# An Ensemble-Based Approach for Generalized Detection of Spoofing Attacks to Automatic Speaker Recognizers

**João Monteiro**, Jahangir Alam, and Tiago H. Falk

MuSAE LAB

CRIM

INRS UNIVERSITÉ DE RECHERCHE

# Outline

1. Introduction and background
    a. Spoofing attacks
    b. Generalized setting
2. Approach description and model details
    a. Modeling approach
    b. Training
3. Evaluation
4. Conclusions

# Introduction and background

# Spoofing attacks

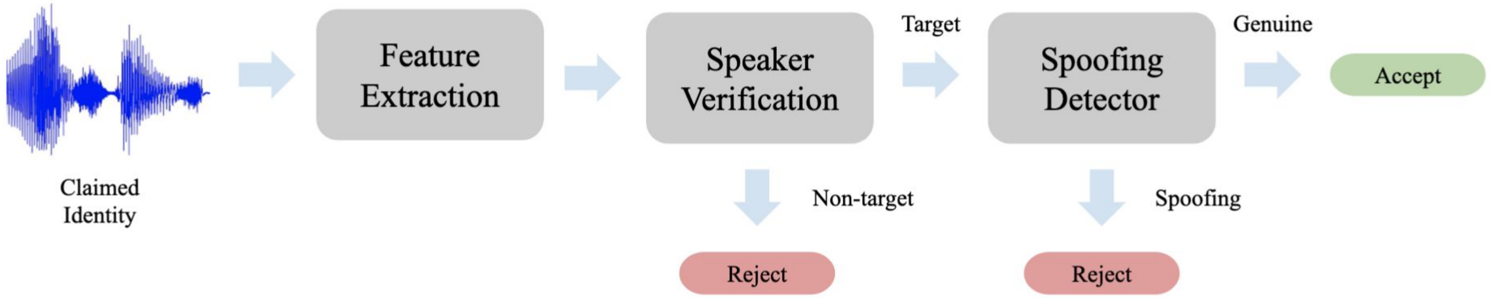
- Speaker recognizers are vulnerable to attacks trivially generated:
  - Replay someone's voice (***Physical access***)
  - Generate someone's voice using text-to-speech or voice conversion approaches (***Logical access***)

- Attack approaches, however, introduce detectable artifacts

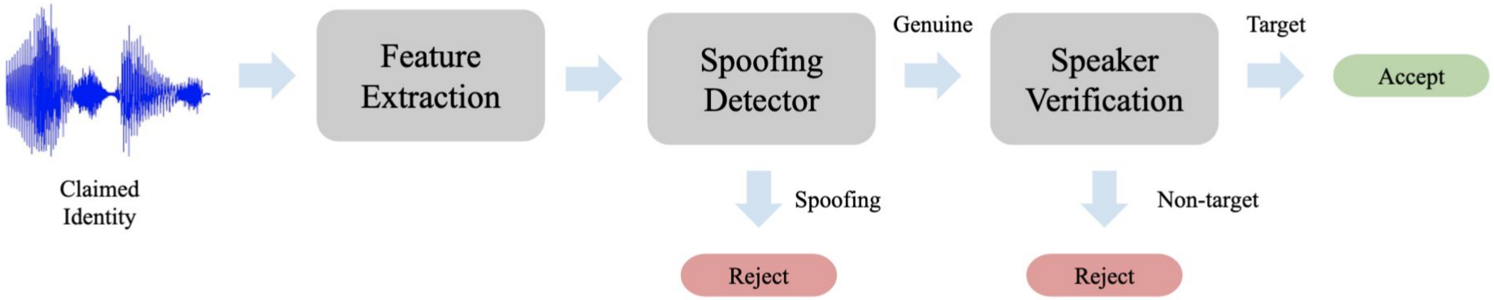
- Recent approaches rely on end-to-end detectors
  - Detectors can then be used in tandem with speaker recognizers

