification, spoofing attacks and countermeasures
- **The ASVspoof challenge series**
  - from where it started from → necessity as motivation
  - and where it arrived → the lesson has actually been learned
- **Voice cloning artefacts: a recent history of detection…**
  - *the constant Q cepstral coefficients (CQCCs) → modeling time-frequency atoms*
  - *improve generalisation and robustness*
    - *RawNet2 → a deep network operating on time waveform*
    - *RawGAT → graph attention networks: pay more focus on time-frequency atoms*
    - *AASIST → an integrated spectro-temporal heterogeneous graph attention networks*
    - *RawBoost → a data augmentation based on signal processing*
    - *SSL → self-supervised learning to learn more generalised representation*
  - …and explainability
    - *SHapley Additive exPlanations (SHAP) → a simple and effective way to get evidence*
- **Links to open-source codes**
- **ASVspoof5: a glimpse into the future**

# The era of voice cloning against voice biometrics

# Automatic Speaker Verification (ASV)

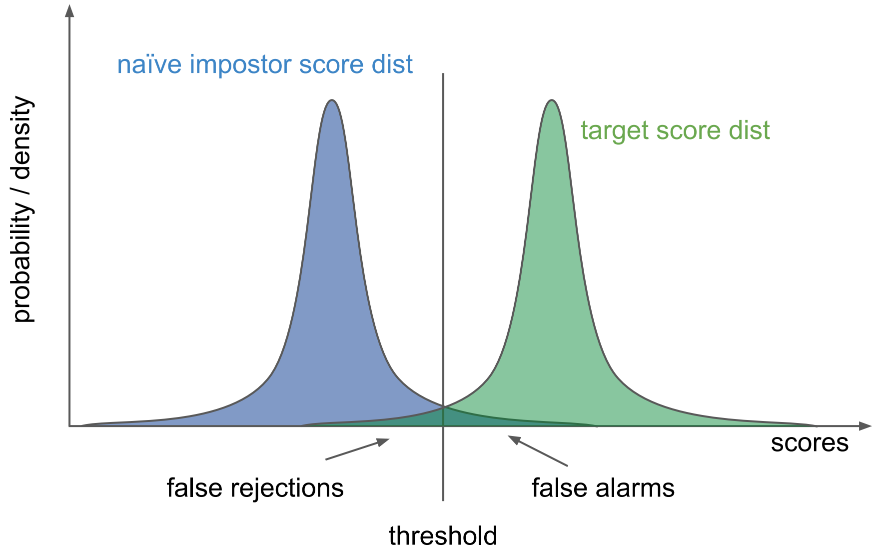



1. **is an ASV system vulnerable to impostors?**
2. **who are actually the impostors?**

**yes** ✔ **no** ☐

# Automatic Speaker Verification (ASV)

- **A binary classification framework**
  - targets who speak with their natural voice
  - naïve impostors who make **no effort** to impersonate the target

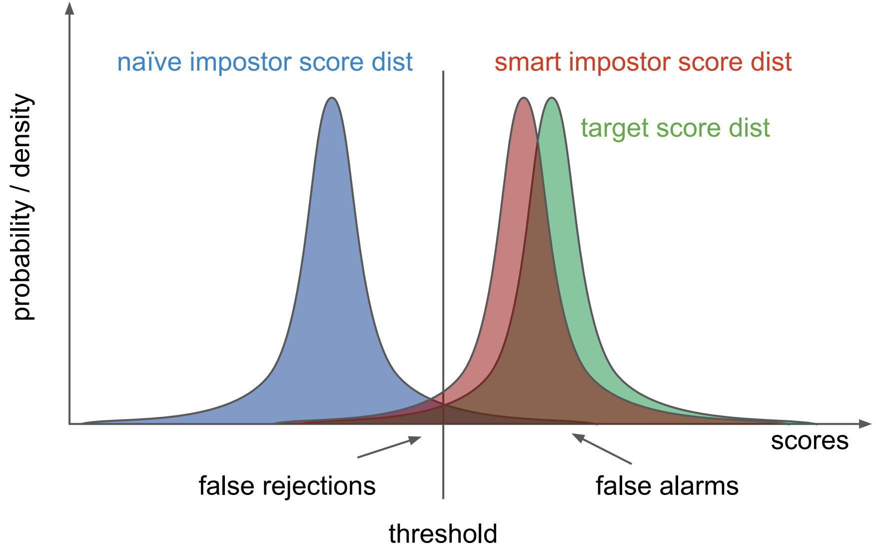



| IF the test users are | decision | |
|---|---|---|
| | **accept** | **reject** |
| targets *bona fide* | **correct** | error (false rejection) |
| naïve impostors *bona fide* | error (false alarm) | **correct** |

# Automatic Speaker Verification (ASV)

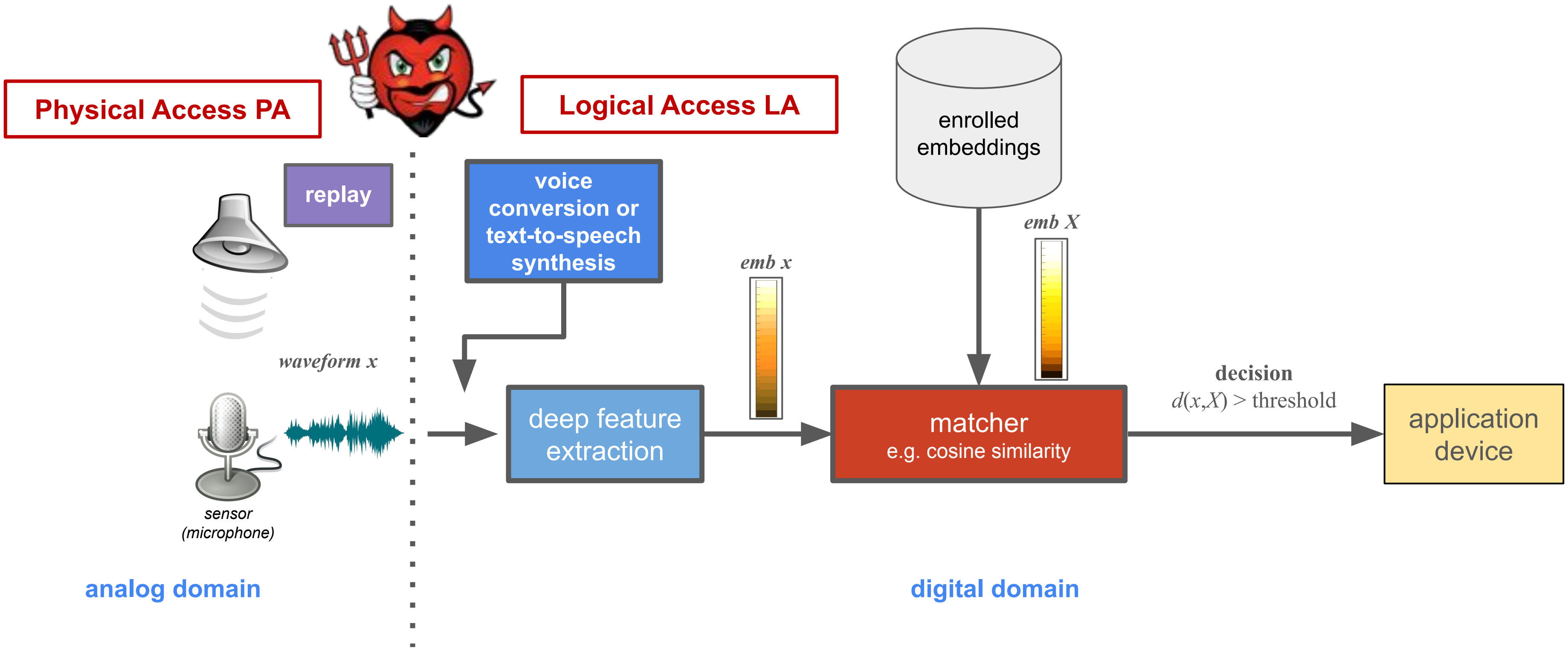
- **A (quasi) binary classification framework**
  - targets who speak with their natural voice
  - naïve impostors who make **no effort** to impersonate the target
  - smart impostors who make **effort** to impersonate the target

the aim of an attacker is to provoke false alarms by increasing ASV classifier scores (nontarget) while avoiding detection




| IF the test users are | decision | |
|---|---|---|
| | **accept** | **reject** |
| targets *bona fide* | **correct** | error (false rejection) |
| naïve impostors *bona fide* | error (false alarm) | **correct** |
| smart impostors *artificial (spoof)* | error (false alarm) | **correct** |

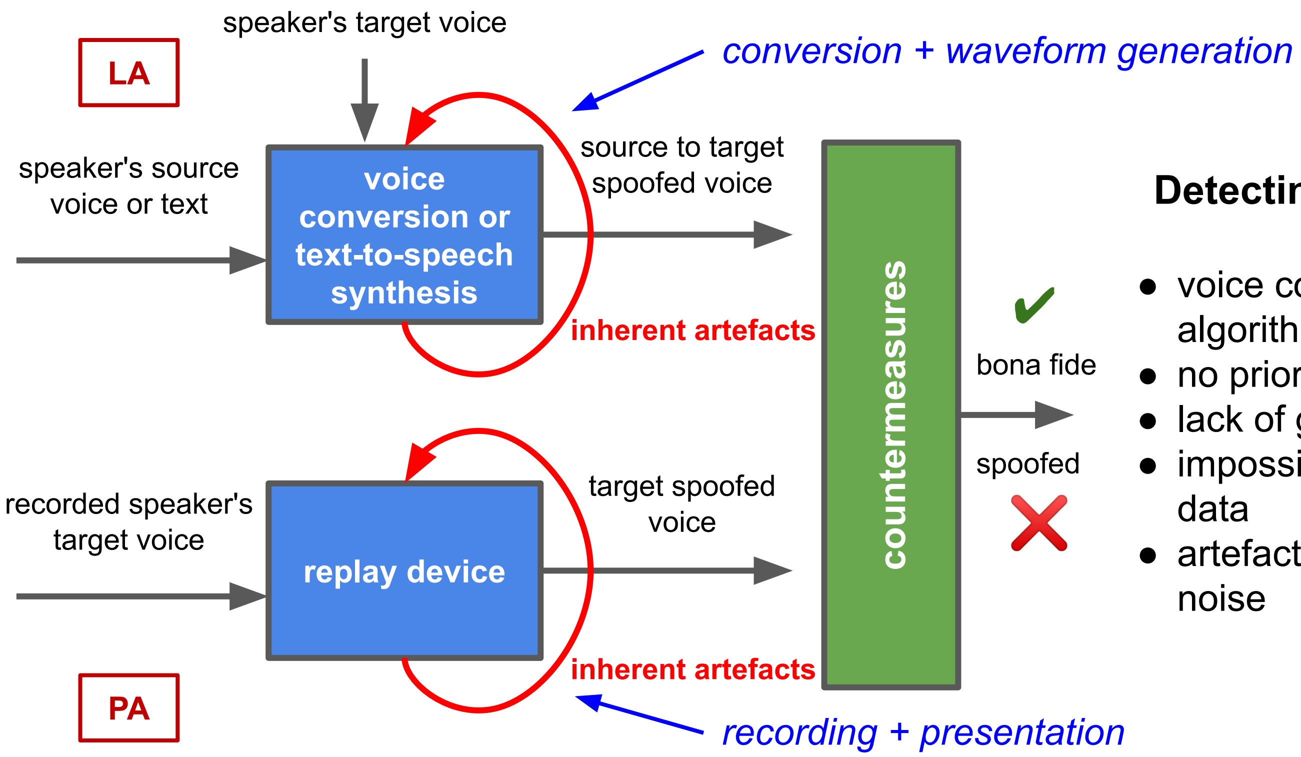
# Spoofing/presentation attacks




- persons masquerading as others in order to gain illegitimate access to sensitive or protected resources
- a.k.a. presentation attacks [ISO/IEC 30107-1:2016]

# Countermeasures to spoofing

- **Spoofing inevitably adds artefacts to the speech signal**



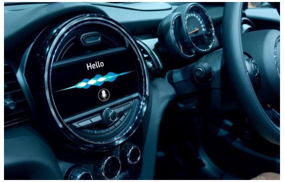

**Detecting artefacts is a difficult task**

- voice conversion or speech synthesis algorithms are continuously evolving
- no prior knowledge is given
- lack of generalisation
- impossible to collect all the spoofing data
- artefacts can be obfuscated by (real-life) noise

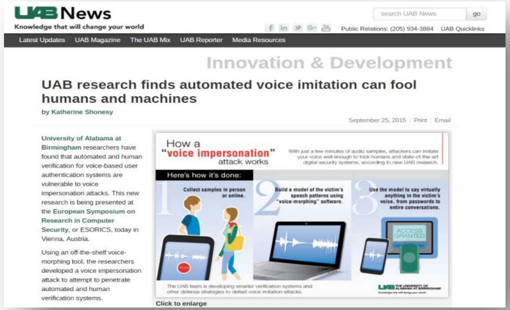
# Security in voice biometrics is becoming a necessity

## Voice-driven interactive services are everywhere today



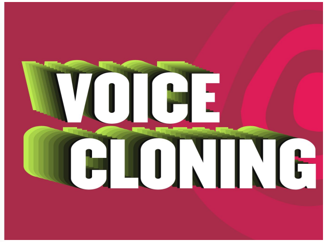




## In the eyes of the press


UAB News
Knowledge that will change your world
Innovation & Development
UAB research finds automated voice imitation can fool humans and machines
by Katherine Shonesy
September 25, 2015
University of Alabama at Birmingham researchers have found that automated and human verification for voice-based user authentication systems are vulnerable to voice impersonation attacks. This new research is being presented at the European Symposium on Research in Computer Security, or ESORICS, today in Vienna, Austria.
Using an off-the-shelf voice-morphing tool, the researchers developed a voice impersonation attack to attempt to penetrate automated and human verification systems.

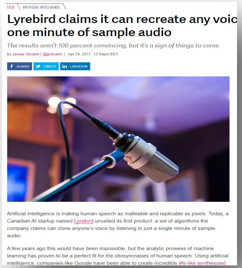
Click to enlarge


The Telegraph
Technology
This robot speech simulator can imitate anyone's voice
The machine has mimicked Barack Obama

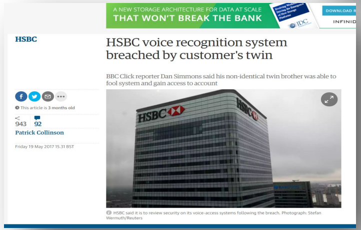




BBC NEWS
Technology
Adobe Voco 'Photoshop-for-voice' causes concern
A new application that promises to be the "Photoshop of speech" is raising ethical and security concerns.


Lyrebird claims it can recreate any voice using one minute of sample audio
The results aren't 100 percent convincing, but it's a sign of things to come


Google News Initiative partners with Google AI to help 'deep fake' audio detection research


HSBC
HSBC voice recognition system breached by customer's twin
BBC Click reporter Dan Simmons said his non-identical twin brother was able to fool system and gain access to account
Patrick Collinson


The Economist
Cloning voices
Imitating people's speech patterns precisely could bring trouble
You took the words right out of my mouth

# The ASVspoof challenge series

# History of the challenges

special session organised

ASVspoof 2015 LA

ASVspoof 2017 PA

ASVspoof 2019 LA & PA

ASVspoof 2021 LA, PA & DF

2013 2014 2015 2017 2019 2021

**small, purpose collected datasets**

**adapted, standard datasets**

**common datasets, metrics, protocols, LA scenario,**

**PA scenarios, broader attack variability**

**LA & PA scenarios, neural waveform models, simulated replay**

**LA, PA & DF scenarios, telephony encoding and transmission variability, real physical environments**

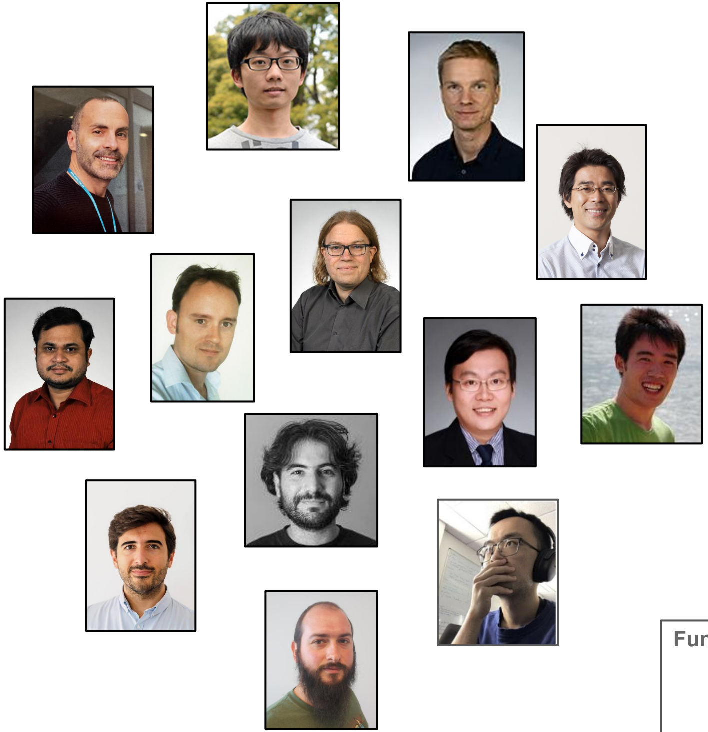
ASVspoof

Automatic Speaker Verification and Spoofing Countermeasures Challenge

https://www.asvspoof.org/

ASVspoof 2021 Workshop

# ASVspoof members (2015-2021)



ASVspoof

Automatic Speaker Verification and Spoofing Countermeasures Challenge

THE UNIVERSITY OF EDINBURGH

EURECOM Sophia Antipolis

UNIVERSITY OF EASTERN FINLAND

Institute for Infocomm Research

I²R

Inria informatics mathematics

NEC

NII 大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics

NUANCE

**Funding agencies**

ACADEMY OF FINLAND

AGENCE NATIONALE DE LA RECHERCHE ANR

JST Japan Science and Technology Agency

Agency for Science, Technology and Research SINGAPORE

# Logical access - a constantly growing threat

- **ASVspoof 2015 [1]**
    - 106 English speakers, disjoint in train / dev / eval sets
    - 10 TTS & VC methods, including known / unknown in eval set
- **ASVspoof 2019 LA [2]**
    - 107 English speakers, disjoint in train / dev / eval sets
    - 19 TTS & VC methods, including known / unknown in eval set
- **ASVspoof 2021 LA [3]**
    - same attacks as in 2019
    - addition of new speakers
    - addition of channel and transmission disturbance

ASVspoof
Automatic Speaker Verification and Spoofing Countermeasures Challenge

**Logical Access LA**

[1] Z. Wu et al., "ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge," in IEEE Journal of Selected Topics in Signal Processing, 2017.
[2] X. Wang et al., “ASVspoof 2019: a large-scale public database of synthesized, converted and replayed speech,” in Computer Speech and Language, 2020.
[3] X. Liu et al., “ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild,” under revision. https://arxiv.org/pdf/2210.02437.pdf

# Logical access - a constantly growing threat

- **ASVspoof 2015**
  - S10 → based on concatenation of speech units from a pre-recorded database

| | Subset | Waveform generation | Spoofing method | Feature representation | |
|---|---|---|---|---|---|
| train & dev | Genuine | None | None | N.A. | known algorithms |
| | S1 | STRAIGHT vocoder | Frame-selection voice conversion | Mel-cepstrum, Band aperiodicity, $F_0$ | |
| | S2 | STRAIGHT vocoder | Slope shifting voice conversion | Mel-cepstrum, Band aperiodicity, $F_0$ | |
| | S3 | STRAIGHT vocoder | HMM-based speech synthesis | Mel-cepstrum, Band aperiodicity, $F_0$ | |
| | S4 | STRAIGHT vocoder | HMM-based speech synthesis | Mel-cepstrum, Band aperiodicity, $F_0$ | |
| | S5 | MLSA vocoder | GMM-based voice conversion | Mel-cepstrum, $F_0$ | |
| eval | S6 | STRAIGHT vocoder | GMM-based voice conversion | Mel-cepstrum, Band aperiodicity, $F_0$ | unknown algorithms |
| | S7 | STRAIGHT vocoder | GMM-based voice conversion | Line spectrum pair, $F_0$ | |
| | S8 | STRAIGHT vocoder | Tensor-based voice conversion | Mel-cepstrum, Band aperiodicity, $F_0$ | |
| | S9 | STRAIGHT vocoder | KPLS-based voice conversion | Mel-cepstrum, Band aperiodicity, $F_0$ | |
| | S10 | Diphone concatenation | Unit selection-based speech synthesis | Waveform | |

# Logical access - a constantly growing threat

- **ASVspoof 2019 LA**

| | | Input | Input processor | Duration | Conversion | Speaker represent. | Outputs | Waveform generator | Post process |
|---|---|---|---|---|---|---|---|---|---|
| train & dev | A01 | Text | NLP | HMM | AR RNN* | VAE* | MCC, F0 | WaveNet* | |
| | A02 | Text | NLP | HMM | AR RNN* | VAE* | MCC, F0, BAP | WORLD | |
| | A03 | Text | NLP | FF* | FF* | One hot embed. | MCC, F0, BAP | WORLD | |
| | A04 | Text | NLP | - | CART | - | MFCC, F0 | Waveform concat. | |
| | A05 | Speech (human) | WORLD | - | VAE* | One hot embed. | MCC, F0, AP | WORLD | |
| | A06 | Speech (human) | LPCC/MFCC | - | GMM-UBM | - | LPC | Spectral filtering + OLA | |
| eval | A07 | Text | NLP | RNN* | RNN* | One hot embed. | MCC, F0, BA | WORLD | GAN* |
| | A08 | Text | NLP | HMM | AR RNN* | One hot embed. | MCC, F0 | Neural source-filter* | |
| | A09 | Text | NLP | RNN* | RNN* | One hot embed. | MCC, F0 | Vocaine | |
| | A10 | Text | CNN+bi-RNN* | Attention* | AR RNN + CNN* | d-vector (RNN)* | Mel-spectrograms | WaveRNN* | |
| | A11 | Text | CNN+bi-RNN* | Attention* | AR RNN + CNN* | d-vector (RNN)* | Mel-spectrograms | Griffin-Lim [13] | |
| | A12 | Text | NLP | RNN* | RNN* | One hot embed. | F0+linguistic features | WaveNet* | |
| | A13 | Speech (TTS) | WORLD | DTW | Moment matching* | - | MCC | Waveform filtering | |
| | A14 | Speech (TTS) | ASR* | - | RNN* | - | MCC, F0, BAP | STRAIGHT | |
| | A15 | Speech (TTS) | ASR* | - | RNN* | - | MCC, F0 | WaveNet* | |
| | A16 | Text | NLP | - | CART | - | MFCC, F0 | Waveform concat. | |
| | A17 | Speech (human) | WORLD | - | VAE* | One hot embed. | MCC, F0 | Waveform filtering | |
| | A18 | Speech (human) | MFCC/i-vector | - | Linear | PLDA | MFCC | MFCC vocoder | |
| | A19 | Speech (human) | LPCC/MFCC | - | GMM-UBM | - | LPC | Spectral filtering + OLA | |

