End-to-end Detection of Attacks to Automatic Speaker Recognizers with Time-attentive Light Convolutional Neural Networks

João Monteiro[1,2], Jahangir Alam[1,2], and Tiago H. Falk[1]

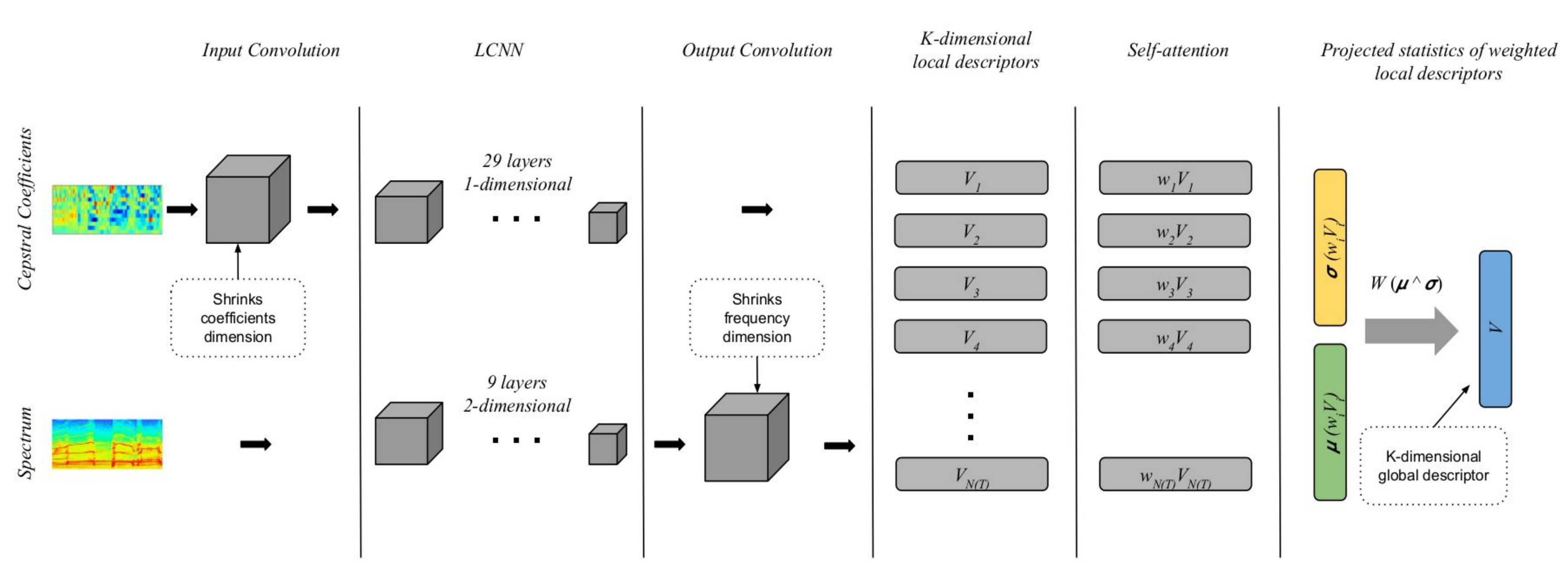
1-Institut National de la Recherche Scientifique (INRS-EMT)
2-Centre de Recherche Informatique de Montréal (CRIM)

Introduction

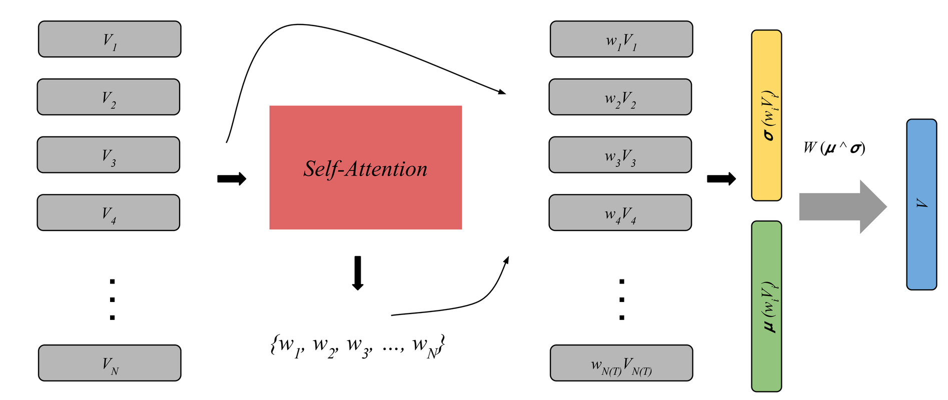
- We introduce an end-to-end setting for detection of spoofing attacks to speaker recognizers
  - End-to-end: Speech features directly mapped into scores indicating how likely the input is to be an attack
  - Single step training
- Both 2-dimensional convolutional models and time convolutions are evaluated on the data introduced for the ASVSpoof 2019

General setting

- Encoding of input audio into local descriptors
  - LCNNs are employed:
    - Fast to train
    - MFM activation
      - Variation of Maxout
      - Unlikely to overfit
  - 1-dimensional convolutions over the time dimension for the case of cepstral coefficients
  - 2-dimensional frequency-time convolutional models for the case fo spectral representations
- Attentive strategy for pooling into a global descriptor
  - Model learns how to discard uninformative frames
  - Allows processing of inputs with varying length
- Projection of statistics of weighted local descriptors is finally given to a fully connected classification layer