# Spoofing attacks

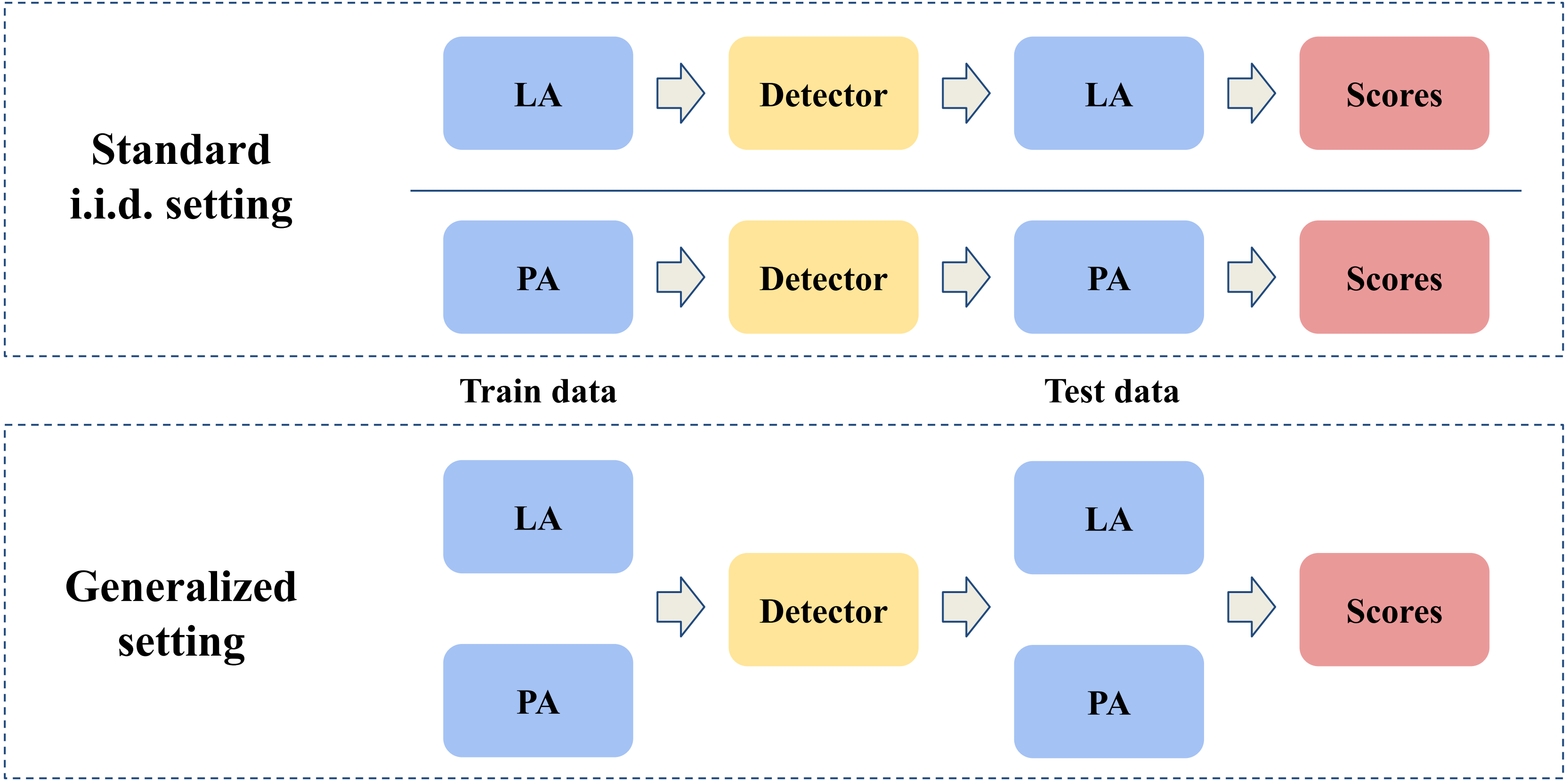
(A)




(B)