# Logical access - a constantly growing threat

- **ASVspoof 2019 LA**

| | | Input | Input processor | Duration | Conversion | Speaker represent. | Outputs | Waveform generator | Post process |
|---|---|---|---|---|---|---|---|---|---|
| train & dev | A01 | Text | NLP | HMM | AR RNN* | VAE* | MCC, F0 | WaveNet* | |
| | A02 | Text | NLP | HMM | AR RNN* | VAE* | MCC, F0, BAP | WORLD | |
| | A03 | Text | NLP | FF* | FF* | One hot embed. | MCC, F0, BAP | WORLD | |
| | A04 | Text | NLP | - | CART | - | MFCC, F0 | Waveform concat. | |
| | A05 | Speech (human) | WORLD | - | VAE* | One hot embed. | MCC, F0, AP | WORLD | |
| | A06 | Speech (human) | LPCC/MFCC | - | GMM-UBM | - | LPC | Spectral filtering + OLA | |
| eval | A07 | Text | NLP | RNN* | RNN* | One hot embed. | MCC, F0, BA | WORLD | GAN* |
| | A08 | Text | NLP | HMM | AR RNN* | One hot embed. | MCC, F0 | Neural source-filter* | |
| | A09 | Text | NLP | RNN* | RNN* | One hot embed. | MCC, F0 | Vocaine | |
| | A10 | Text | CNN+bi-RNN* | Attention* | AR RNN + CNN* | d-vector (RNN)* | Mel-spectrograms | WaveRNN* | |
| | A11 | Text | CNN+bi-RNN* | Attention* | AR RNN + CNN* | d-vector (RNN)* | Mel-spectrograms | Griffin-Lim [13] | |
| | A12 | Text | NLP | RNN* | RNN* | One hot embed. | F0+linguistic features | WaveNet* | |
| | A13 | Speech (TTS) | WORLD | DTW | Moment matching* | - | MCC | Waveform filtering | |
| | A14 | Speech (TTS) | ASR* | - | RNN* | - | MCC, F0, BAP | STRAIGHT | |
| | A15 | Speech (TTS) | ASR* | - | RNN* | - | MCC, F0 | WaveNet* | |
| | A16 | Text | NLP | - | CART | - | MFCC, F0 | Waveform concat. | |
| | A17 | Speech (human) | WORLD | - | VAE* | One hot embed. | MCC, F0 | Waveform filtering | |
| | A18 | Speech (human) | MFCC/i-vector | - | Linear | PLDA | MFCC | MFCC vocoder | |
| | A19 | Speech (human) | LPCC/MFCC | - | GMM-UBM | - | LPC | Spectral filtering + OLA | |

same algorithms

# Logical access - a constantly growing threat

- **ASVspoof 2019 LA**

| | Input | Input processor | Duration | Conversion | Speaker represent. | Outputs | Waveform generator | Post process |
|---|---|---|---|---|---|---|---|---|
| A01 | Text | NLP | HMM | AR RNN* | VAE* | MCC, F0 | WaveNet* | |
| A02 | Text | NLP | HMM | AR RNN* | VAE* | MCC, F0, BAP | WORLD | |
| A03 | Text | NLP | FF* | FF* | One hot embed. | MCC, F0, BAP | WORLD | |
| A04 | Text | NLP | - | CART | - | MFCC, F0 | Waveform concat. | |
| A05 | Speech (human) | WORLD | - | VAE* | One hot embed. | MCC, F0, AP | WORLD | |
| A06 | Speech (human) | LPCC/MFCC | - | GMM-UBM | - | LPC | Spectral filtering + OLA | |
| A07 | Text | NLP | RNN* | RNN* | One hot embed. | MCC, F0, BA | WORLD | GAN* |
| A08 | Text | NLP | HMM | AR RNN* | One hot embed. | MCC, F0 | Neural source-filter* | |
| A09 | Text | NLP | RNN* | RNN* | One hot embed. | MCC, F0 | Vocaine | |
| A10 | Text | CNN+bi-RNN* | Attention* | AR RNN + CNN* | d-vector (RNN)* | Mel-spectrograms | WaveRNN* | |
| A11 | Text | CNN+bi-RNN* | Attention* | AR RNN + CNN* | d-vector (RNN)* | Mel-spectrograms | Griffin-Lim [13] | |
| A12 | Text | NLP | RNN* | RNN* | One hot embed. | F0+linguistic features | WaveNet* | |
| A13 | Speech (TTS) | WORLD | DTW | Moment matching* | - | MCC | Waveform filtering | |
| A14 | Speech (TTS) | ASR* | - | RNN* | - | MCC, F0, BAP | STRAIGHT | |
| A15 | Speech (TTS) | ASR* | - | RNN* | - | MCC, F0 | WaveNet* | |
| A16 | Text | NLP | - | CART | - | MFCC, F0 | Waveform concat. | |
| A17 | Speech (human) | WORLD | - | VAE* | One hot embed. | MCC, F0 | Waveform filtering | |
| A18 | Speech (human) | MFCC/i-vector | - | Linear | PLDA | MFCC | MFCC vocoder | |
| A19 | Speech (human) | LPCC/MFCC | - | GMM-UBM | - | LPC | Spectral filtering + OLA | |

train & dev

eval

varied or improved algorithms

# Logical access - a constantly growing threat

- **ASVspoof 2019 LA**

train & dev

eval

| | Input | Input processor | Duration | Conversion | Speaker represent. | Outputs | Waveform generator | Post process |
|---|---|---|---|---|---|---|---|---|
| A01 | Text | NLP | HMM | AR RNN* | VAE* | MCC, F0 | WaveNet* | |
| A02 | Text | NLP | HMM | AR RNN* | VAE* | MCC, F0, BAP | WORLD | |
| A03 | Text | NLP | FF* | FF* | One hot embed. | MCC, F0, BAP | WORLD | |
| A04 | Text | NLP | - | CART | - | MFCC, F0 | Waveform concat. | |
| A05 | Speech (human) | WORLD | - | VAE* | One hot embed. | MCC, F0, AP | WORLD | |
| A06 | Speech (human) | LPCC/MFCC | - | GMM-UBM | - | LPC | Spectral filtering + OLA | |
| A07 | Text | NLP | RNN* | RNN* | One hot embed. | MCC, F0, BA | WORLD | GAN* |
| A08 | Text | NLP | HMM | AR RNN* | One hot embed. | MCC, F0 | Neural source-filter* | |
| A09 | Text | NLP | RNN* | RNN* | One hot embed. | MCC, F0 | Vocaine | |
| A10 | Text | CNN+bi-RNN* | Attention* | AR RNN + CNN* | d-vector (RNN)* | Mel-spectrograms | WaveRNN* | |
| A11 | Text | CNN+bi-RNN* | Attention* | AR RNN + CNN* | d-vector (RNN)* | Mel-spectrograms | Griffin-Lim [13] | |
| A12 | Text | NLP | RNN* | RNN* | One hot embed. | F0+linguistic features | WaveNet* | |
| A13 | Speech (TTS) | WORLD | DTW | Moment matching* | - | MCC | Waveform filtering | |
| A14 | Speech (TTS) | ASR* | - | RNN* | - | MCC, F0, BAP | STRAIGHT | |
| A15 | Speech (TTS) | ASR* | - | RNN* | - | MCC, F0 | WaveNet* | |
| A16 | Text | NLP | - | CART | - | MFCC, F0 | Waveform concat. | |
| A17 | Speech (human) | WORLD | - | VAE* | One hot embed. | MCC, F0 | Waveform filtering | |
| A18 | Speech (human) | MFCC/i-vector | - | Linear | PLDA | MFCC | MFCC vocoder | |
| A19 | Speech (human) | LPCC/MFCC | - | GMM-UBM | - | LPC | Spectral filtering + OLA | |

unknown algorithms

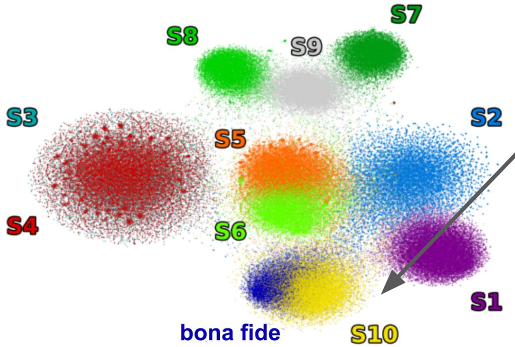
# ASVspoof 2019 LA - data providers

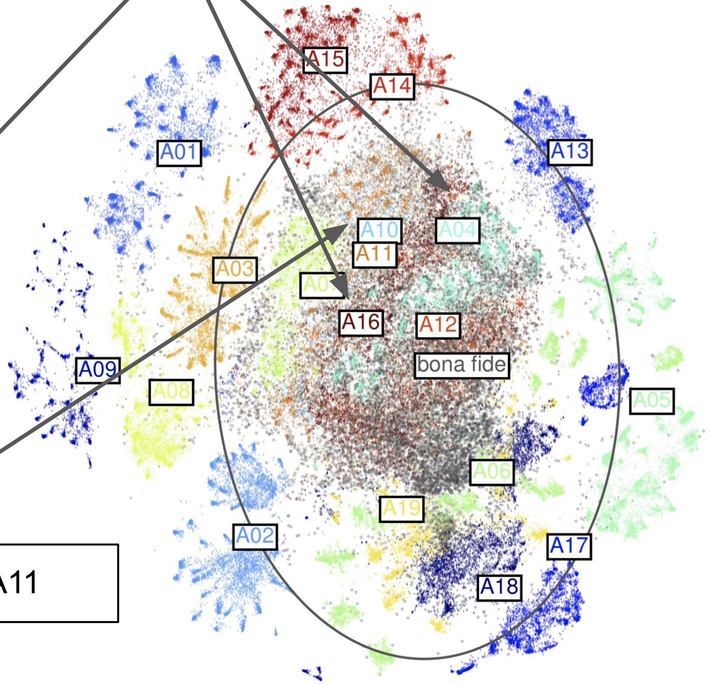
Aalto University
ACADEMIA SINICA 1928
ADAPT Engaging Content Engaging People
DFKI
EURECOM Sophia Antipolis
Google
Google AI
Google News Initiative
HOYA
科大讯飞 iFLYTEK
NAGOYA UNIVERSITY
NII Inter-University Research Institute Corporation / Research Organization of Information and Systems National Institute of Informatics
NTT
Trinity College Dublin Coláiste na Tríonóide, Baile Átha Cliath The University of Dublin
UNIVERSITÄT DES SAARLANDES
UNIVERSITY OF EASTERN FINLAND
UNIVERSITÉ D'AVIGNON
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE
THE UNIVERSITY of EDINBURGH
中国科学技术大学 University of Science and Technology of China

# Attack-wise speaker clustering

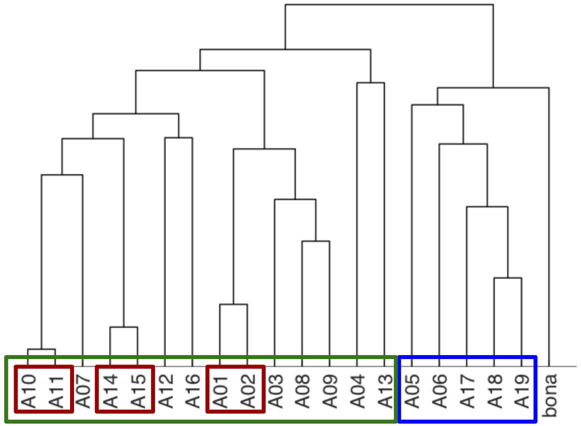
## ASVspoof 2015

unit selection S10, A16 and A04




TACOTRON A10 and A11




TTS systems

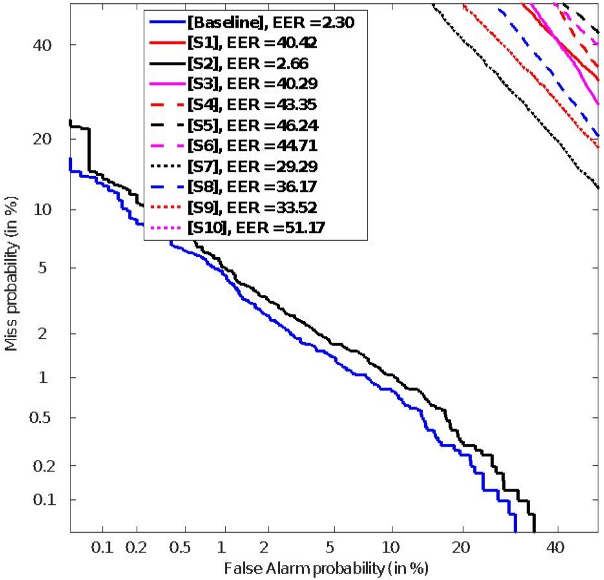
VC systems using human speech source data

same TTS architecture

## ASVspoof 2019 LA

# ASVspoof 2015 assessment & results

- **ASV vulnerability and primary submission results for the ASVspoof 2015 challenge**
  - the best EER for S10 is 8.49%



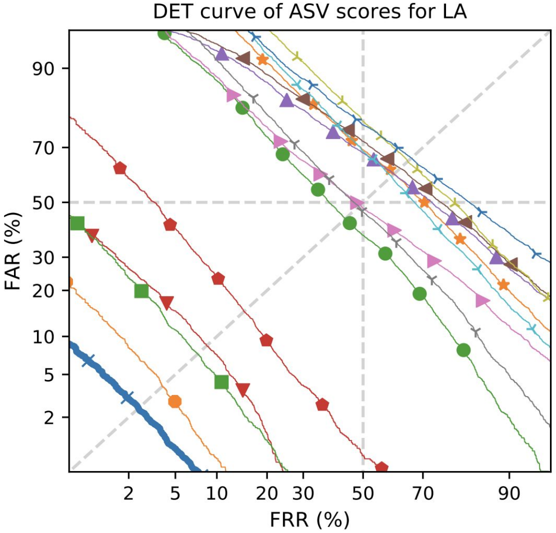

| System ID | Average Equal Error Rates (EERs) [%] | | | | |
|---|---|---|---|---|---|
| | Known | Unknown | | | All |
| | AVG S1-S5 | AVG S6-S9 | S10 | AVG S6-S10 | AVG |
| A | 0.408 | 0.394 | 8.490 | **2.013** | **1.211** |
| B | 0.008 | 0.009 | 19.571 | 3.922 | 1.965 |
| C | 0.058 | 0.098 | 24.601 | 4.998 | 2.528 |
| D | **0.003** | 0.003 | 26.142 | 5.231 | 2.617 |
| E | 0.041 | 0.085 | 26.393 | 5.347 | 2.694 |
| F | 0.358 | 0.453 | 28.581 | 6.078 | 3.218 |
| G | 0.405 | 0.304 | 30.021 | 6.247 | 3.326 |
| H | 0.670 | 0.042 | 37.068 | 6.041 | 3.355 |
| I | 0.005 | 0.839 | 32.651 | 7.447 | 3.726 |
| J | 0.025 | 0.033 | 40.708 | 8.168 | 4.097 |
| K | 0.210 | 0.195 | 43.638 | 8.883 | 4.547 |
| L | 0.412 | 7.310 | 35.890 | 13.026 | 6.719 |
| M | 8.528 | 17.423 | 31.574 | 20.253 | 14.391 |
| N | 7.874 | 15.580 | 43.991 | 21.262 | 14.568 |
| O | 17.723 | 14.532 | 41.519 | 19.929 | 18.826 |
| P | 21.206 | 15.763 | 46.102 | 21.831 | 21.518 |

[.] Z. Wu et al., "ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge," in IEEE Journal of Selected Topics in Signal Processing, 2017.

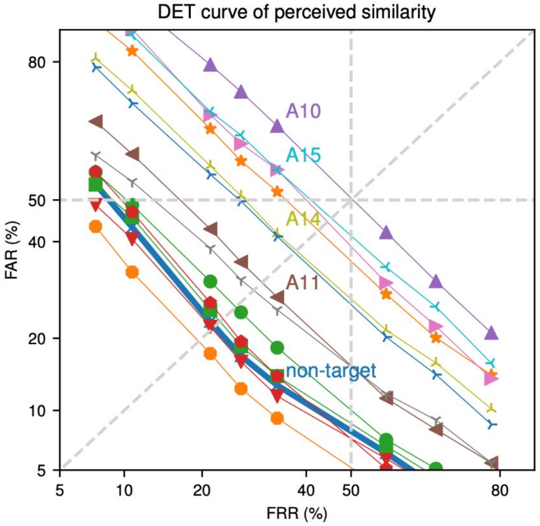
# ASVspoof 2019 LA assessment

- **ASV vulnerability vs human assessment [1]**
  - A10 (TACOTRON) is perceived with very **good** quality and very **similar** to the target
  - A17 (VC) is perceived with **bad** quality and very **different** from the target
  - similarly, A17 does **NOT** foul the ASV

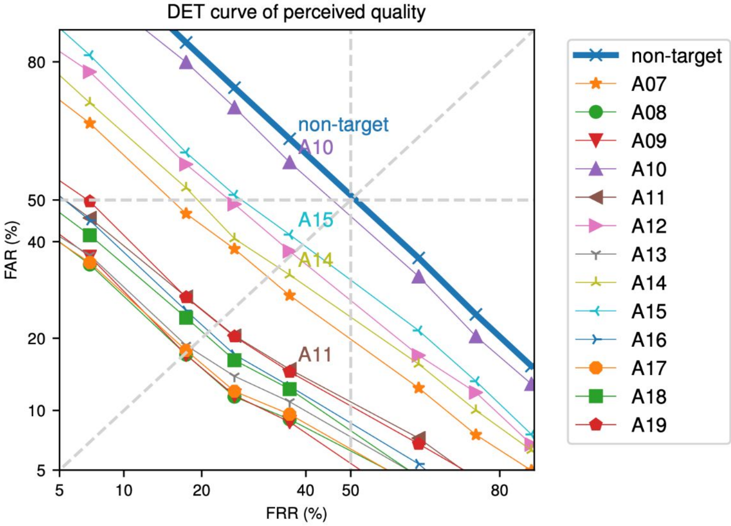
**ASV**




**human SV**




**human PQ**

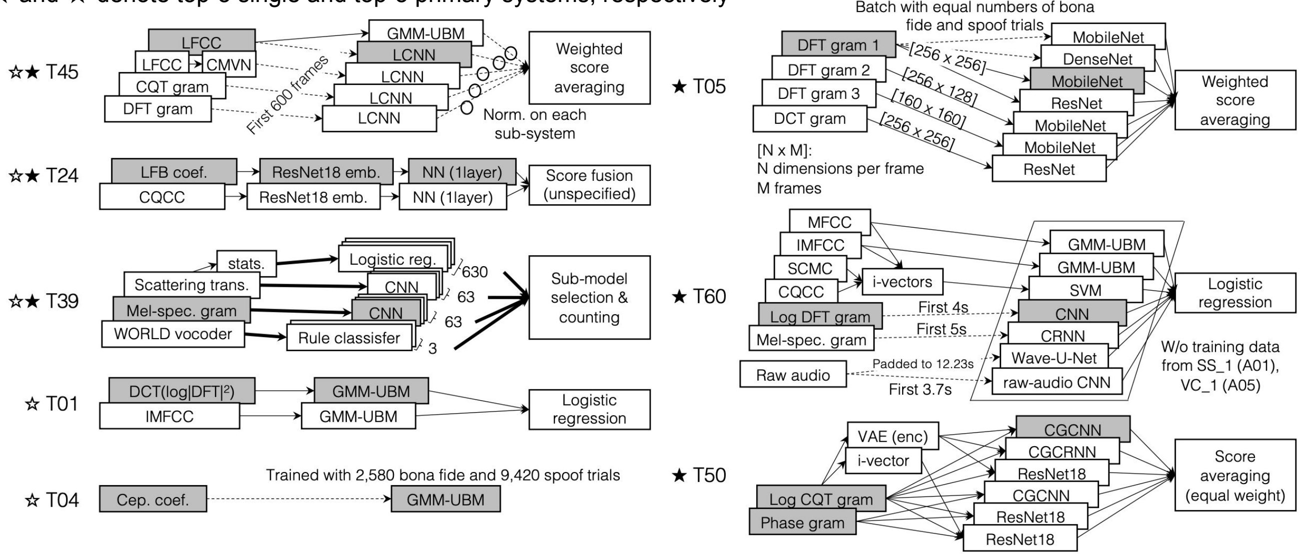



[1] Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi Kinnunen, Ville Vestman, Massimiliano Todisco, Hector Delgado, Md Sahidullah, Junichi Yamagishi, Kong Aik Lee, "ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech", IEEE Transactions on Biometrics, Behavior, and Identity Science

# ASVspoof 2019 LA submitted CM systems & results

- **Top-5 single and primary system submissions**
  - single systems (grey blocks) perform poorly
  - primary systems perform well → drawback: consist of fusion of many system
  - A17 the worst attack → not detectable by CMs