Attentive temporal pooling



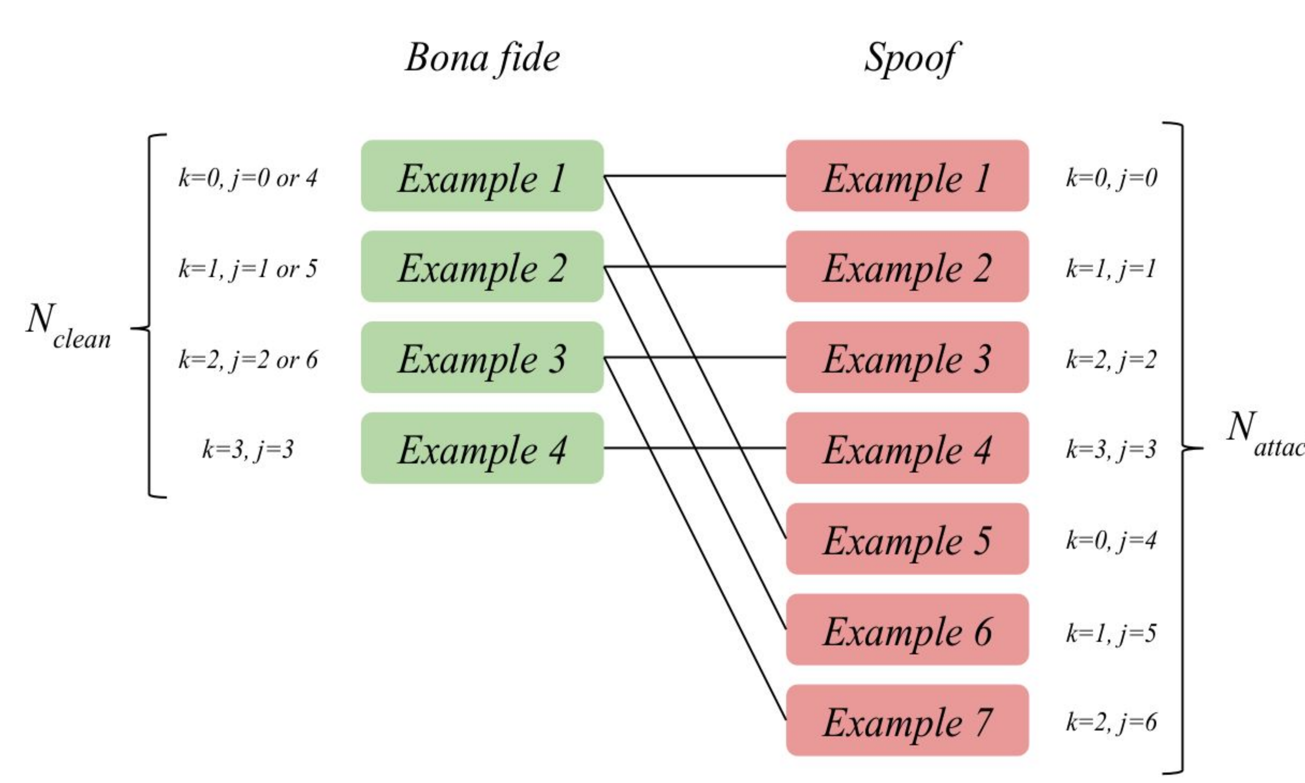
$a_i = \tanh(AV_i)$

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

$V = W[\mu(w_iV_i) \wedge \sigma(w_iV_i)]$

Training details

- *Sampling training examples:*
  - Ensure balanced minibatches



- *Online augmentation:* A random window of fixed duration is sampled every time an example is selected. Additionally, minibatches are cropped into a random duration prior to feeding in the model
- Training is carried out with Stochastic Gradient Descent using mini-batches of size 16 and 32 for the cases of spectral and cepstral coefficients, respectively. Polyak's momentum is also employed

Evaluation data

- Data Introduced for the ASVSpoof 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

Results

- *Logical Access*

	Feature-Model	EER(%)	t-DCF
ASVspoof benchmarks	LFCC-GMM	2.71	0.0663
	CQCC-GMM	0.43	0.0123
Internal baselines	CQCC-GMM	0.39	0.0110
	i-vector-PLDA	0.70	0.0210
Proposed	CQCC-LCNN29	1.07	0.0321
	LFCC-LCNN29	**0.20**	**0.0048**

- *Physical Access*

	Feature-Model	EER(%)	t-DCF
ASVspoof benchmarks	LFCC-GMM	11.96	0.2554
	CQCC-GMM	9.87	0.1953
Internal baselines	CQCC-GMM	9.70	0.1842
	i-vector-PLDA	9.17	0.2310
Proposed	CQCC-LCNN29	2.93	0.0752
	Spec-LCNN9	2.00	0.0488
	ProdSpec-LCNN9	**0.87**	**0.0232**

Conclusions

- We introduced variations of the LCNN architecture augmented with a self-attention mechanism so as to perform end-to-end detection of spoofing attacks
  - Introduced approach outperforms classical settings involving GMM classifiers
- In future work we intend to investigate the ability of end-to-end models in generalizing across attack strategies

joaomonteirof@gmail.com

http://musaelab.ca/