# Generalized setting



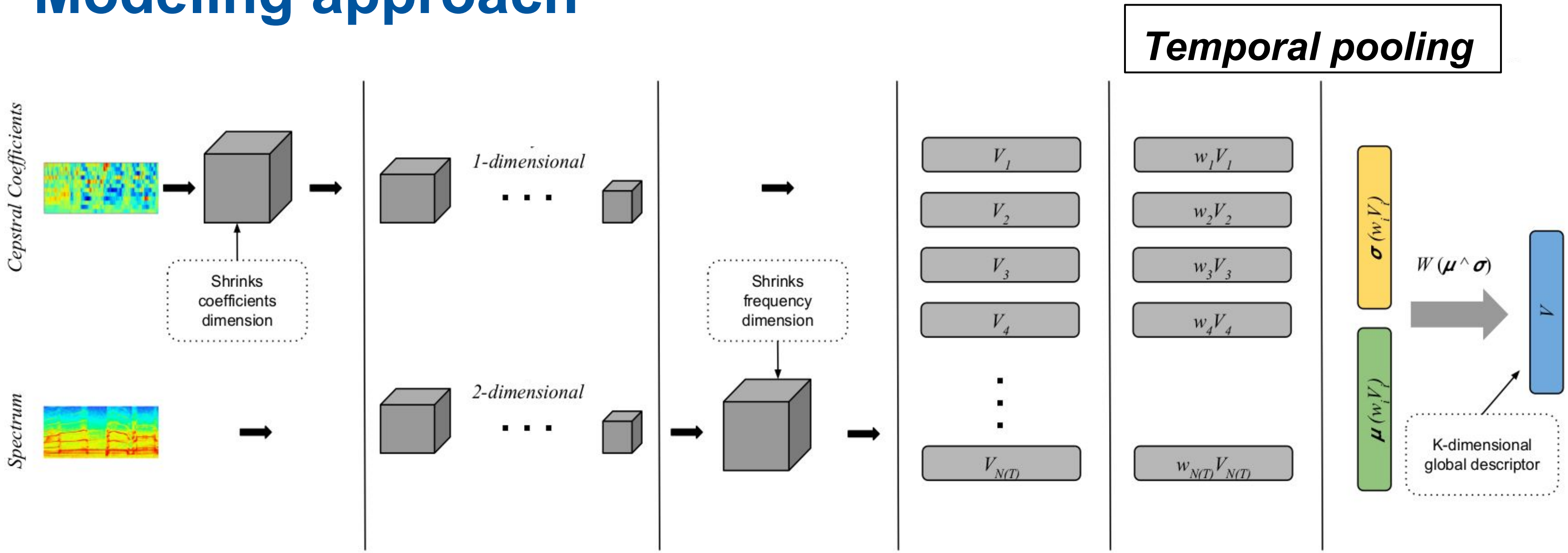
# Generalized setting

- Some recent approaches and benchmarks for detection of spoofing attacks do not reflect real life use cases:
    - Real detectors do not know in advance which approach the attacker will use
    - Detectors should be able to detect both LA and PA attacks

- We thus tackle that issue by:
    - Training detectors known to work well for LA/PA
    - Further training a third model which predicts the coefficient of a convex combination between the outputs of the other models