☆ and ★ denote top-5 single and top-5 primary systems, respectively




(a) Single systems

| TID | min t-DCF | EER [%] | Max min t-DCF (AID) |
|---|---|---|---|
| T45 | 0.1562 | 5.06 | 0.9905 (A17) |
| T24 | 0.1655 | 4.04 | 0.8499 (A17) |
| T39 | 0.1894 | 7.01 | 1.000 (A17) |
| T01 | 0.1937 | 5.97 | 0.7667 (A17) |
| T04 | 0.1939 | 5.74 | 0.7837 (A17) |
| B01 | 0.2839 | 9.57 | 0.9901 (A17) |
| B02 | 0.2605 | 8.09 | 0.6571 (A17) |
| Perfect | 0.0627 | 0.0 | 0.4218 (A17) |

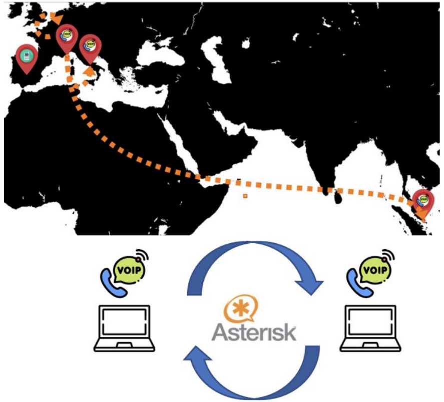
(b) Primary systems

| TID | min t-DCF | EER [%] | Max min t-DCF (AID) |
|---|---|---|---|
| T05 | 0.0692 | 0.22 | 0.4418 (A17) |
| T45 | 0.1104 | 1.86 | 0.7778 (A17) |
| T60 | 0.1331 | 2.64 | 0.8803 (A17) |
| T24 | 0.1518 | 3.45 | 0.8546 (A17) |
| T50 | 0.1671 | 3.56 | 0.8471 (A17) |

[.] Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi Kinnunen, Ville Vestman, Massimiliano Todisco, Hector Delgado, Md Sahidullah, Junichi Yamagishi, Kong Aik Lee, "ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech", IEEE Transactions on Biometrics, Behavior, and Identity Science

# ASVspoof 2021 LA spoof detection in the wild

- Data is transmitted across telephony or VoIP networks with various coding and transmission effects
- Variability resulting from encoding, transmission, distortion of devices

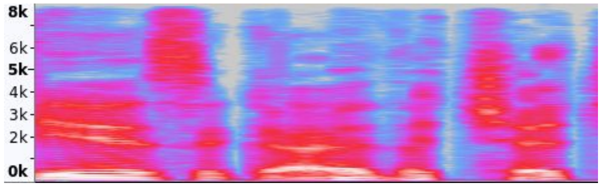



- varied codecs
- actual VoIP channel
- actual PSTN channel

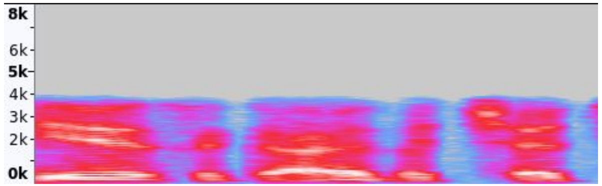
unknown codecs

| Cond. | Codec | Sampling rate | Transmission | Bitrate |
|---|---|---|---|---|
| LA-C1 | - | 16 kHz | - | 250 kbps |
| LA-C2 | a-law | 8 kHz | VoIP | 64 kbps |
| LA-C3 | unk. + $\mu$-law | 8 kHz | PSTN + VoIP | - / 64 kbps |
| LA-C4 | G.722 | 16 kHz | VoIP | 64 kbps |
| LA-C5 | $\mu$-law | 8 kHz | VoIP | 64 kbps |
| LA-C6 | GSM | 8 kHz | VoIP | 13 kbps |
| LA-C7 | OPUS | 16 kHz | VoIP | VBR 16 kbps |

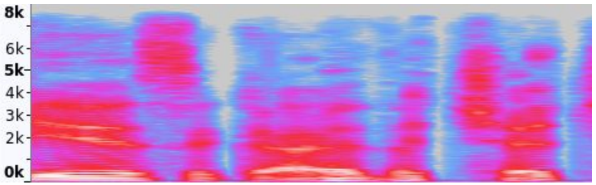
none




PSTN

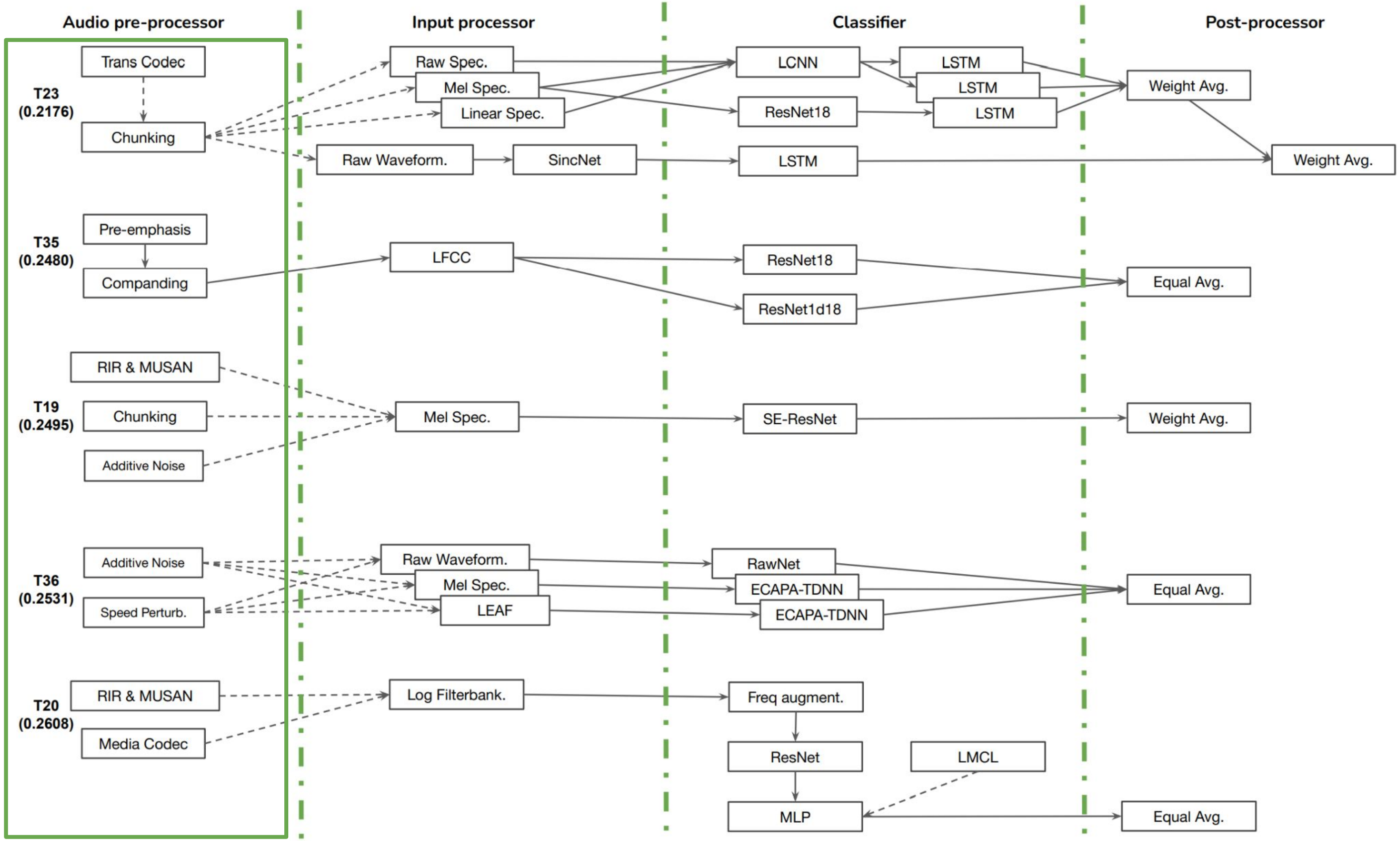



OPUS




[.] X. Liu et al., “ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild,” under revision. https://arxiv.org/pdf/2210.02437.pdf

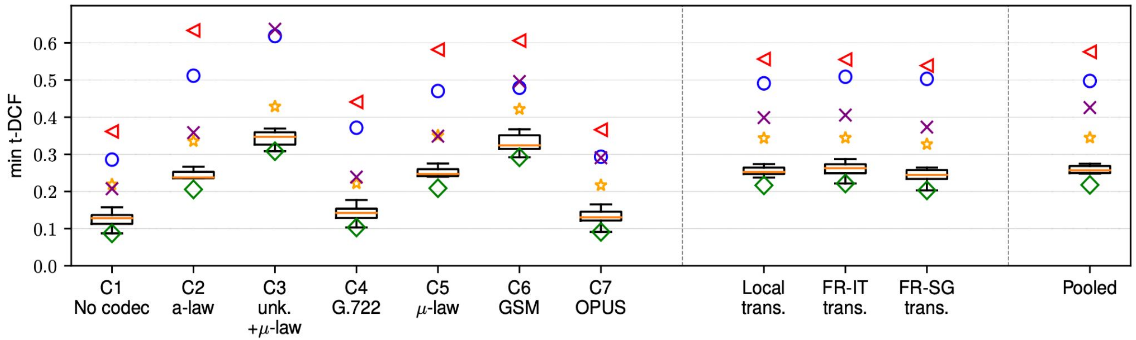
# ASVspoof 2021 LA submitted CM systems & results



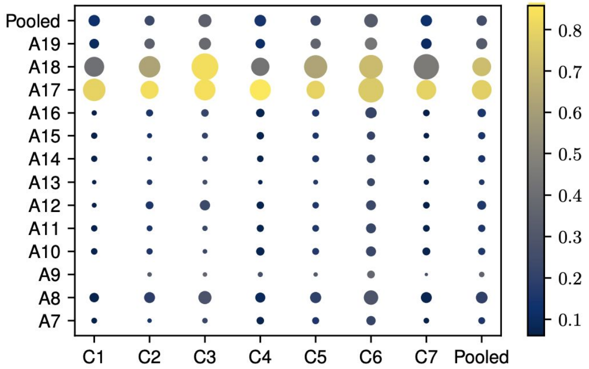

| | | Evaluation set | |
|---|---|---|---|
| # | ID | t-DCF | EER |
| 1 | T23 | 0.2176 | 1.32 |
| 2 | T35 | 0.2480 | 2.77 |
| 3 | T19 | 0.2495 | 3.13 |
| 4 | | 0.2500 | 2.81 |
| 5 | T36 | 0.2531 | 3.10 |
| 6 | T20 | 0.2608 | 3.21 |
| 7 | T08 | 0.2672 | 3.62 |
| 8 | T16 | 0.2689 | 3.63 |
| 9 | | 0.2725 | 3.61 |
| 10 | T04 | 0.2747 | 5.58 |
| 11 | T06 | 0.2853 | 5.66 |
| 12 | | 0.2880 | 5.01 |
| 13 | T03 | 0.2882 | 4.66 |
| 14 | | 0.2893 | 5.70 |
| 15 | T31 | 0.3094 | 5.46 |
| 16 | T17 | 0.3279 | 7.19 |
| 17 | T07 | 0.3310 | 8.23 |
| 18 | T30 | 0.3362 | 8.89 |
| 19 | B03 | 0.3445 | 9.26 |
| 20 | T02 | 0.3445 | 7.79 |
| 21 | T14 | 0.3451 | 8.98 |
| 22 | T11 | 0.3666 | 7.19 |
| 23 | T34 | 0.4059 | 13.45 |
| 24 | B04 | 0.4257 | 9.50 |

# ASVspoof 2021 LA results by disturbance

- **Top-10 system submissions submissions decomposed over different factors**
  - (no codec, G.722, OPUS) is lower than for narrowband conditions (a-law, PSTN, u-law and GSM) → importance of information at higher frequencies
  - among the narrowband conditions, lower bit rates and uncontrolled transmission (GSM and unk. + μlaw) lead to worse performance
  - transmission routes have little impact upon CM performance
  - A17 remains the worst attack with A18




top-1 submission (◇), B01 (o), B02 (◁), B03 (☆), and B04 (×).




[.] Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi Kinnunen, Ville Vestman, Massimiliano Todisco, Hector Delgado, Md Sahidullah, Junichi Yamagishi, Kong Aik Lee, "ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech", IEEE Transactions on Biometrics, Behavior, and Identity Science

# Top system architecture among the challenges

- **2015 → 2019**
  - *pre-processor* → no substantial difference; no pre-processing
  - *features* → from a compact (MFCC) to complete representation (STFT); exploring other compact features (LFCC, CQCC, IMFCC)
  - *classifiers* → from GMM towards deeper classifiers (ResNet, CNN)
  - *post-processor* → no substantial difference; normalisation and score average fusion

| | **ASVspoof 2015** | **ASVspoof 2019 LA** |
|---|---|---|
| ***pre-processor*** | max_norm none pre-emphasis | vocoder none |
| ***features*** | i-vectors RLMS MEL MGD MFCC STFT LBP LPC Wavelet instantaneous_frequency CFCC LFCC CFCCIF phase | MEL i-vectors VAE CQT CQCC STFT phase LFCC LFB MFCC DCT IMFCC SCATTER SCMC WAVEFORM |
| ***classifiers*** | MLP GMM linear_regression SVM GMM-UBM deep_belief_network | GMM-UBM ResNet WaveUNet CRNN linear_regression DenseNet SVM RC CNN MobileNet LCNN CGCNN MLP |
| ***post-processor*** | score_average mahal_dist SVM PCA CMVN CMN grid_search none | score_average CMVN none sub-model_selection linear_regression |

# Top system architecture among the challenges

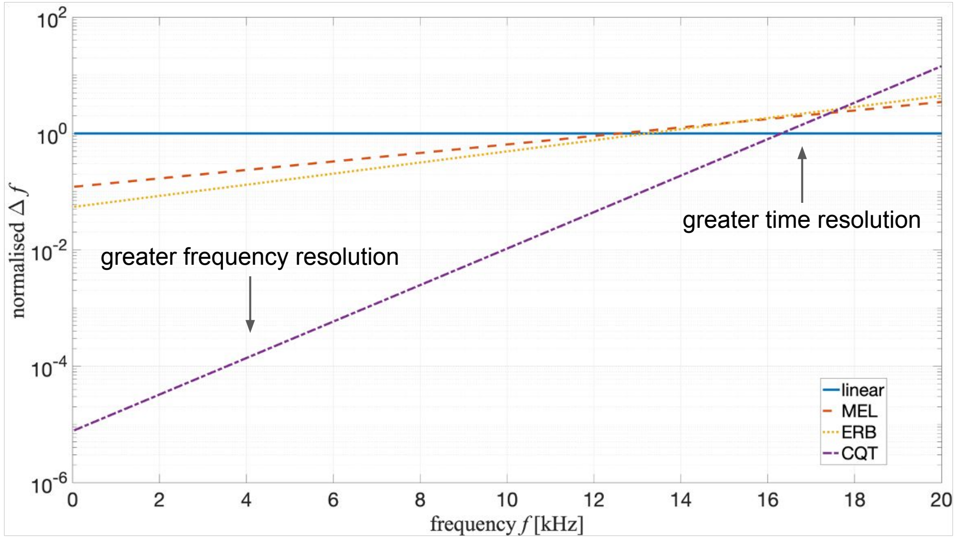
- **2019 → 2021**
  - *pre-processor* → very substantial difference; toward data augmentation and signal pre-processing
  - *features* → from a complete (STFT) to a more auditory-based representation (MEL) and time-domain waveform
  - *classifiers* → no substantial difference;  addition of some state-of-the-art systems for ASV (ECAPA-TDNN)
  - *post-processor* → no substantial difference; normalisation and score average fusion

| | **ASVspoof 2019 LA** | **ASVspoof 2021 LA** |
|---|---|---|
| ***pre-processor*** | vocoder none | CODECS additive_noise COMPAND time_perturbation convolutive_noise |
| ***features*** | MEL CQT SCATTER i-vectors CQCC SCMC VAE STFT LFB MFCC phase DCT WAVEFORM LFCC IMFCC | WAVEFORM SincNet LFCC MEL STFT LEAF |
| ***classifiers*** | GMM-UBM CGCNN WaveUNet CRNN ResNet linear_regression DenseNet SVM RC CNN MLP MobileNet LCNN | LSTM MLP RawNet ResNet LCNN ECAPA-TDNN |
| ***post-processor*** | score_average CMVN none sub-model_selection linear_regression | score_weighted_average score_average |

# Voice cloning artefacts
## a recent history of detection and explainability

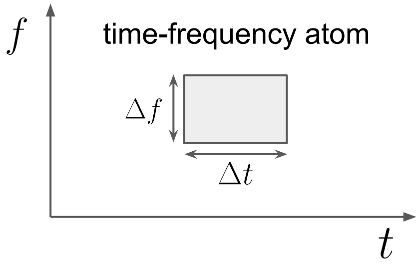
# The constant Q cepstral coefficients (CQCCs)

- **Can we explain why the CQCC [1] front-end performs well for certain attacks?**
  - based on Constant Q transform (CQT) [2]
    - reflect more closely human perception
    - humans do not perceive frequencies on a linear scale




uncertainty principle

$$\Delta f \cdot \Delta t \geq \frac{1}{2\pi}$$

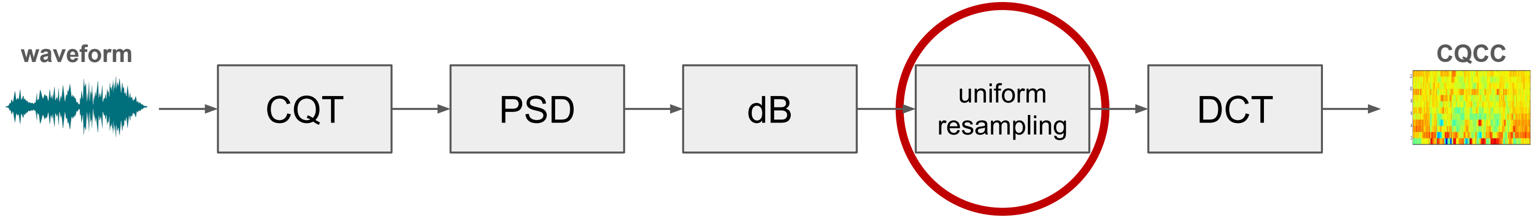




[1] M. Todisco, H. Delgado, N. Evans, “Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification,” Computer Speech & Language, 2017.
[2] J. Brown, “Calculation of a constant Q spectral transform,” Journal of the Acoustical Society of America, vol. 89, no. 1, pp. 425– 434, January 1991.