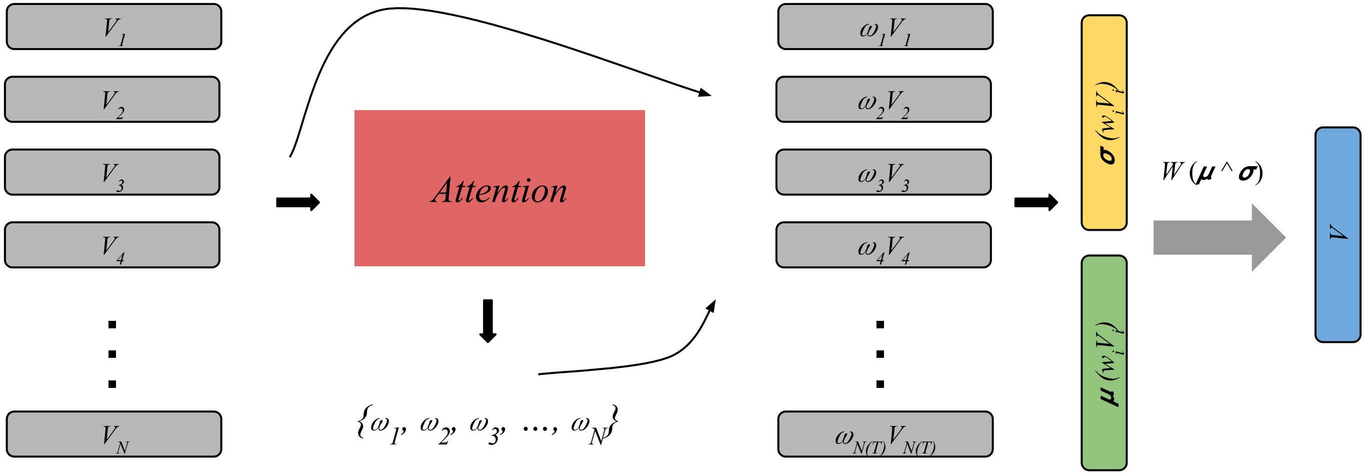
# Approach description and model details

# Modeling approach




- Different approach depending on input feature type
- $V$ is then projected into a final output score through an affine transformation learned along with the complete model

# Temporal pooling



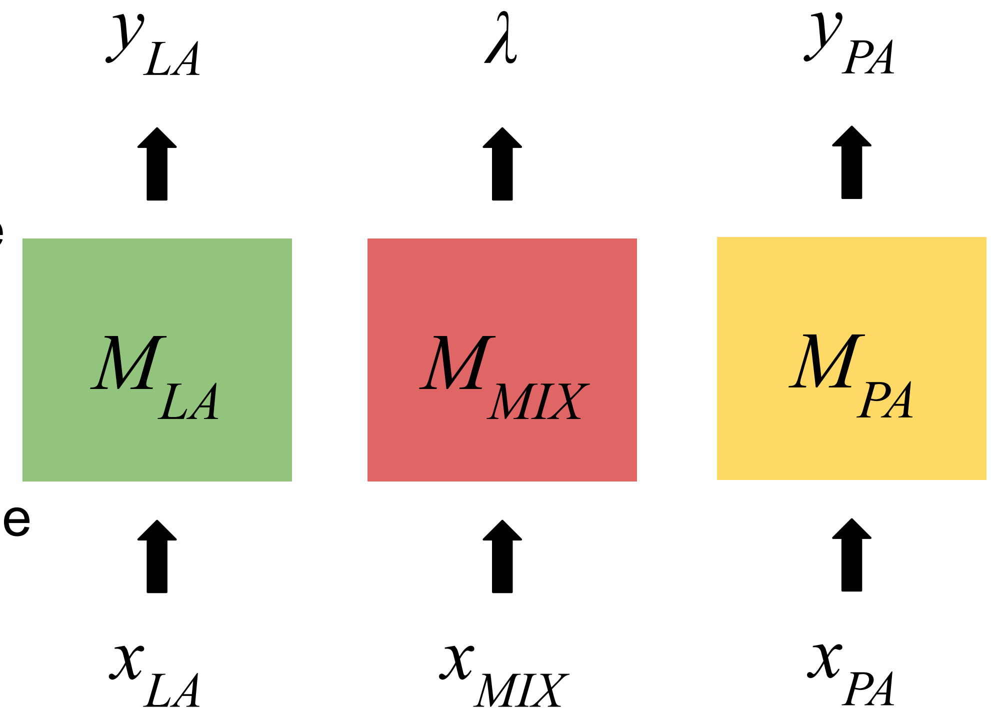

- *Summarizes* a sequence of local descriptors
- Allows processing of inputs of arbitrary length

$$a_i = \tanh(AV_i)$$

$$w_i = \frac{e^{a_i}}{\sum_{i=1}^{N(T)} e^{a_i}}$$

# Modeling approach

- Three independent models and features
- Mixture model learns to combine outputs of other models
- We chose LFCCs for the LA model and product spectra for the PA and mixture models