# The constant Q cepstral coefficients (CQCCs)

- **CQCCs pipeline**
  - cepstral computation → CQT and DCT have different scale (geometric vs linear)
  - uniform resample → equal weighting to information across the full spectrum



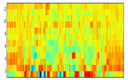
# The constant Q cepstral coefficients (CQCCs)

- **CQCCs pipeline**
  - cepstral computation → CQT and DCT have different scale (geometric vs linear)
  - uniform resample → equal weighting to information across the full spectrum

waveform

→ CQT → PSD → dB → uniform resampling → DCT →

CQCC

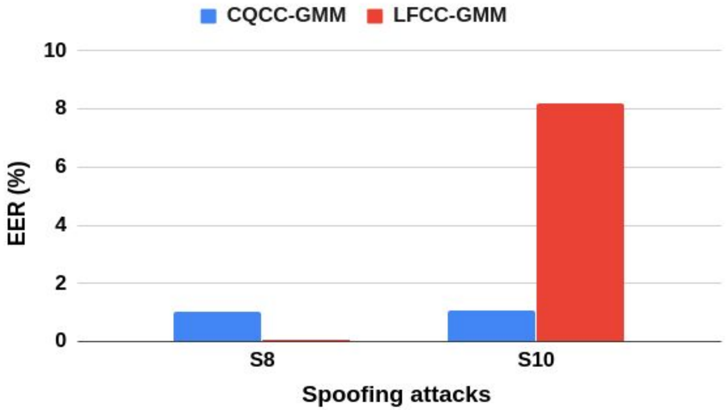


| | | | | |
|---|---|---|---|---|
| **different** | **same** | **same** | **no need** | **same** |
| STFT | = | = | ❌ | = |

**Linear frequency cepstral coefficients (LFCC)**

# CQCCs vs LFCCs

- **ASVspoof 2015 database**
  - substantial variations in the performance
    - CQCC-GMM best detected: S10
    - LFCC-GMM best detected: S8

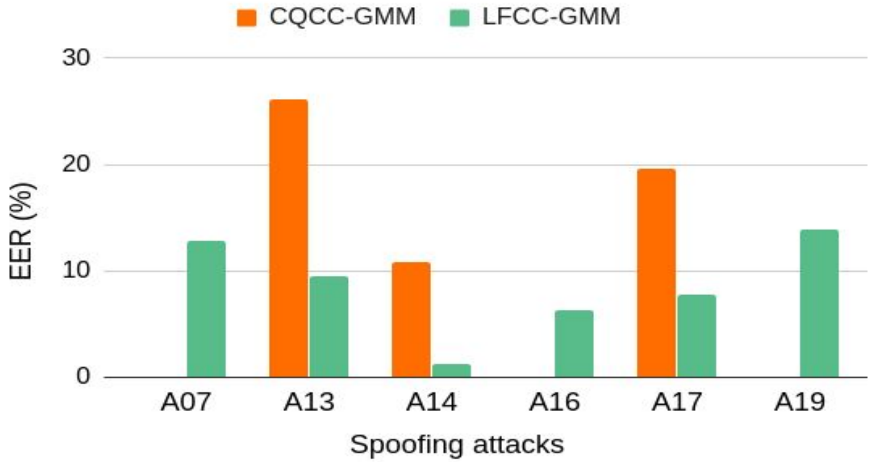



| System | S8 | S10 |
|---|---|---|
| GMM-CQCC | 1.033 | 1.065 |
| GMM-LFCC | 0.074 | 8.185 |

# CQCCs vs LFCCs

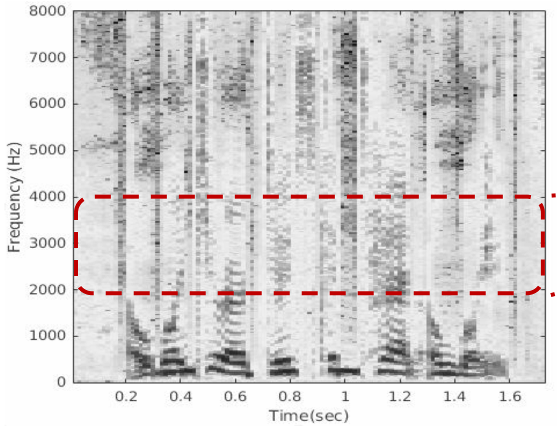
- **ASVspoof 2019 LA database**
  - substantial variations in the performance
    - CQCC-GMM best detected: A07, A16 and A19
    - LFCC-GMM best detected: A13, A14 and A17

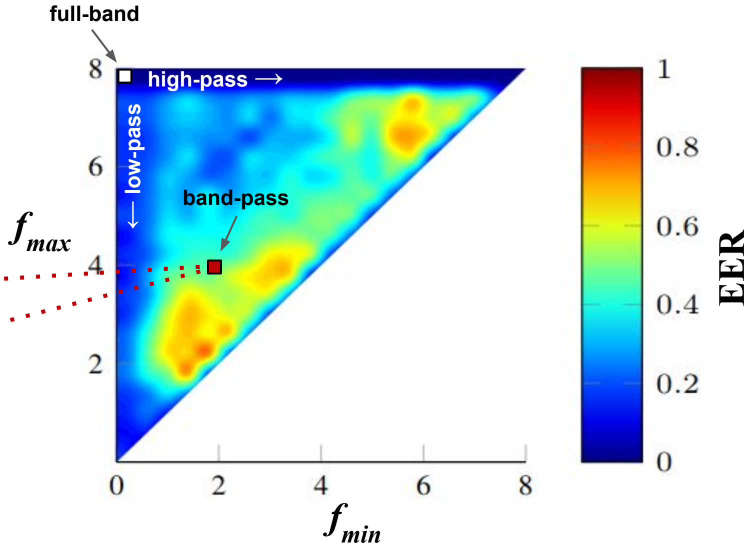
| System | **A07** | A08 | A09 | A10 | A11 | A12 | **A13** | **A14** | A15 | **A16** | **A17** | A18 | **A19** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GMM-CQCC | **0.00** | 0.04 | 0.14 | 15.16 | 0.08 | 4.74 | **26.15** | **10.85** | 1.26 | **0.00** | **19.62** | 3.81 | **0.04** |
| GMM-LFCC | **12.86** | 0.37 | 0.00 | 18.97 | 0.12 | 4.92 | **9.57** | **1.22** | 2.22 | **6.31** | **7.71** | 3.58 | **13.94** |



# Sub-band analysis: a method for explainability [1]

- **Research hypotheses**
  - spoofing artefacts can be localised in the spectrum, e.g. high-band, mid-band or low-band
  - cepstral analysis smooths information across the full band and dilutes localised information
  - more reliable detection with features that emphasize information at the sub-band level
- **2D heat map representation**
  - sub-band dependence on EER

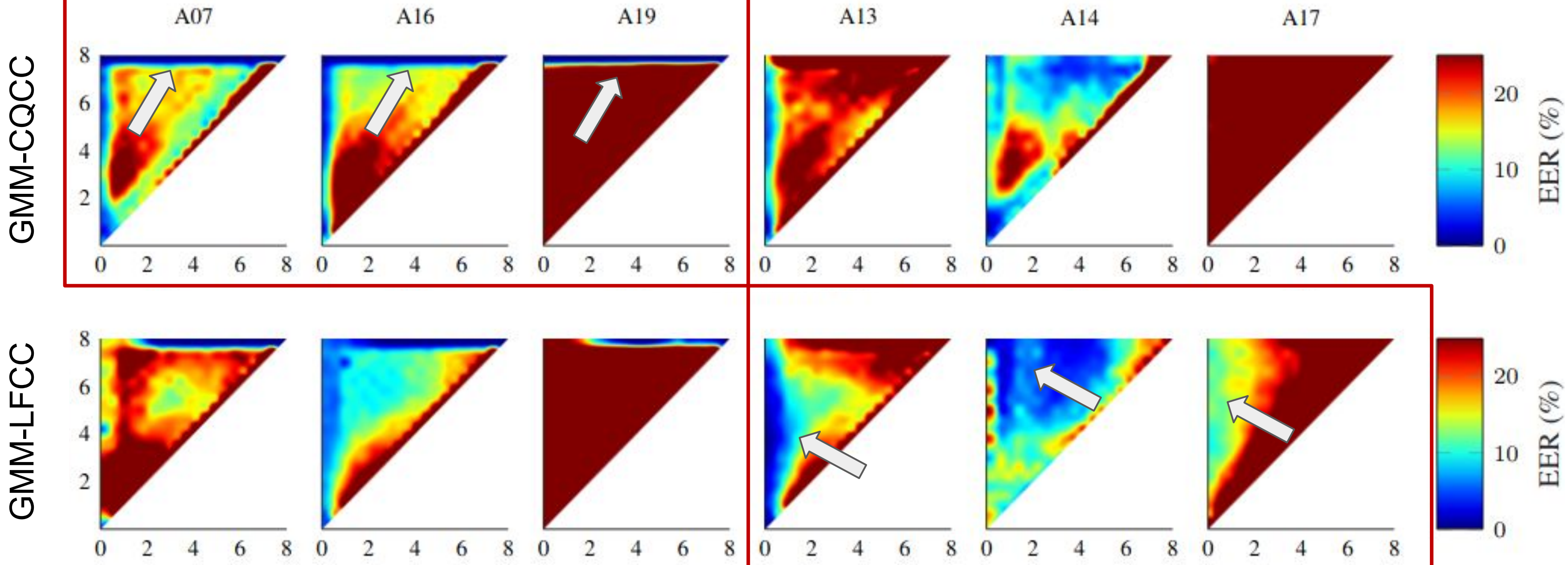



[1] H. Tak et al., "An explainability study of the constant Q cepstral coefficient spoofing countermeasure for automatic speaker verification," in Proc. Speaker Odyssey Workshop, 2020.

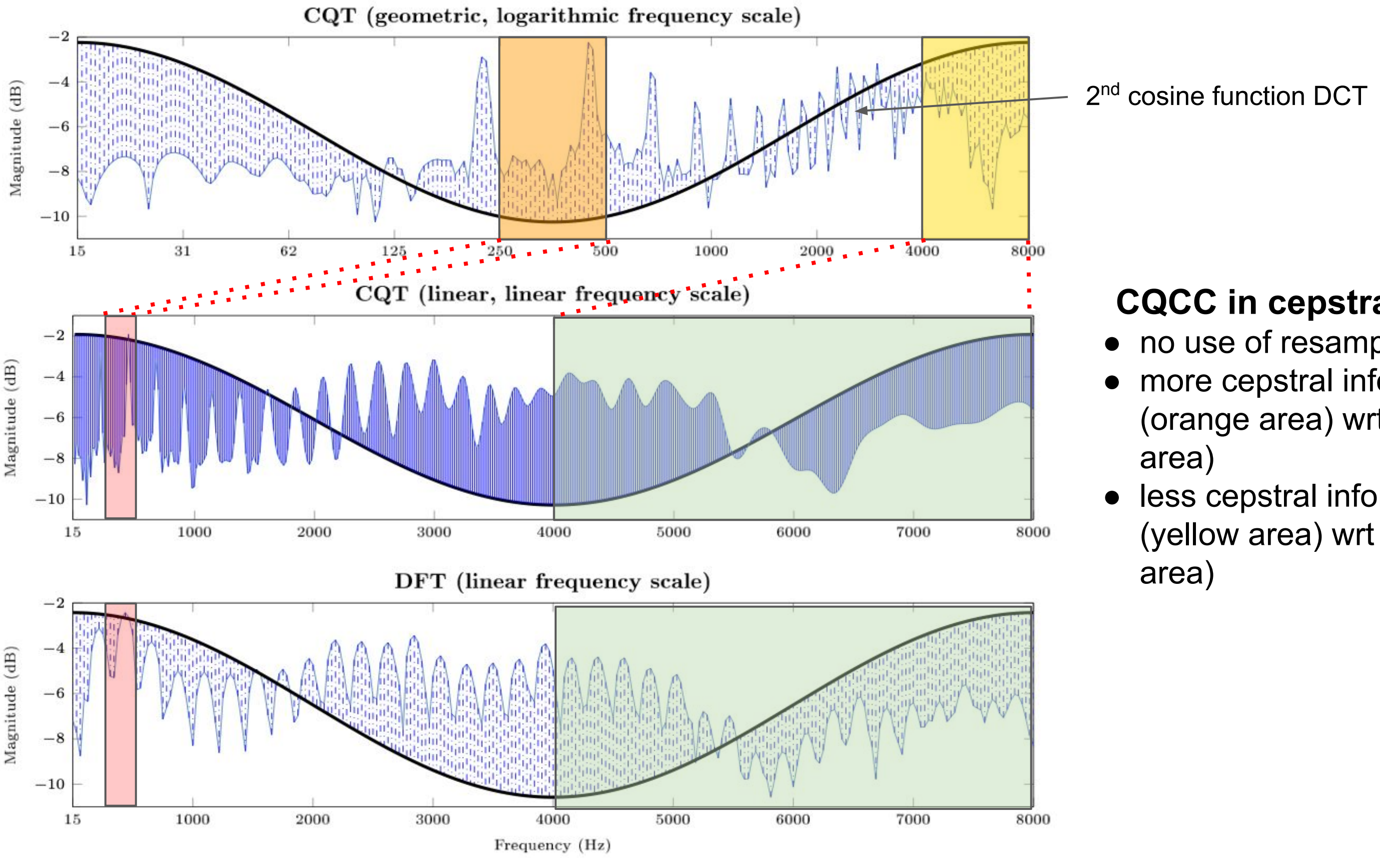
# Sub-band analysis: a method for explainability

**mainly high frequencies → a higher temporal resolution helps**




**mainly mid and low frequencies → a lower frequency resolution helps**

# Time-frequency resolution vs cepstral computation



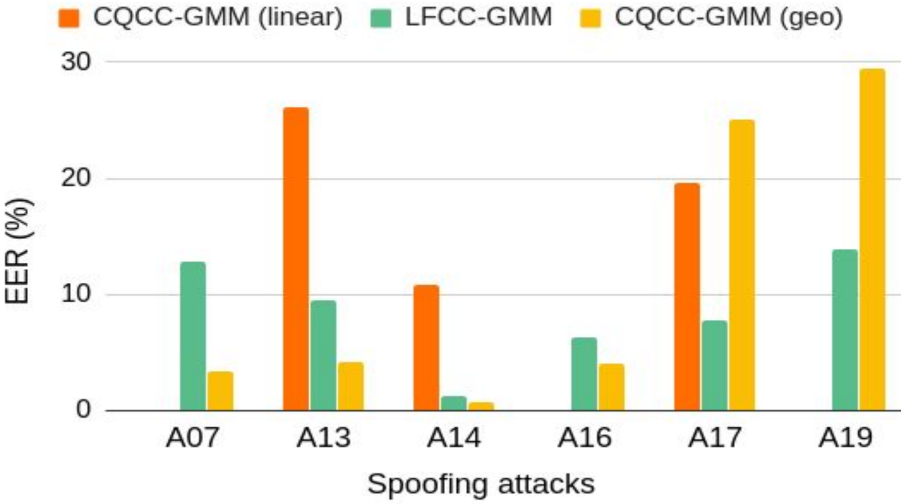

**CQCC in cepstral geometric representation**

- no use of resampling
- more cepstral information at low frequencies (orange area) wrt CQCC linear and DFT (pink area)
- less cepstral information at high frequencies (yellow area) wrt CQCC linear and DFT (green area)

# CQCCs vs LFCCs

- **ASVspoof 2019 LA database**
  - CQCC-GMM (geometric-scale) best detected: A13 and A14

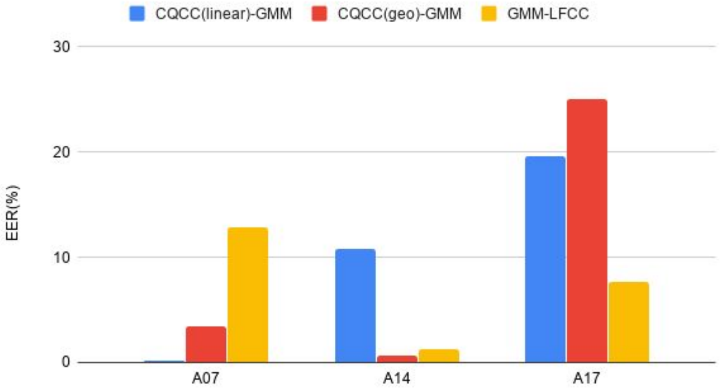
| System | **A07** | A08 | A09 | A10 | A11 | A12 | **A13** | **A14** | A15 | **A16** | **A17** | A18 | **A19** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GMM-CQCC (linear) | **0.00** | 0.04 | 0.14 | 15.16 | 0.08 | 4.74 | **26.15** | **10.85** | 1.26 | **0.00** | **19.62** | 3.81 | **0.04** |
| GMM-LFCC | **12.86** | 0.37 | 0.00 | 18.97 | 0.12 | 4.92 | **9.57** | **1.22** | 2.22 | **6.31** | **7.71** | 3.58 | **13.94** |
| GMM-CQCC (geometric) | **3.39** | 0.34 | 0.46 | 6.86 | 4.62 | 3.58 | **4.23** | **0.67** | 1.52 | **4.00** | **25.04** | 19.63 | **29.46** |

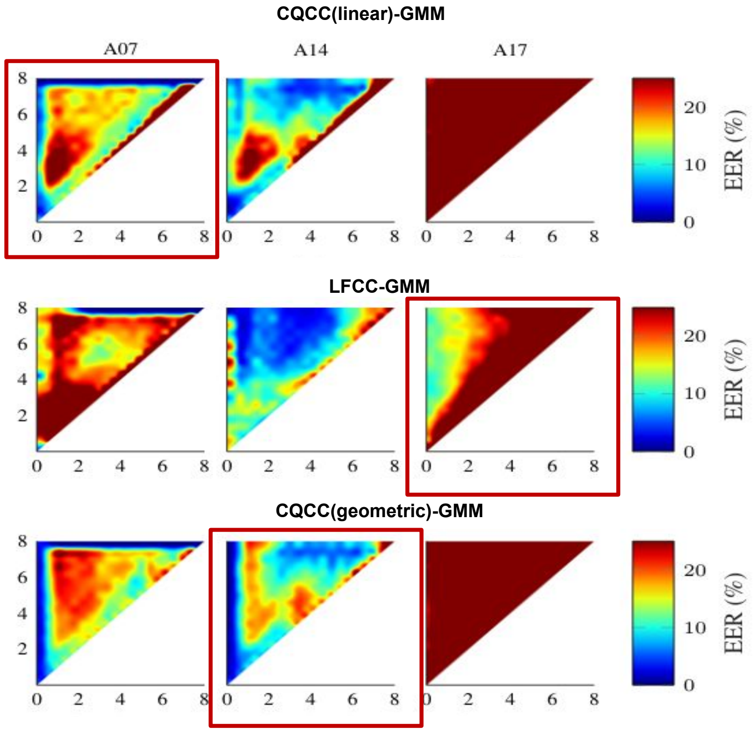


# Findings and key messages

- no silver bullet works
- different attacks exhibit artefacts within different subbands
- better potential to capture these with front-ends which emphasise information in the relevant frequency band

| System | A07 | A14 | A17 |
|---|---|---|---|
| GMM-CQCC (linear) | 0.00 | 10.85 | 19.62 |
| GMM-LFCC | 12.86 | 1.22 | 7.71 |
| GMM-CQCC (geometric) | 3.39 | 0.67 | 25.04 |

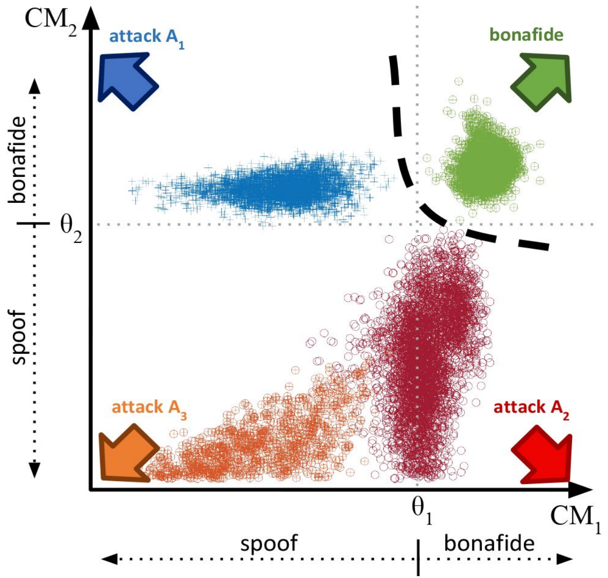


# Fusion of subband-based classifiers?

- **Time-frequency resolution matters a lot!**
  - hypothesis 1
    - what if we increase LFCC resolution?

- **Ensemble of subband-based classifiers work?**
  - hypothesis 2
    - an ensemble of subband classifiers, each tuned to the detection of different attacks in different sub-bands, should give more reliable detection
  - hypothesis 3
    - non-linear, rather than linear fusion of subband classifiers will better exploit complementarity




[.] H. Tak et al., “Spoofing Attack Detection using the Non-linear Fusion of Sub-band Classifiers,” in Proc. INTERSPEECH, 2020.

# High resolution LFCC

- **Optimisation of the spectral resolution at full-band level**
  - nothing simpler: 30 ms window with a 15 ms shift using 1024 FFT points

- **Optimisation of the number of filterbanks**
  - Bhattacharyya distance to optimise the number of filters in a linear filterbank

Table 1: *min t-DCF, EER and Bhattacharyya distance between bona fide and spoofed score distributions for different numbers of subband filters $N$. Baseline configuration illustrated in bold; selected configuration in italics.*

| Filters ($N$) | min t-DCF | EER (%) | $D_B$ |
|---|---|---|---|
| **20** | **0.2110** | **2.71** | **0.1338** |
| 30 | 0.0000 | 0.79 | 0.1706 |
| 40 | 0.0000 | 0.00 | 0.1770 |
| 50 | 0.0000 | 0.00 | 0.1785 |
| 60 | 0.0000 | 0.00 | 0.1793 |
| *70* | *0.0000* | *0.00* | *0.1826* |
| 80 | 0.0000 | 0.00 | 0.1788 |
| 90 | 0.0000 | 0.00 | 0.1823 |
| 100 | 0.0000 | 0.00 | 0.1830 |
| 120 | 0.0000 | 0.00 | 0.1820 |

$$D_B(b,s) = \frac{1}{4} ln\left(\frac{1}{4}\left(\frac{\sigma_b^2}{\sigma_s^2}+\frac{\sigma_s^2}{\sigma_b^2}+2\right)\right)+\frac{1}{4}\left(\frac{(\mu_b-\mu_s)^2}{\sigma_b^2+\sigma_s^2}\right)$$

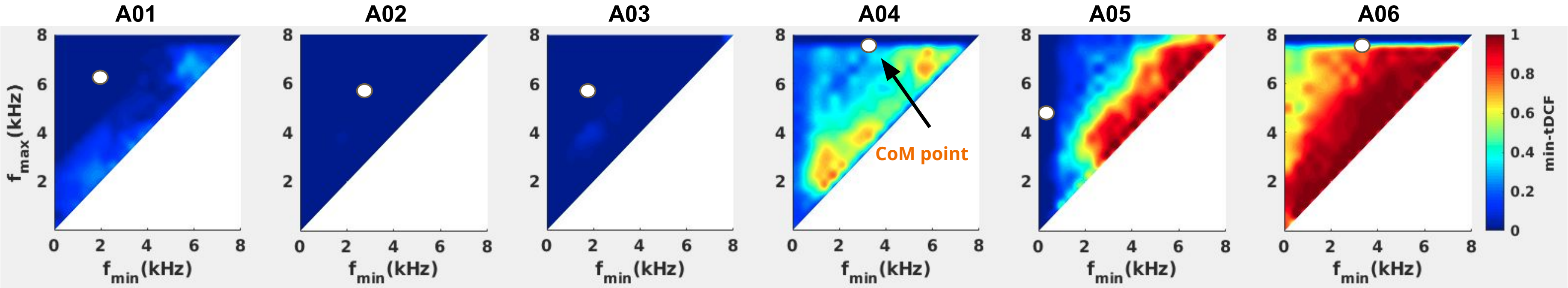
where subscripts $b$ and $s$ indicate parameters for bona fide and spoofed score distributions and where $\mu$ and $\sigma$ refer to the means and standard deviations respectively.

| System | EER(%) |
|---|---|
| CQCC-GMM (baseline) | 9.57 |
| LFCC-GMM (baseline) | 8.09 |
| **HR-LFCC-GMM** | **3.50** |

# Subband classifiers: how to choose them

- **Specific subband selection using Centre-of-Mass (CoM) approach for each spoof attack**
  - CoM is a rudimentary means of dealing with a noisy surface containing multiple minima
  - …but it works!
  - the coordinates $\bar{R} = [f_{\min}^{CoM}, f_{\max}^{CoM}]$ of the CoM satisfy the condition $\sum_{i=1}^{n} m_i(r_i - \bar{R}) = 0$

$$R = \frac{1}{M}\sum_{i=1}^{n} m_i r_i$$

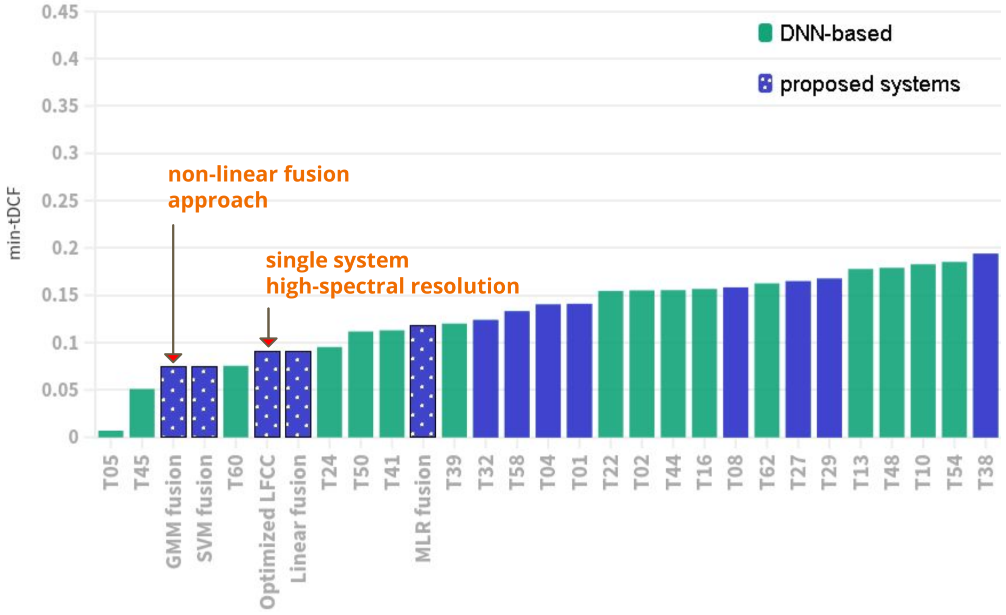



2D heat map representation of subband analysis results for the six different spoof attacks on development partition

# Results on ASVspoof 2019 LA

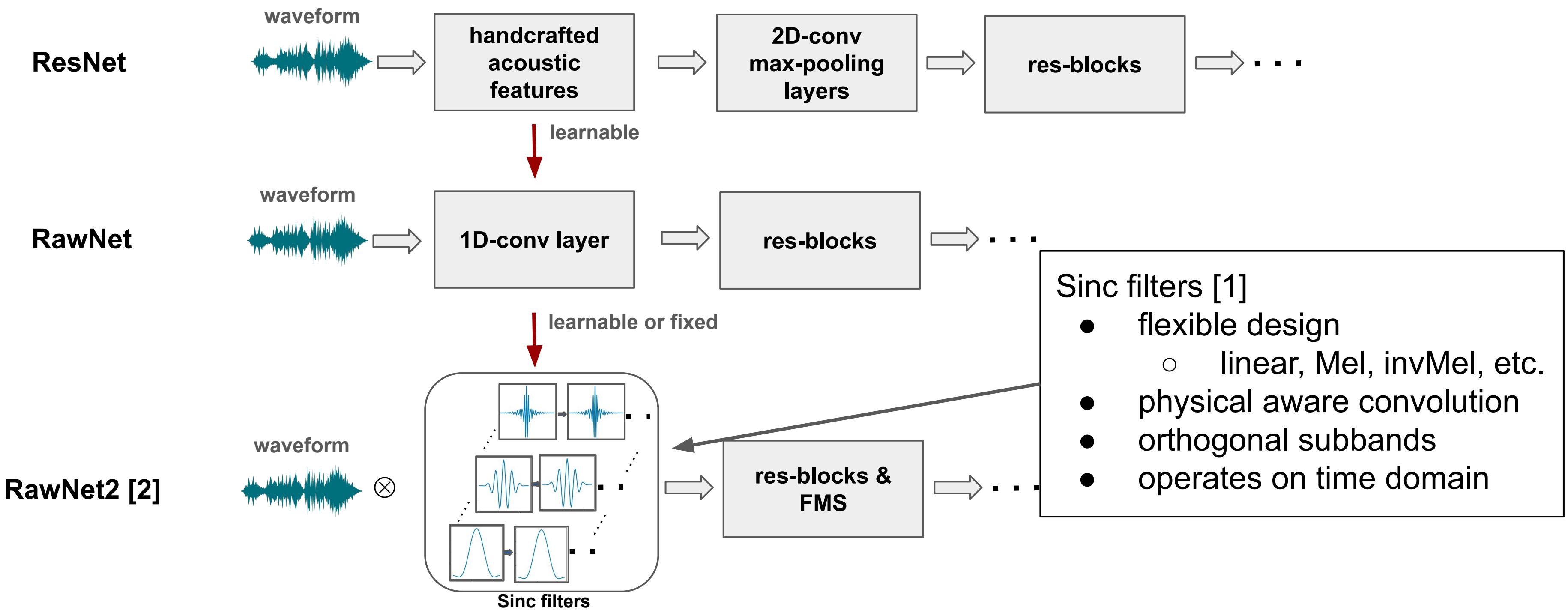
| System | EER (%) | min-tDCF |
|---|---|---|
| **HR-LFCC-GMM** | 3.50 | 0.0904 |
| **Ensemble GMM-fusion** | **2.92** | **0.0740** |
| **Ensemble SVM-fusion** | 2.92 | 0.0748 |
| **Ensemble Linear fusion** | 3.38 | 0.0911 |

Performance comparisons with 48 challenge competing systems




[.] H. Tak et al., "Spoofing Attack Detection using the Non-linear Fusion of Sub-band Classifiers," in Proc. INTERSPEECH, 2020.

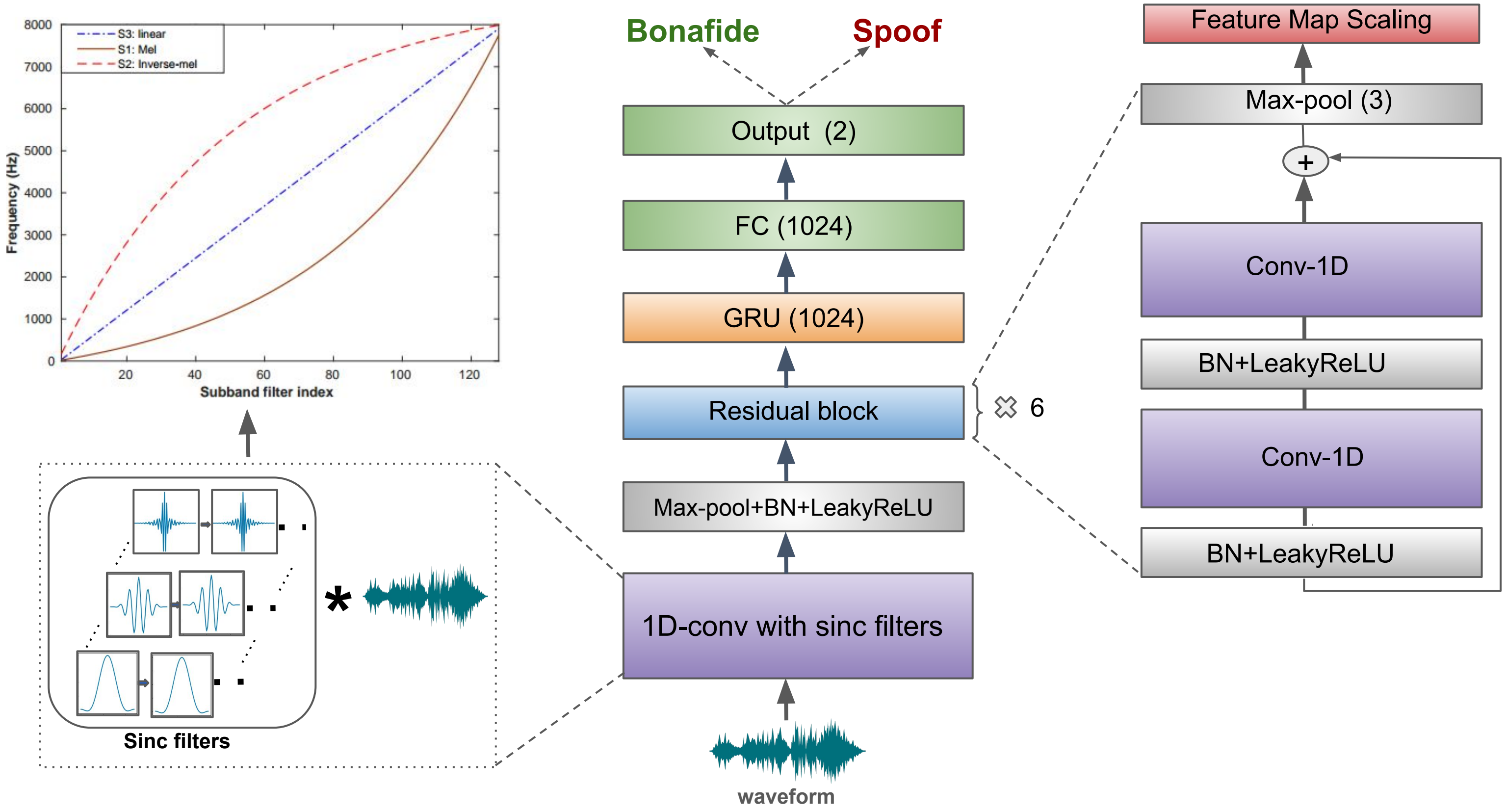
# From ResNet to RawNet2 passing by Sinc filters





[1] M. Ravanelli, Y. Bengio, “Speaker recognition from raw waveform with sincnet,” in IEEE Proc. Spoken Language Technology Workshop (SLT), 2018.
[2] H. Tak, et al.,“End-to-end anti-spoofing with RawNet2,” in Proc. IEEE ICASSP, 2021

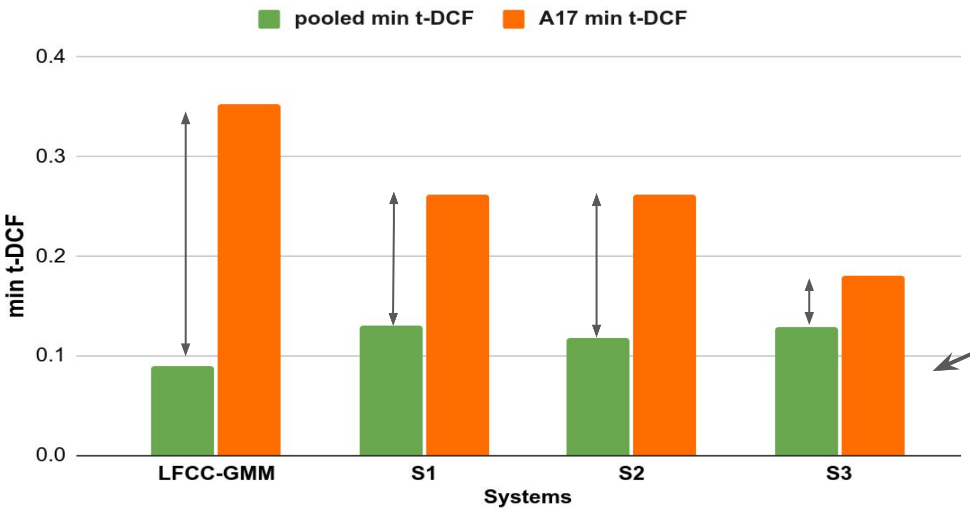
# RawNet2 - end2end approach in the time domain




[1] Jung et al., “Improved RawNet with Feature Map Scaling for Text-independent Speaker Verification using Raw Waveforms, Interspeech 2020.

# RawNet2 - end2end approach in the time domain

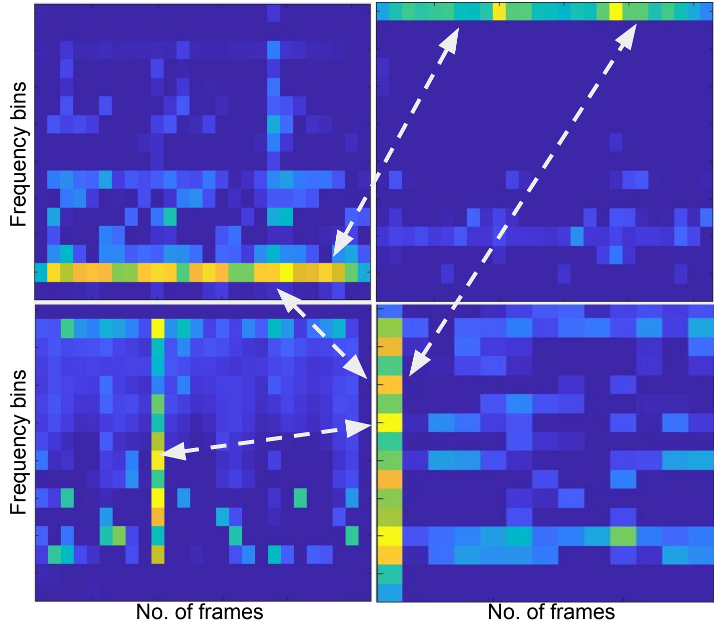
- **Comparison of overall pooled performance and worst-case A17 spoof attack**
  - time-domain processing facilitates the detection of the most challenging A17 attack
  - achieves close to state-of-the-art performance while operating upon raw audio signals in truly end-to-end fashion




- **HR-LFCC-GMM**
- **S1**: fixed **Mel-scaled** sinc filters
- **S2**: fixed **Inverse mel-scaled** sinc filters
- **S3**: fixed **linear-scaled** sinc filters

# Graph attention (GAT) to model T-F atoms

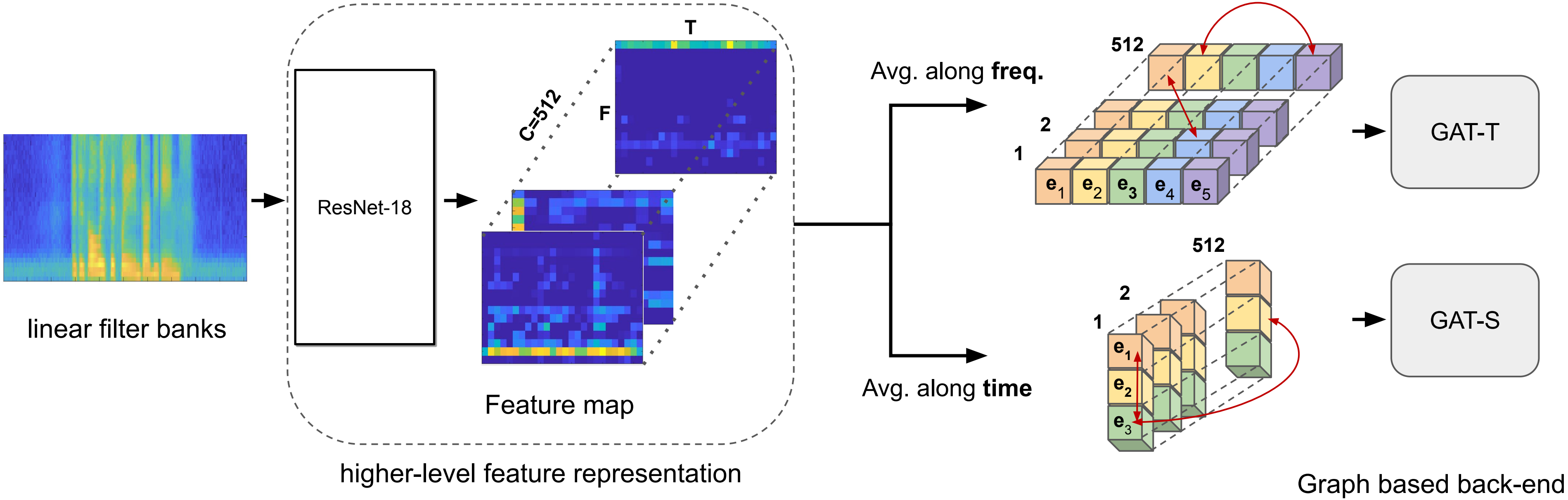
- what we know → artefacts lie in specific subbands or temporal frames [1,2,3]
- conventional attention mechanisms do not explicitly model these relationships
- modelling the relationship between the evidence spanning different sub-bands and time intervals
- to leverage the potential of GAT for modeling relationships in spectral or temporal domain [4,5]




feature-map representations from deep residual network


[1] H. Tak et al., “An explainability study of the constant Q cepstral coefficient spoofing countermeasure for automatic speaker verification,” in Proc. Speaker Odyssey Workshop, 2020.
[2] B. Chettri et al., “Subband Modeling for Spoofing Detection in Automatic Speaker Verification,” in Proc. Speaker Odyssey Workshop, 2020.
[3] H. Tak et al., “Spoofing Attack Detection using the Non-linear Fusion of Sub-band Classifiers,” in Proc. INTERSPEECH, 2020.
[4] P. Velickovic et al., “Graph attention networks,” in Proc. ICLR, 2018.
[5] J.-w. Jung et al., “Graph attention networks for speaker verification,” in Proc. ICASSP, 2021

# Graph attention (GAT) to model T-F atoms




[6] H. Tak, Jee-weon Jung, J. Patino, M. Todisco and N. Evans, "Graph attention network for anti-spoofing," in Proc. INTERSPEECH 2021

# Graph attention networks (GAT)

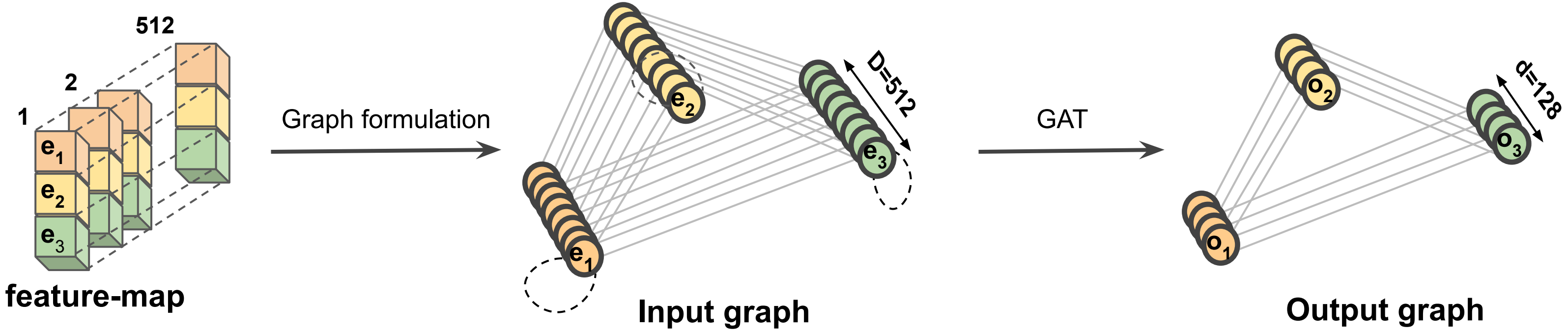
**GAT modelling**

**Input**: a set of node features

$e = \{ e_1, e_2, \ldots, e_N \}, \; e_n \in \mathbb{R}^D$

**Output**: a set of new node features including neighboring information

$o = \{ o_1, o_2, \ldots, o_N \}, \; o_n \in \mathbb{R}^d$





[4] P. Velickovic et al., “Graph attention networks,” in Proc. ICLR, 2018.