## Training

- Loss: binary cross entropy over combined outputs
- Unbalanced data: clean examples are oversampled; every mini batch is balanced
- Training is carried out with Stochastic Gradient Descent using mini-batches of effective size 16. Polyak's acceleration is also employed

# Evaluation

# Evaluation

- Data introduced for the ASVSpoof 2019 challenge. Two sub-challenges:
    - Logical access: attacks created with speech synthesis
    - Physical access: attacks created with simulated replay

| | # Speakers | # Recordings | | | |
|---|---|---|---|---|---|
| | | Logical Access | | Physical Access | |
| | | Bona fide | Spoof | Bona fide | Spoof |
| Training | 20 | 2580 | 22800 | 5400 | 48600 |
| Development | 20 | 2548 | 22296 | 5400 | 24300 |

# Evaluation - LA

| System Description | | Dev. | | Eval. | |
|---|---|---|---|---|---|
| | | **EER** | **min-tDCF** | **EER** | **min-tDCF** |
| *Privileged* [1] | CQCC-GMM | 0.43% | 0.0123 | 9.57% | 0.2366 |
| | LFCC-GMM | 2.71% | 0.0663 | 8.09% | 0.2116 |
| *Privileged* | LFCC-ResNet | 0.04% | 0.0004 | **6.38%** | **0.1423** |
| *Pooled data* | LFCC | 0.08% | 0.0023 | 14.38% | 0.3231 |
| | ProdSpec | 0.01% | 0.0002 | 12.77% | 0.2448 |
| | MGDCC | 0.27% | 0.0066 | 13.13% | 0.2953 |
| *Proposed - ResNet* | LFCC | 0.08% | 0.0021 | 15.84% | 0.3476 |
| | ProdSpec | 0.03% | 0.0002 | 15.73% | 0.2725 |
| | Lambda | 0.04% | 0.0004 | 13.12% | 0.2962 |
| | Mixture | **0.01%** | **0.0002** | 9.87% | 0.1890 |

# Evaluation - PA

| System Description | | Dev. EER | Dev. min-tDCF | Eval. EER | Eval. min-tDCF |
|---|---|---|---|---|---|
| *Privileged* [1] | CQCC-GMM | 9.87% | 0.1953 | 11.04 | 0.2454 |
| | LFCC-GMM | 11.96% | 0.2554 | 13.54 | 0.3017 |
| *Privileged* | ProdSpec-ResNet | 0.87% | 0.0232 | 1.98% | **0.0579** |
| *Pooled data* | LFCC | 2.39% | 0.0835 | 2.96% | 0.1017 |
| | ProdSpec | 0.85% | 0.0251 | 4.31% | 0.1538 |
| | MGDCC | 3.89% | 0.1174 | 5.99% | 0.1858 |
| *Proposed - ResNet* | LFCC | 1.87% | 0.0656 | 3.99% | 0.1408 |
| | ProdSpec | 3.80% | 0.1111 | 4.94% | 0.1479 |
| | Lambda | 1.32% | 0.0317 | 2.29% | 0.0641 |
| | Mixture | **0.78%** | **0.0275** | **1.75%** | 0.0606 |

# Conclusions

- Simple pooling strategies are not enough to recover the performance of specialized privileged detectors

- Proposed mixture approach is able to recover some of the lost performance when one moves from the standard i.i.d. to the generalized case
    - Outperformed the privileged baseline for the PA case

- Evaluation of mixture scores yields better performance than individual mixture components

- Future work: New underlying models as well as speech representations

# Thank you

joao.monteiro@emt.inrs.ca

https://github.com/joaomonteirof/e2e_antispoofing