# Graph attention (GAT) results

- **Performance comparisons with other CM techniques on ASVspoof 2019 LA**

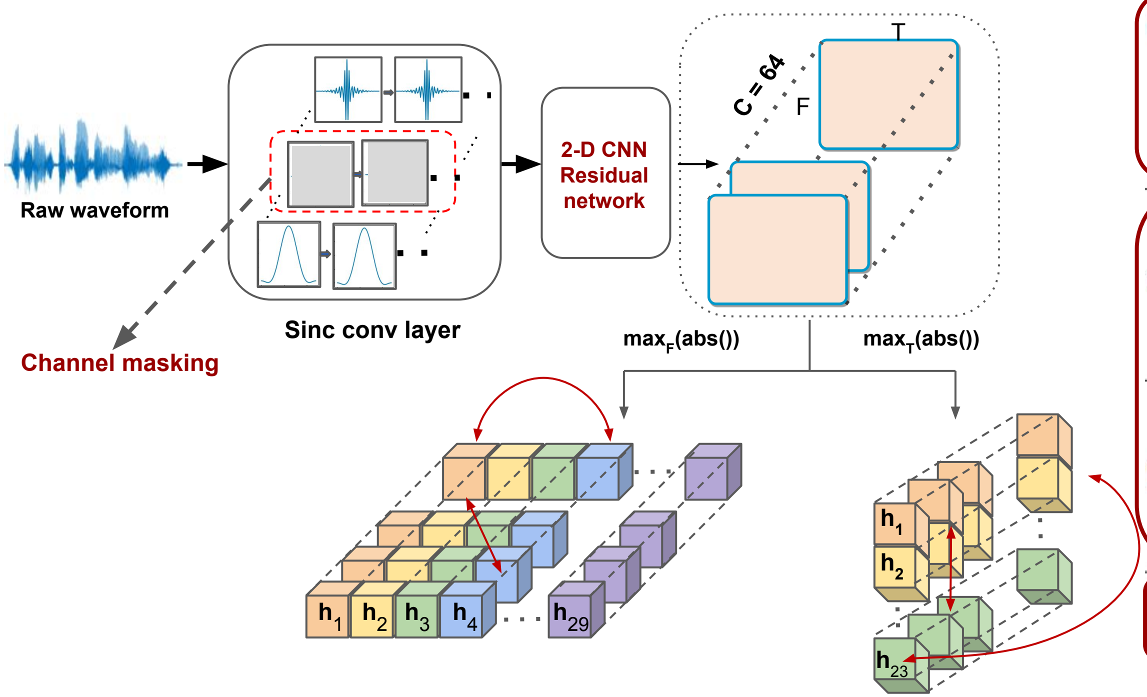
| CM Systems | Pooled min t-DCF | Pooled EER (%) |
|---|---|---|
| **HR-LFCC-GMM** | 0.090 | **3.50** |
| **RawNet2 + RawBoost** | 0.155 | 5.31 |
| **GAT-T** | **0.089** | 4.71 |
| **GAT-S** | 0.091 | 4.48 |
| **Resnet18-SP** | 0.114 | 6.82 |
| **Resnet18-SAP** | 0.138 | 7.11 |
| **ResNet18-ASP** | 0.127 | 6.22 |

- Limitations
  - spectral and temporal relationship is separated → no communication
  - averaging nodes in aggregator might not be informative

**a single E2E GAT model might be useful**

# E2E feature learning from raw waveform

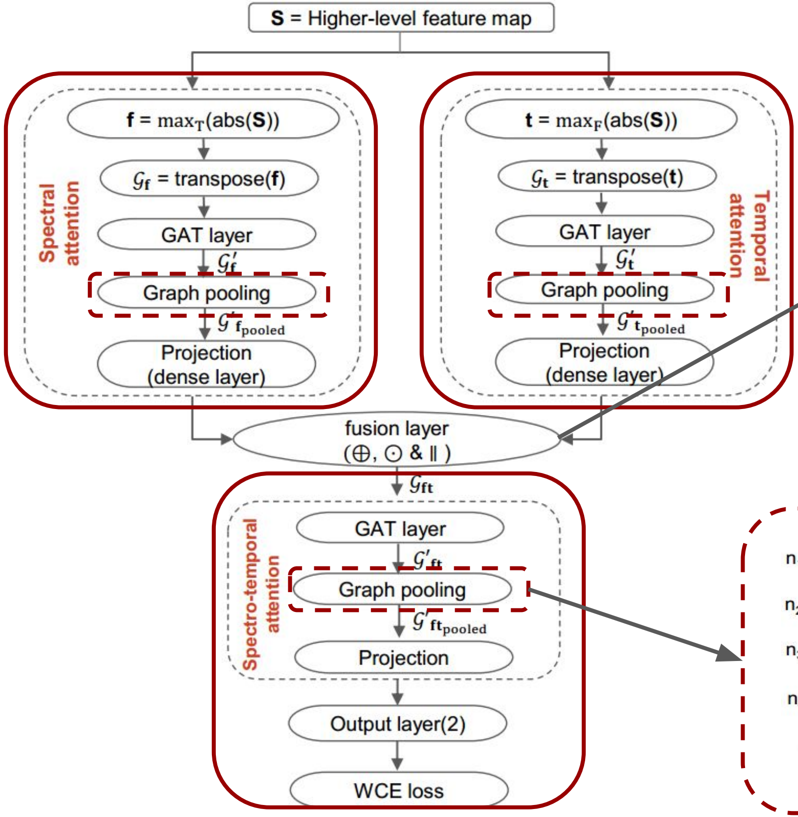
- modeling the relationship between subbands and temporal segments in E2E fashion
- employing graph pooling to improve performance



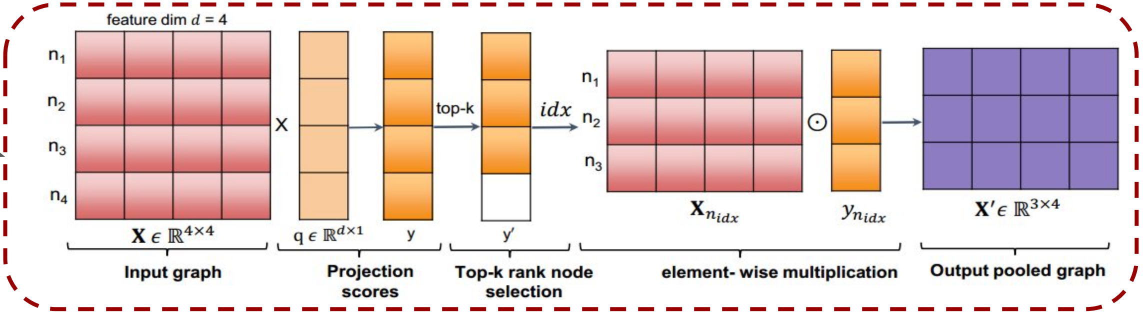

| Layer | Input: 64600 samples | | Output shape |
|---|---|---|---|
| Sinc layer | Conv-1D(129,1,70) | | (70,64472) |
| | add channel (TF representation) | | (1,70,64472) |
| | Maxpool-2D(3) | | (1,23,21490) |
| | BN & SeLU | | |
| Res block | Conv-2D((2,3),1,32)<br>BN & SeLU<br>Conv-2D((2,3),1,32)<br>Maxpool-2D((1,3)) | × 2 | (32,23,2387) |
| Res block | Conv-2D((2,3),1,64)<br>BN & SeLU<br>Conv-2D((2,3),1,64)<br>Maxpool-2D((1,3)) | × 4 | (64,23,29) |

| Spectral-attention | Temporal-attention |
|---|---|
| $\max_T(\text{abs}()) = (64, 23)$ | $\max_F(\text{abs}()) = (64, 29)$ |

# Spectro-temporal GAT (RawGAT-ST)




1. Element-wise multiplication (GAT-S ⊙ GAT-T )
2. Element-wise addition (GAT-S ⊕ GAT-T )
3. Concatenation (GAT-S || GAT-T )




[1] H. Gao, S. Ji, “Graph u-nets,” in international conference on machine learning (PMLR), 2019.

# Spectro-temporal GAT (RawGAT-ST)

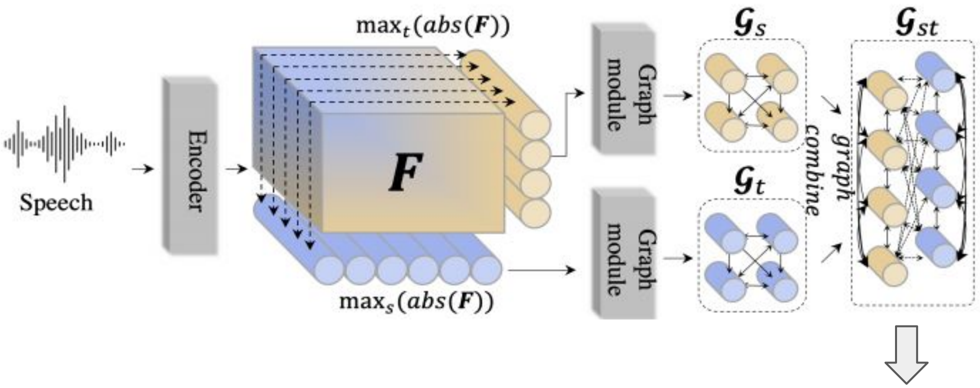
- **Performance comparisons with other CM techniques on ASVspoof 2019 LA**
  - RawGAT-ST-mul [1] shows 78% relative reduction in min t-DCF over RawNet2 system

| CM systems | min t-DCF | EER(%) |
|---|---|---|
| HR-LFCC-GMM | 0.090 | 3.50 |
| RawNet2 + RawBoost | 0.155 | 5.54 |
| **RawGAT-ST-mul** | **0.033** | **1.06** |
| RawGAT-ST-add | 0.037 | 1.15 |
| RawGAT-ST-concat | 0.038 | 1.23 |

[1] H. Tak et al., "End-to-End Spectro-Temporal Graph Attention Networks for Speaker Verification Anti-Spoofing and Speech Deepfake Detection," in ASVspoof 2021 workshop.

# AASIST model for anti-spoofing

- explore heterogeneous graph attention network to model the heterogeneous relationship between spectral and temporal domains.
- relationship between different types of nodes (spectral & temporal) and edges.
- to learn the importance between a node and its meta-path based neighboring nodes.




S - S (pre-define edge values, solid arrow)
S - T (learnable edge attention weights, dashed arrow)
T - S (learnable edge attention weights,dashed arrow)
T - T (pre-define edge values,solid arrow)

**Performance on ASVspoof 2019 LA**

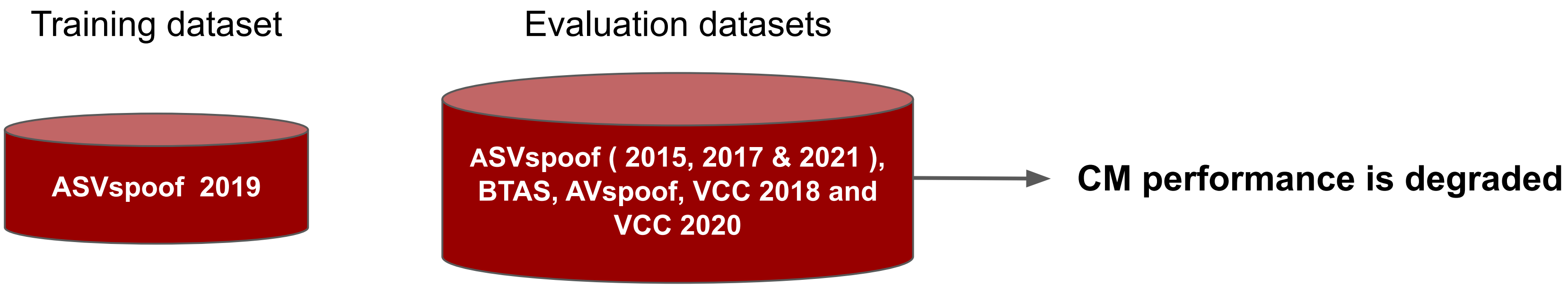
→ **AASIST model [2]** shows **25% relative reduction in min t-DCF** over RawGAT-ST-mul

→ **EER: from 1.06% to 0.83%**

[1] X. Wang et al., “Heterogeneous Graph Attention Network,” in The World Wide Web Conference, 2019.
[2] J. Jung et al., “AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks,” submitted in ICASSP 2022.

# Database mismatch: challenges and solutions

- Challenges
  - Lack of generalisation and domain mismatch between training and testing data [1,2].
  - Lack of sufficiently representative training data.





[1] D. Paul, M. Sahidullah, et al.," Generalization of spoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora", in Proc. ICASSP 2017.
[2] R. K Das, H. Li, "Assessing the scope of generalized countermeasures for anti-spoofing", in Proc. ICASSP 2020.

IBERSPEECH22 GRANADA 14-16 NOV

- **RawNet2 and AASIST trained on 2019 and tested on 2021**

| | No codec C1 | a-law C2 | unk. + $\mu$-law C3 | G.722 C4 | $\mu$-law C5 | GSM C6 | OPUS C7 | Pooled |
|---|---|---|---|---|---|---|---|---|
| **A07** | 0.88 | 3.65 | 14.54 | 1.56 | 3.49 | 7.08 | 3.51 | 5.57 |
| **A08** | 3.24 | 6.01 | 15.54 | 4.45 | 5.76 | 10.17 | 6.62 | 8.87 |
| **A09** | 0.76 | 3.34 | 17.28 | 1.29 | 3.15 | 6.55 | 3.45 | 5.29 |
| **A10** | 0.97 | 3.48 | 12.2 | 1.61 | 3.53 | 6.83 | 3.35 | 5.33 |
| **A11** | 0.97 | 3.85 | 13.06 | 1.74 | 3.91 | 8.52 | 3.48 | 5.65 |
| **A12** | 0.97 | 3.86 | 12.12 | 1.77 | 3.75 | 7.52 | 3.7 | 5.95 |
| **A13** | 0.73 | 2.71 | 10.61 | 1.13 | 2.68 | 3.78 | 2.3 | 3.77 |
| **A14** | 1.04 | 4.01 | 14.79 | 1.88 | 3.92 | 9.46 | 3.62 | 6.19 |
| **A15** | 1.03 | 3.85 | 12.91 | 1.79 | 3.65 | 8.05 | 3.53 | 5.91 |
| **A16** | 1.18 | 3.85 | 11.88 | 1.85 | 3.69 | 6.79 | 3.87 | 5.6 |
| **A17** | 12.01 | 11.61 | 28.78 | 12.51 | 11.21 | 20.77 | 18.05 | 19.36 |
| **A18** | 20.84 | 21 | 36.93 | 22.08 | 20.37 | 31.12 | 20.75 | 27.32 |
| **A19** | 2.58 | 5.24 | 14.2 | 3.36 | 4.5 | 9.61 | 5.36 | 7.93 |
| **Pooled** | 5.84 | 6.56 | $\boxed{16.72}$ | 6.41 | 6.33 | $\boxed{10.65}$ | 7.98 | **9.49** |

| CM systems | EER(%) |
|---|---|
| RawNet2 | 9.49 |
| AASIST | 11.47 |

# Augmentation and self-supervised models

- use of larger and diverse representative training database
  - Advantage: better generalisation
  - Disadvantage: It’s impractical - never enough

to use **self-supervised models** trained on large diverse datasets to learn more generalised representation

**Data augmentation**

**combining both techniques to improve generalizability of CMs**

[1] D. Paul, M. Sahidullah, et al.,” Generalization of spoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora”, in Proc. ICASSP 2017.
[2] R. K Das, H. Li, “Assessing the scope of generalized countermeasures for anti-spoofing”, in Proc. ICASSP 2020.

# Data augmentation

- **Why data augmentation (DA) is important for machine learning?**
  - increasing training data by introducing more variability
  - reduce model overfitting
  - improves generalization and robustness to out-of-domain data

- **DA methods**
  - SpecAugment
  - WavAugment
  - Codec
  - Multimedia & codec transformations
  - RIR convolutive noise
  - MUSAN database additive noise

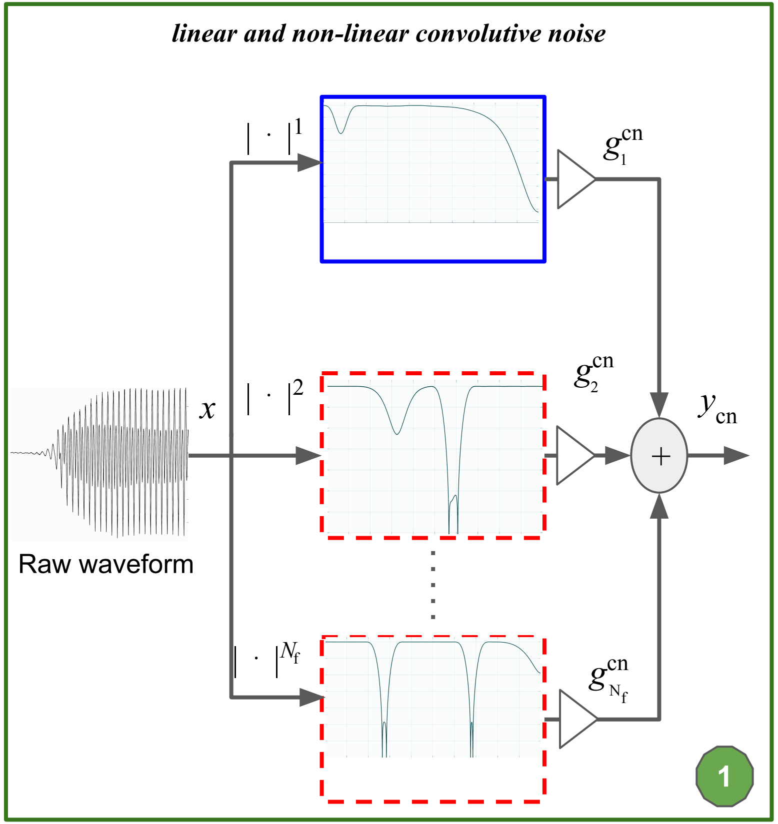
[.] D. S. Park, W. Chan et al., “SpecAugment: A simple data augmentation method for automatic speech recognition,” in Proc. INTERSPEECH, 2019.
[.] E. Kharitonov, M. Riviere et al., “Data augmenting contrastive ` learning of speech representations in the time domain,” in Proc. IEEE SLT, 2021.

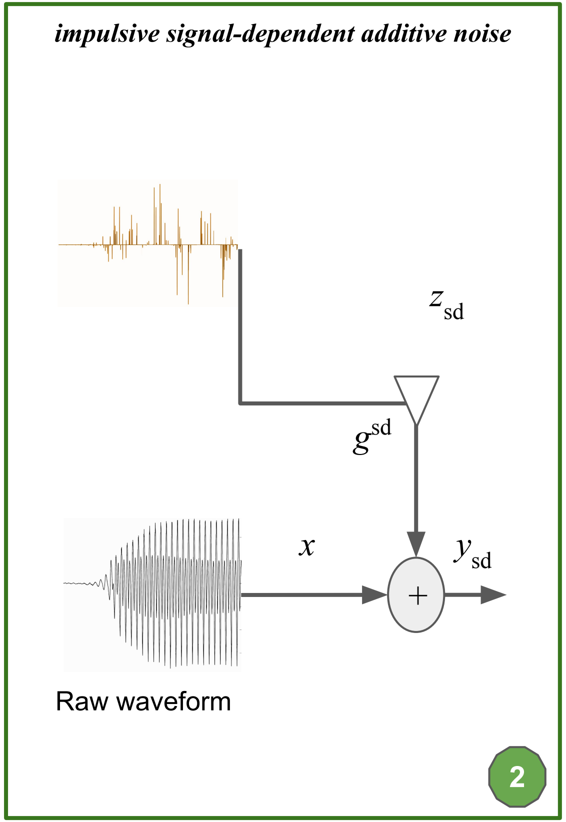
# RawBoost data augmentation [1]

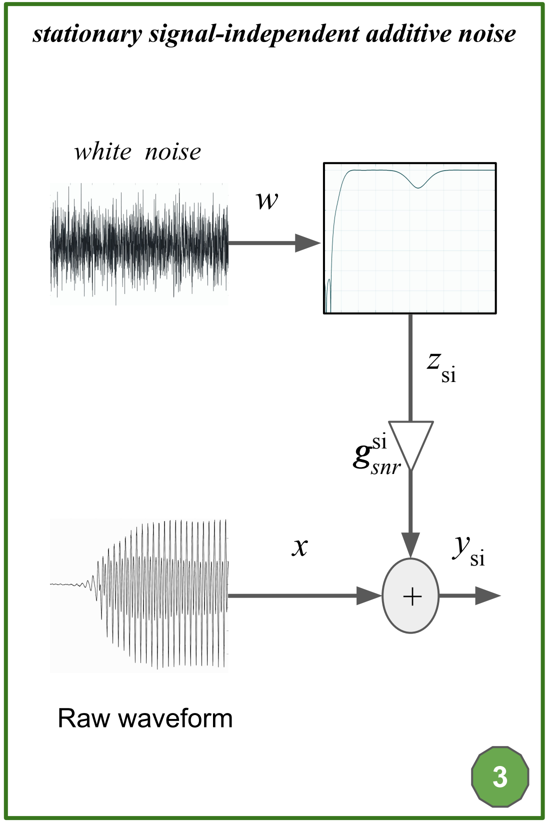
- **RawBoost**
  - a raw data boosting and augmentation method
  - no additional data sources
  - operate directly upon raw waveform inputs
  - to address
    - lack of generalisation
    - channel and transmission nuisance
    - compression
  - with 3 algorithms
    - 1. linear and non-linear convolutive noise
    - 2. impulsive signal-dependent additive noise
    - 3. stationary signal-independent additive noise

[1] H. Tak,et al., “RawBoost: A Raw Data Boosting and Augmentation Method applied to Automatic Speaker Verification Anti-Spoofing,” accepted in ICASSP, 2022.

# RawBoost algoritms



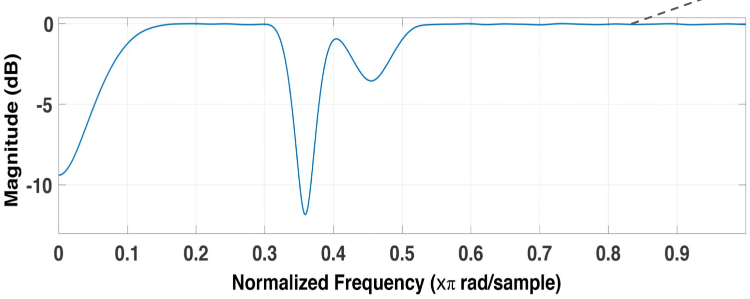



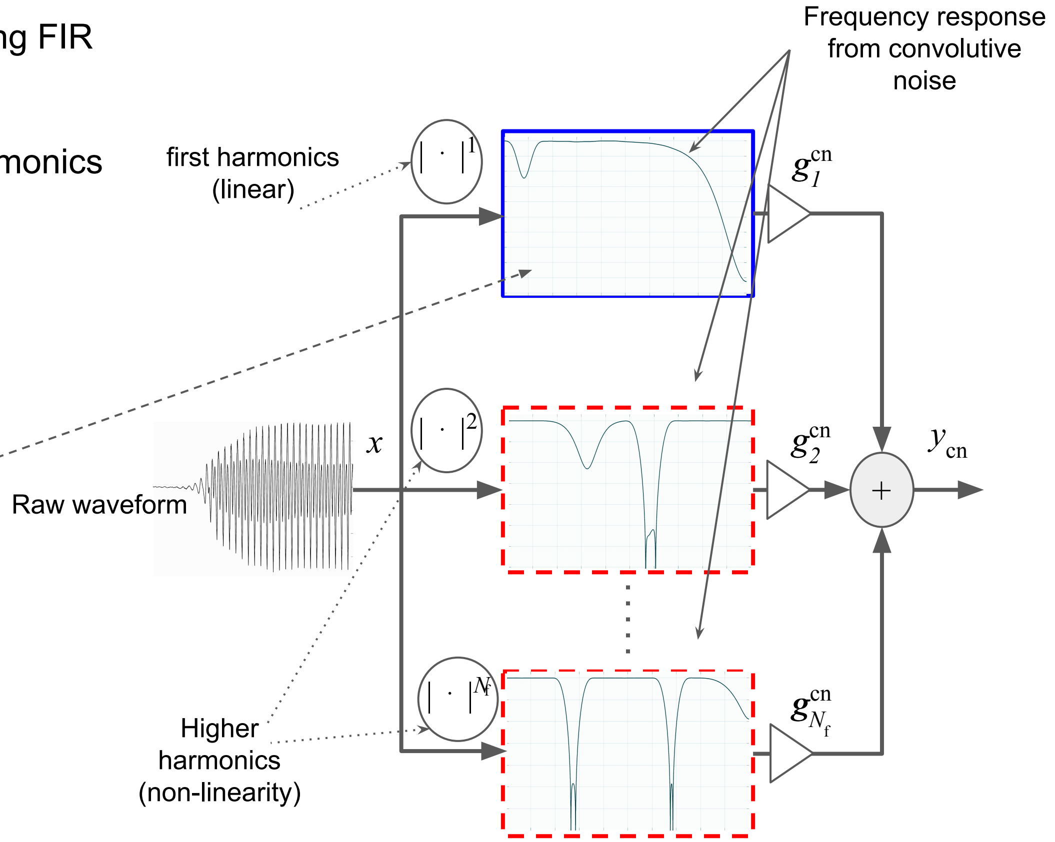
**Convolutive noise**: time domain notch filtering (using FIR filter)

**Hammerstein systems**: generate higher-order harmonics (non-linearity)

$$y_{\mathrm{cn}}[n] = \sum_{j=1}^{N_{\mathrm{f}}} g_j^{\mathrm{cn}} \sum_{i=0}^{N_{\mathrm{fir}_j}} b_{i_j} \cdot x^j[n-i]$$

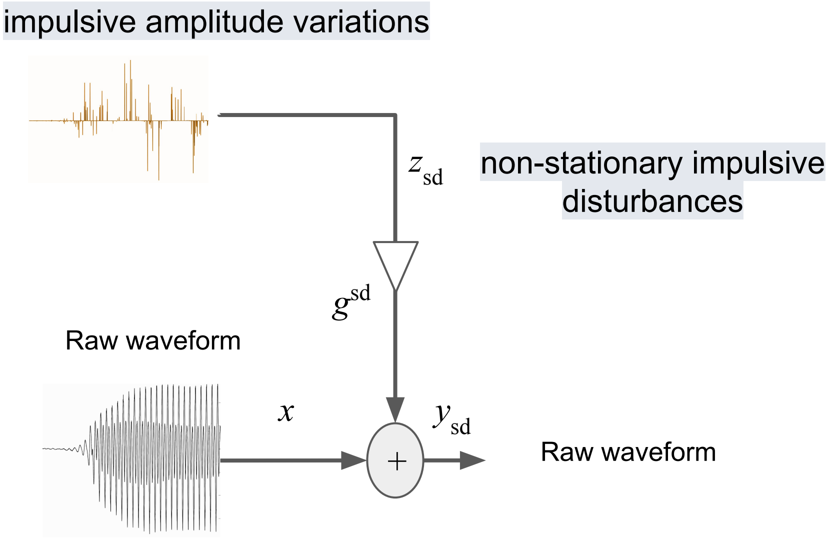
Magnitude response of a multi-band filter



# RawBoost 2 impulsive signal-dependent additive noise

- **Impulsive noise** usually generated through non-linear processes in microphones and amplifiers devices.
- We change the samples (chosen at random) with an amount proportional to the value of the sample itself.




$$y_{\mathrm{sd}}[n] = x[n] + z_{\mathrm{sd}}[n]$$

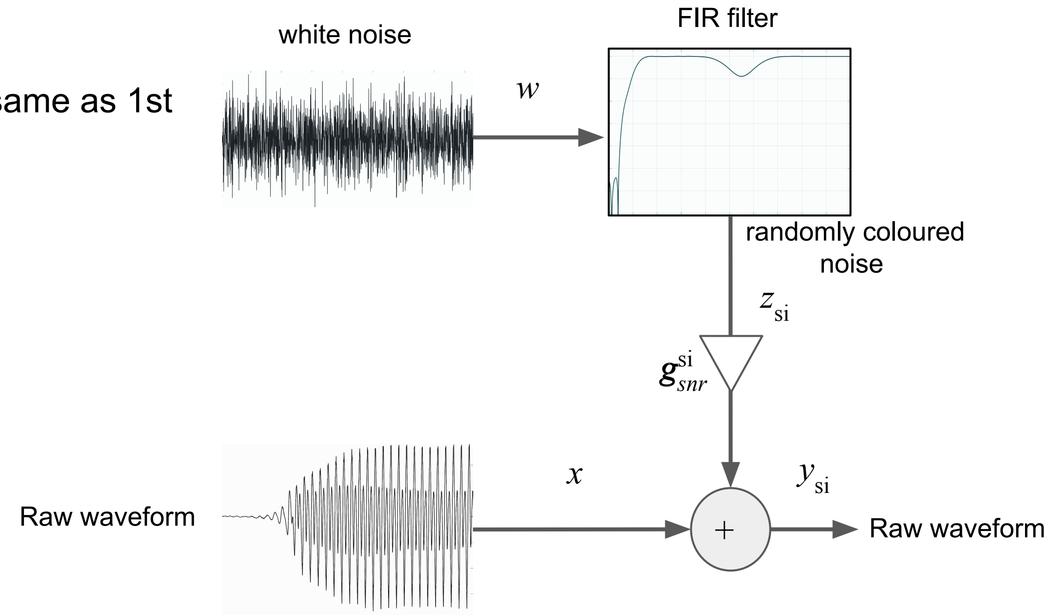
$$z_{\mathrm{sd}}[n] = \begin{cases} g^{\mathrm{sd}} \cdot D_R\{-1,1\}[n] \cdot x[n], & \textit{if } \; n = \{p_1, p_2, ..., p_P\} \\ 0, & \text{otherwise} \end{cases}$$

# RawBoost 3 stationary signal-independent additive noise

- introduce through poorly joined cable connections, transmission channels effects, electromagnetic interference.
- **colored additive noise** using FIR filtering (same as 1st algo.) with a randomly chosen SNR.

$$y_{\mathrm{si}}[n] = x[n] + g^{\mathrm{si}}_{snr} \cdot z_{\mathrm{si}}[n]$$

$$g^{\mathrm{si}}_{snr} = \frac{10^{\frac{SNR}{20}}}{\|z_{\mathrm{si}}\|^2 \cdot \|x\|^2}$$



# RawBoost configuration

- **RawNet2 with RawBoost DA**
- **RawBoost DA applied on-the-fly to existing training and development ASVspoof 2019 LA**
- **Comparisons with standard data augmentation techniques**
  - SpecAugment [1]
  - WavAugment [2]
- **RawBoost configuration**
  - RawBoost parameter values for each of the three different techniques
  - values within expressed ranges are selected at random (uniform distributions)

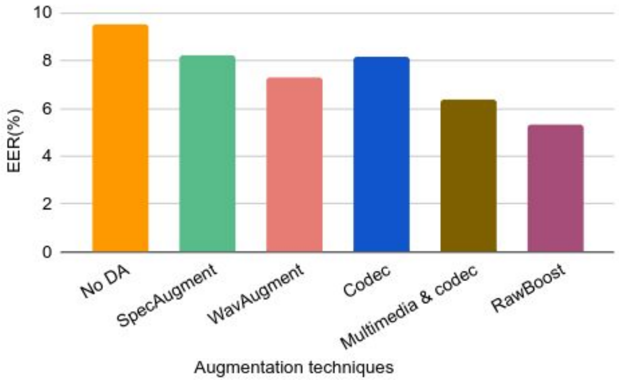
| Parameters | Notch filter | $N_{fir}$ coefficients | Non-linearity ($N_f$) | $f_c$ [Hz] | Δf [Hz] | $g^{cn}_1$ [dB] | $g^{cn}_{2\text{-}Nf}$ [dB] | $P_{relative}$ [%] | $g^{sd}$ | SNR [dB] |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5 | [10,100] | 5 | [20,4000] | [100,1000] | [0,0] | [-5,-20] | - | - | - |
| 2 | - | - | - | - | - | - | - | [0,10] | 2 | - |
| 3 | 5 | [10,100] | 1 | [20,4000] | [100,1000] | - | - | - | - | [10,40] |

[1] D. S. Park, W. Chan et al., “SpecAugment: A simple data augmentation method for automatic speech recognition,” in Proc. INTERSPEECH, 2019.
[2] E. Kharitonov, M. Riviere et al., “Data augmenting contrastive ` learning of speech representations in the time domain,” in Proc. IEEE SLT, 2021.

# RawBoost performances on ASVspoof 2021 LA

- **Performance comparisons with other DA techniques**

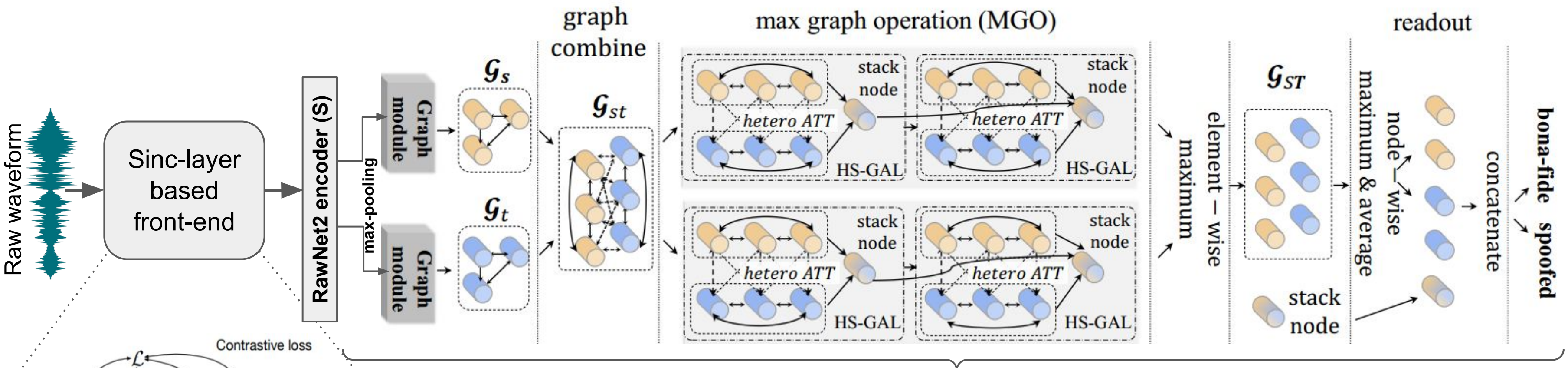
| CM system | Augmentation | min t-DCF | EER(%) |
|---|---|---|---|
| RawNet2 (Baseline) | No augmentation | 0.4257 | 9.49 |
| RawNet2 | SpecAugment | 0.3418 | 8.25 |
| RawNet2 | WavAugment | 0.3435 | 7.32 |
| RawNet2 | Codec | 0.3297 | 8.17 |
| RawNet2 | Multimedia & codec transformations | 0.3168 | 6.36 |
| **RawNet2** | **RawBoost** | **0.3099** | **5.31** |



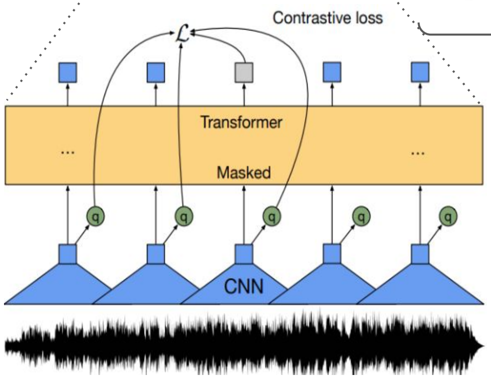

| CM systems | Augm | min t-DCF | EER(%) |
|---|---|---|---|
| AASIST | - | 0.5081 | 11.47 |
| **AASIST** | **RawBoost** | **0.2804** | **3.89** |

**RawBoost** is model agnostic!

# Self-supervised learning & AASIST



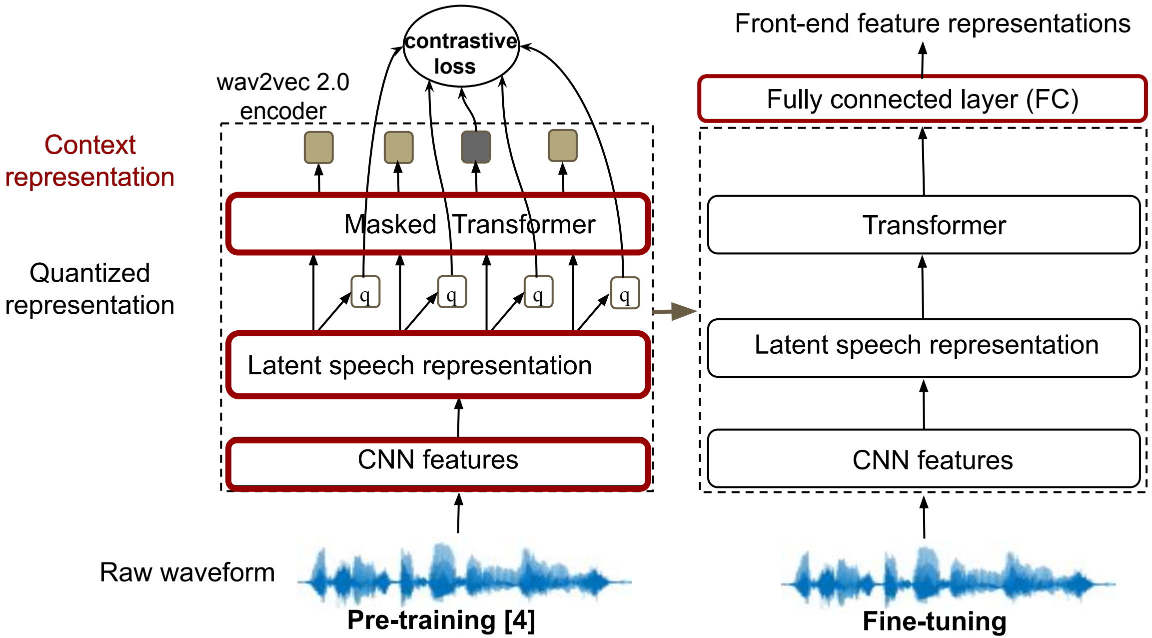

**Graph attention based AASIST framework [1]**




wav2vec 2.0 is a framework of self-supervised learning which had recently achieved prominent results in many speech related tasks


[1] J. Jung, H. Heo, H. Tak  et al., “AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks,” in Proc. ICASSP, 2022.
[2] A. Babu, C. Wang, et al., “XLS-R: Self-supervised cross-lingual speech representation learning at scale,” arXiv preprint arXiv:2111.09296, 2021.c

# Wav2vec 2.0 (XLSR) model




Fine-tuning:

- add a simple linear layer on top of the transformer layer and jointly optimize using weighted cross entropy loss with a lower learning rate
- using ASVspoof 2019 training labeled data.

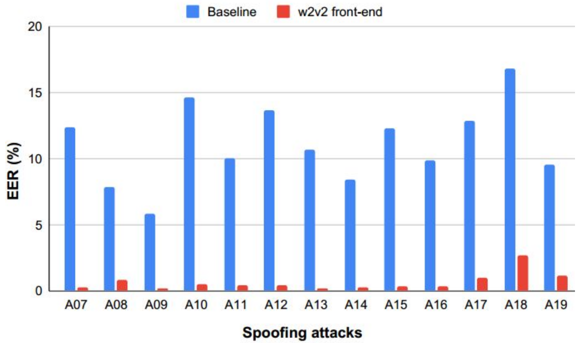
[.] H. Tak, M. Todisco, X. Wang, et al., “Automatic speaker verification spoofing and deepfake detection using Wav2vec 2.0 and data augmentation,” in Proc. Odyssey, 2022, pp. 112–119.

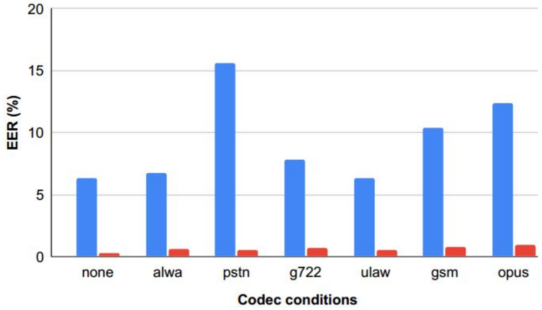
### ASVspoof 2021 LA evaluation set

SA refers to the self-attentive aggregation layer whereas
DA refers to data augmentation

| front-end | SA | DA | Pooled EER | Pooled min t-DCF |
|---|---|---|---|---|
| sinc-layer | × | × | 11.47 (11.95) | 0.5081 (0.5139) |
| wav2vec 2.0 | × | × | 6.15 (6.46) | 0.3577 (0.3587) |
| sinc-layer | ✓ | × | 8.73 (11.61) | 0.4285 (0.5203) |
| wav2vec 2.0 | ✓ | × | 4.48 (6.15) | 0.3094 (0.3482) |
| sinc-layer | ✓ | ✓ | 7.65 (7.87) | 0.3894 (0.3960) |
| **wav2vec 2.0** | ✓ | ✓ | **0.82 (1.00)** | **0.2066 (0.2120)** |

~90% relative improvement




- Baseline: an integrated spectro-temporal graph attention network (AASIST).
- RawBoost Data augmentation applied on-the-fly to existing training database.
- Best single system results on ASVspoof 2021 challenge LA task till date

[1] H. Tak, M. Todisco, X. Wang, et al., “Automatic speaker verification spoofing and deepfake detection using Wav2vec 2.0 and data augmentation,” in Proc. Odyssey, 2022, pp. 112–119.

# SHapley Additive exPlanations (SHAP)

- **An explainability study using SHAP to gain new insights in spoofing detection**
  - use SHAP to estimate the importance of individual speech features for spoofing detection
  - visualise SHAP values for both bona fide and spoofed classes
  - analyse differences in classifier behaviour

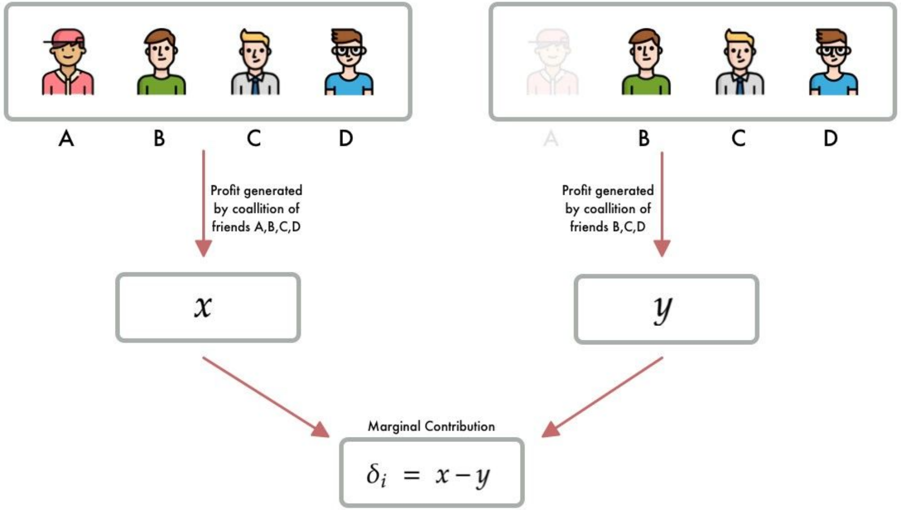
- **Definition**
  - SHAP value $\varphi_i$ can be both negative and positive to reflect the relative (un)importance of a particular feature to a classifier output
  - to obtained $\varphi_i$ , a classifier is trained twice, with and without the inclusion of the feature $i$

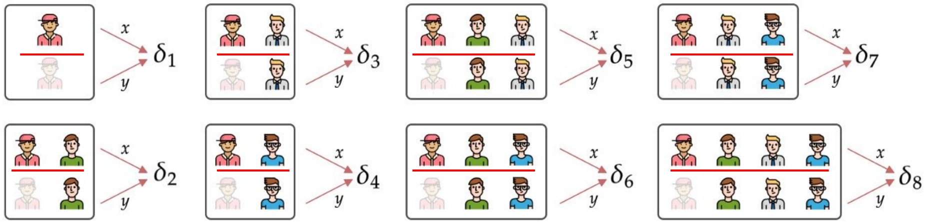
$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! \, (|F| - |S| - 1)!}{|F|!} \delta_i(S)$$

  - where $S$ is a feature subset of full set of features $F$, and $\delta_i$ is the prediction difference of feature $i$ being presented and absent
  - SHAP values are of the same size as the input feature

[1] S. M. Lundberg, S.-i. Lee and D. Fohr, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.

# SHapley Additive exPlanations (SHAP)




The Shapley value for member

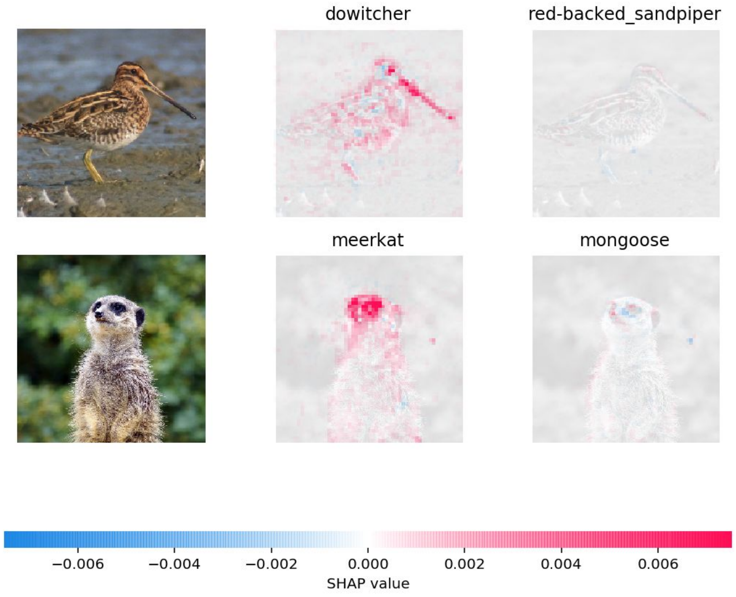
is given by:

$$\phi_i = \frac{\delta_1 + \delta_2 + \delta_3 + \delta_4 + \delta_5 + \delta_6 + \delta_7 + \delta_8}{8}$$

source: F. López, "SHAP: Shapley Additive Explanations," https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3

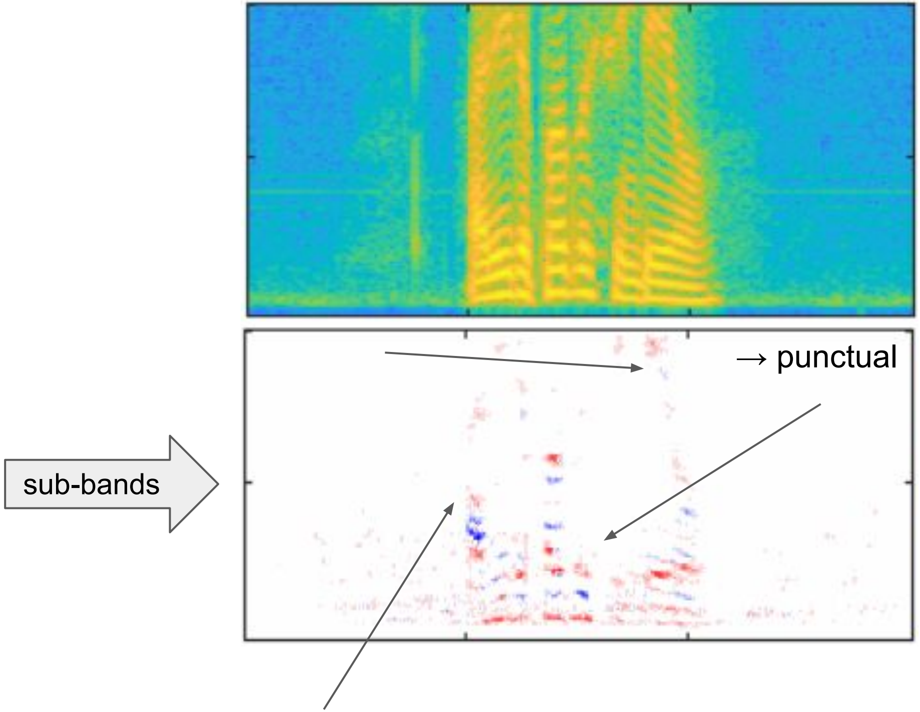
# SHapley Additive exPlanations (SHAP)

- Positive SHAP values represent parts of the image that the network considers important for the detection of that class
- But how to interpret it for speech signals?




source: https://github.com/slundberg/shap

A bonafide file LA_E_3757378 from ASVspoof 2019 LA



# SHAP analysis on ASVspoof 2019 LA [1]

- **Models**
  - 1D- and 2D- Res-TSSDNet [1] with raw waveform and STFT spectrogram as input
  - audio files are fed with original length during inference time to avoid concatenation

- **Post-processing**
  - only the highest 0.2% SHAP values are plotted

- **ASVspoof 2019 LA development partition**
  - examples are shown attack-wisely for the 6 seen attacks in train set

[1] W. Ge, J. Patino, M. Todisco, and N. Evans, “Explaining deep learning models for spoofing and deepfake detection with SHapley Additive exPlanations,” in ICASSP 2022.

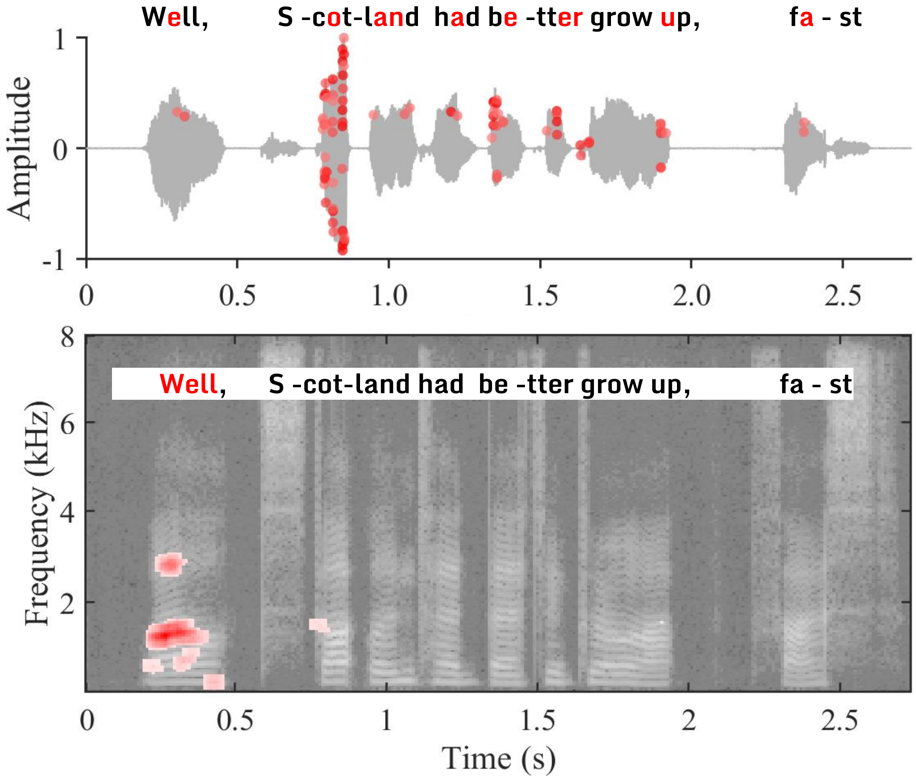
# SHAP analysis on ASVspoof 2019 LA

- **A01 - TTS attack with a WaveNet vocoder**

**mostly vowels**

**Leading 0.5s**

**Low frequency**



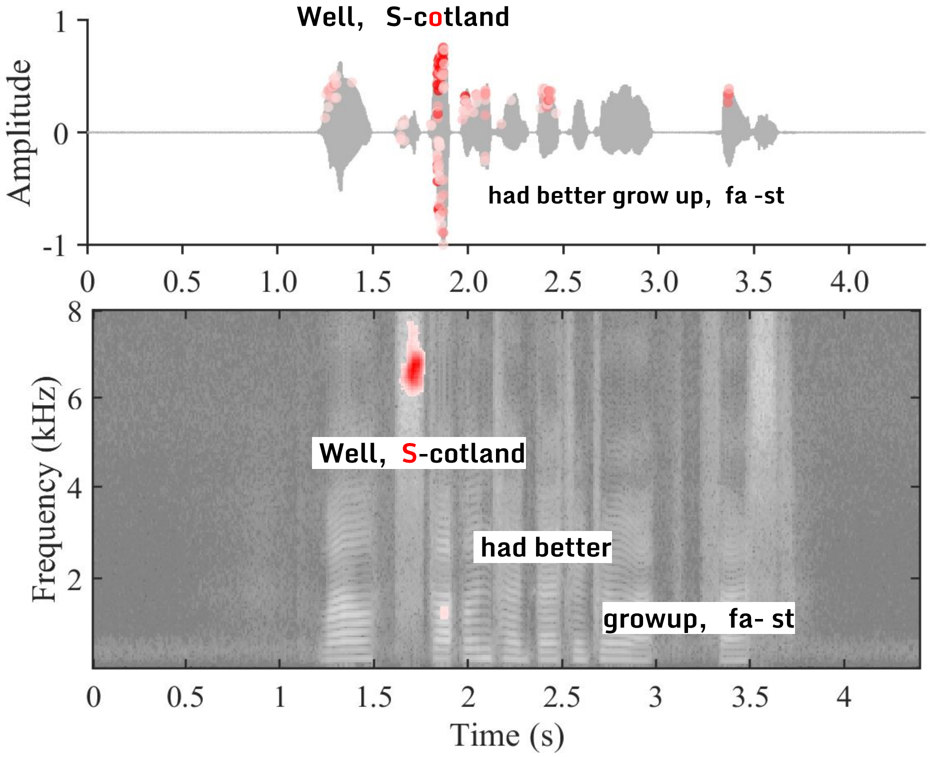
# SHAP analysis on ASVspoof 2019 LA

- **A02 - TTS attack with a WORLD vocoder**

mostly single vowel

unvoiced sound \s\

higher and lower frequency

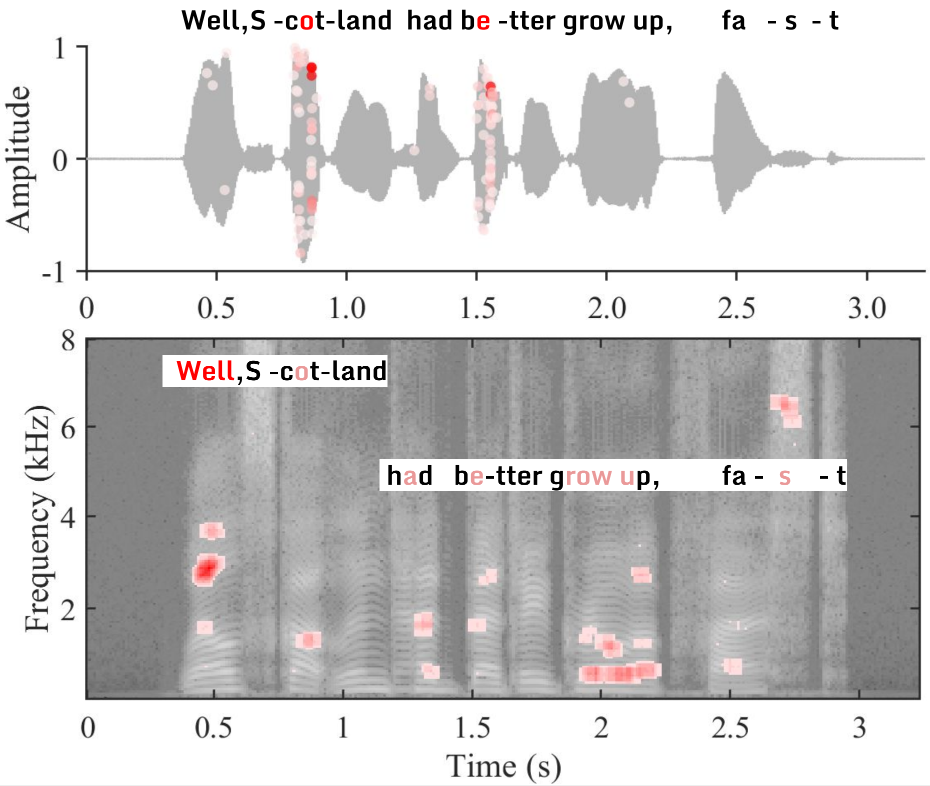


# SHAP analysis on ASVspoof 2019 LA

- **A03 - TTS attack with a WORLD vocoder**

less densely distributed in vowels

lower frequency

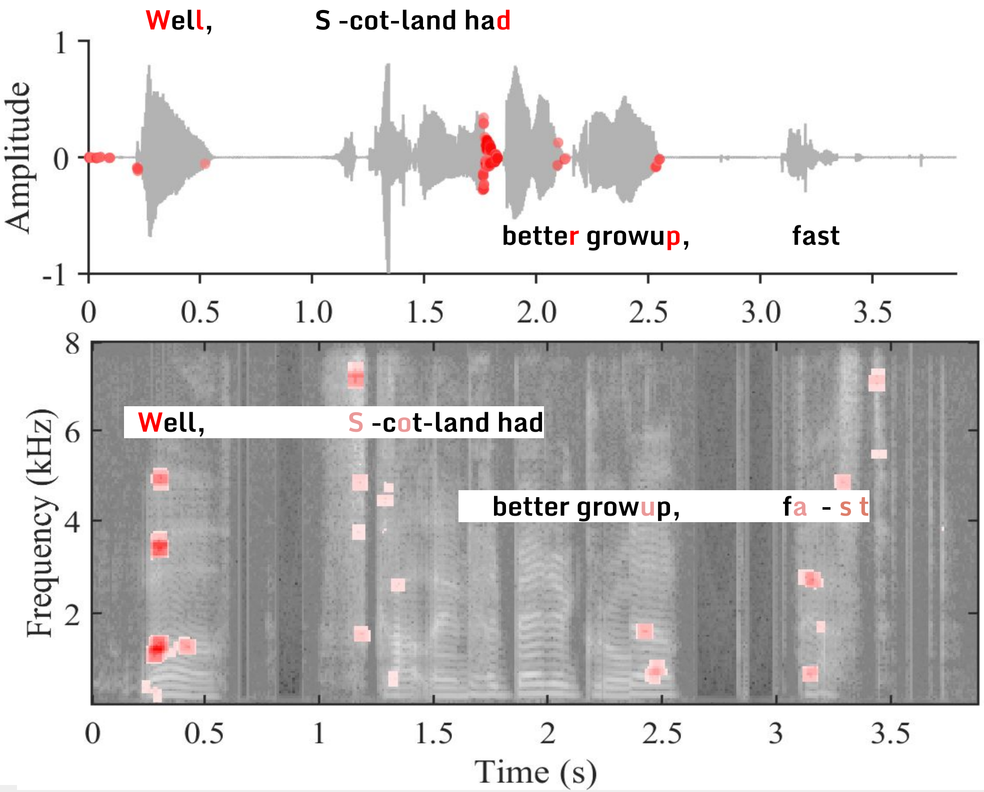


# SHAP analysis on ASVspoof 2019 LA

- **A04 - TTS (waveform concatenation)**

**non-speech, low energy speech segments (voice onsets and offsets)**

**full band, unvoiced speech, clicks**

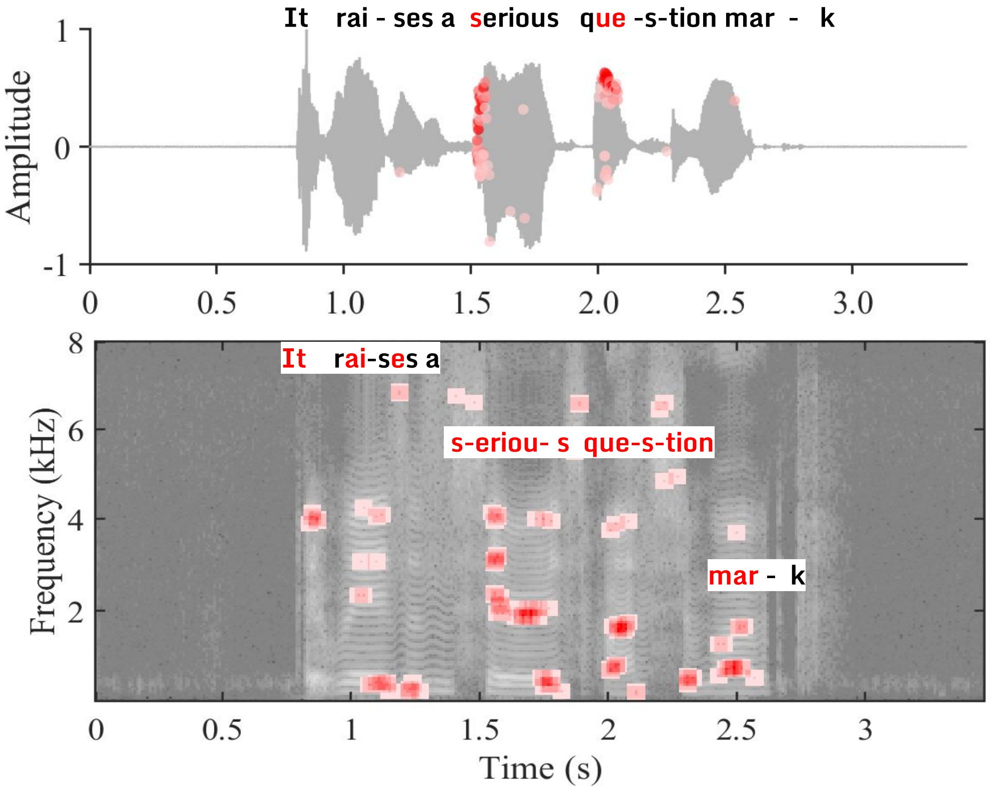


# SHAP analysis on ASVspoof 2019 LA

- **A05 - VC (NN-based)**

voice onset, vowels

full band, higher energy formant frequencies

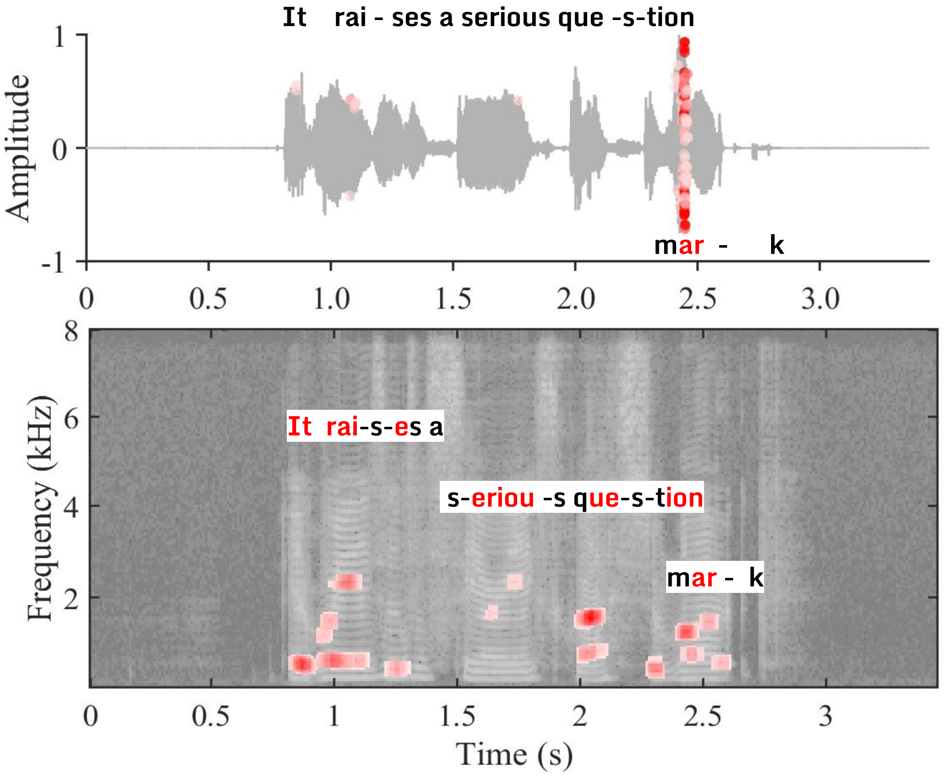


# SHAP analysis on ASVspoof 2019 LA

- **A06 - VC (transfer-function-based)**

vowel distortion

lower frequency



# SHAP analysis on ASVspoof 2019 LA

- Time-domain and spectral domain classifiers use different artefacts
- TTS (A01-03), vowels and lower frequency bands speech are generally more important for spoofing detection
- TTS A04 and VC (A05 and A06), artefacts distribution is different depending on the attack

Table 1: *Artefact description of attacks in ASVspoof 2019 LA train partition.*

| | | Found artefacts | |
|---|---|---|---|
| Attack | Algorithm | Waveform | Spectrogram |
| A01 | TTS | Vowels | Lower frequency bands, leading 0.5s |
| A02 | TTS | Single dominant vowel | Lower & higher frequency bands, unvoiced \s\ |
| A03 | TTS | Less densely distributed in vowels | Lower frequency bands |
| A04 | TTS | Non-speech, low energy speech segments (voice onsets and offsets) | Full spectrum, unvoiced speech, clicks |
| A05 | VC | Voice onset, vowels | Full spectrum, higher energy formant frequencies |
| A06 | VC | Speech distortion | Lower frequency bands |

# Conclusions

- **What we have learned**
  - there is no single countermeasure that works for all attacks
    - artefacts for different attacks have extremely different characteristics
  - performance of CMs degrades in real life scenarios
  - the fusion of several systems to increase complementarity is always necessary
    - not convenient for complexity and power consumption
  - (i) self-supervised models and (ii) data augmentation are good candidate for the detection
    - (i) need of huge, diverse data for training
    - (ii) need to be tailored to the problem to be tackled
- **Unsolved questions**
  - generalisation will be always a problem
    - new unseen attacks are always ready to break countermeasures
  - training on all types of attacks is impossible
- **Some ideas**
  - explainability and interpretability: artefacts seen from a physical point of view can help
    - physical-aware attention mechanism
  - one-class classification → anomaly detection

# Links to open-source codes

- RawNet2 (ASVspoof 2021 challenge baseline)
  - https://github.com/eurecom-asp/rawnet2-antispoofing
  - https://github.com/asvspoof-challenge/2021/tree/main/LA/Baseline-RawNet2

- RawGAT-ST - Spectro-Temporal Graph Attention Network
  - https://github.com/eurecom-asp/RawGAT-ST-antispoofing

- AASIST - Integrated Spectro-Temporal Heterogeneous Graph Attention Network
  - https://github.com/clovaai/aasist

- RawBoost: A Raw Data Boosting and Augmentation Method
  - https://github.com/TakHemlata/RawBoost-antispoofing

- SSL (wav2vec 2.0) for anti-spoofing
  - https://github.com/TakHemlata/SSL_Anti-spoofing

- SHapley Additive exPlanations for anti-spoofing
  - https://github.com/GeWanying/shap-anti-spoofing

# ASVspoof5 (the 2023 edition challenge)

Automatic Speaker Verification and Spoofing Countermeasures Challenge

## ASVspoof5

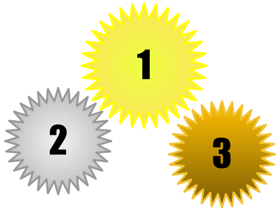
**We Need You!**

**Call For Spoofed/Speech DeepFake Data Contributors**

https://www.asvspoof.org/
info@asvspoof.org

| | |
|---|---|
| Database creation: | ongoing |
| Challenge set-up: | first half, 2023 |
| ASVspoof5 challenge: | second half, 2023 |

- focus on VC and TTS, including adversarial attacks (ASV/CM feedback)
- tentative data for creating attacks: Librispeech, LibriTTS, others (TBD) with noise/channel effects
- as in ASVspoof 2021, attack detection from degraded-quality data
- both CM-only and CM+ASV tasks




1st → 2000$ SGD
2nd → 1000$ SGD
3rd → 500$ SGD

*sponsored by* **A*STAR**

Agency for Science, Technology and Research
SINGAPORE