Attentive Filtering Networks for Audio Replay Attack Detection

Cheng-I Lai[1,2], Alberto Abad[1,3], Korin Richmond[1], Junichi Yamagishi[1,4], Najim Dehak[2], Simon King[1]

[1]Centre for Speech Technology Research, University of Edinburgh, UK
[2]Center for Language and Speech Processing, Johns Hopkins University, USA
[3]INESC-ID / Instituto Superior Técnico, University of Lisbon, Portugal
[4]Digital Content and Media Sciences Research Division, National Institute of Informatics, Japan

Research Problem & Our Objectives

Problem

Automatic speaker verification (ASV) systems are susceptible to malicious spoofing attacks, especially those in the form of audio replay.

(Left) An example of audio replay attack [1]. The left phone (black color) is a smart phone with a voice-unlock function for user authentication; the right phone (white color) replays a pre-recorded speech sample to unlock the left phone.

Objectives

1. Advance previous anti-spoofing research on DNN based system, and
2. Develop a system that automatically acquires and enhances discriminative features in both the time and frequency domain.

ASVspoof 2017 Version 2.0

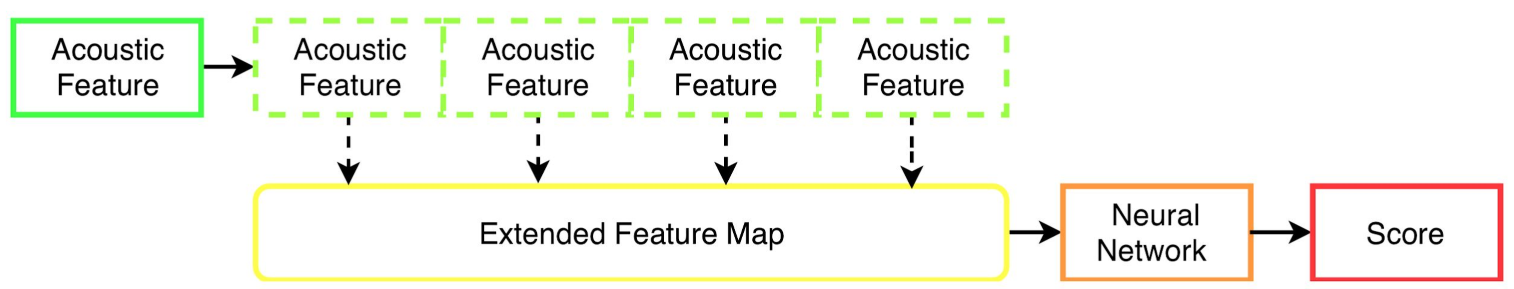
The ASVspoof 2017 corpus is a collection of *bona fide* and *spoofed* utterances. Bona fide utterances are a subset of the *RedDots* corpus, while the spoofed utterances are the result of replaying and recording bona fide utterances using a variety of heterogeneous devices and acoustic environments [2].

Subset	# Spk	# Replay sessions	# Replay Config	#Utterances	
				Bona fide	Replay
Training	10	6	3	1507	1507
Devel.	8	10	10	760	950
Eval.	24	161	57	1298	12008
Total	42	177	61	3565	14465

Unified Feature Map Creation

Acoustic Feature

Log magnitude spectrum (logpsec) is used as the acoustic feature. We kept all frames without applying VAD and applied mean normalization using a 3-s sliding window.



Extended Feature Map

The unified time-frequency map was created by extending all utterances to the length of the longest utterance by repeating their feature maps. The benefits of this feature engineering approach is that there is no need for feature truncation or frame-level score combination.

Code & Contact

Code: **github.com/jefflai108/Attentive-Filtering-Network**
Alternatively, you can reach the author at **clai24@mit.edu**

Attentive Filtering Network

Attentive Filtering

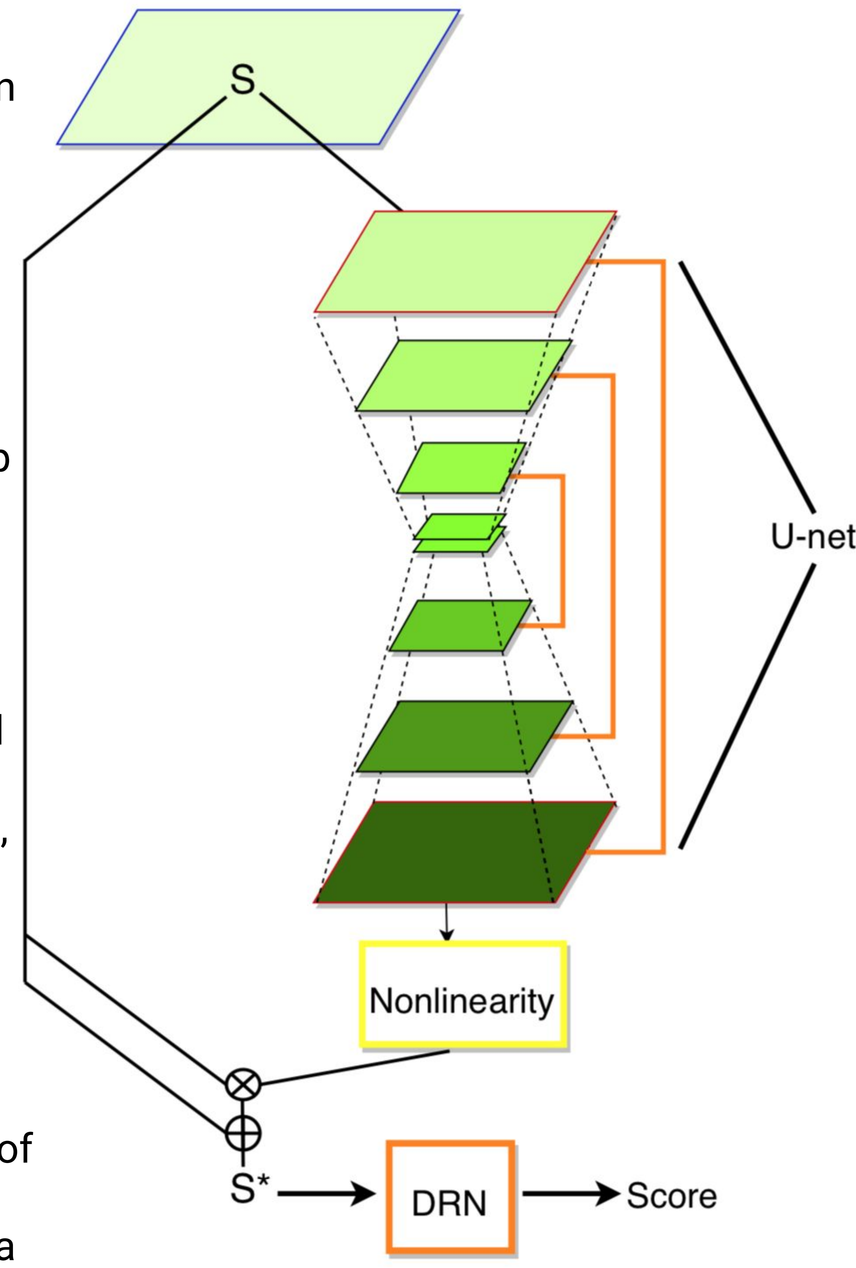
Attentive Filtering (AF) accumulates features in frequency and time domains selectively. AF augments every input feature map $\mathbf{S}$ with an attention heatmap $\mathbf{A_s}$ to produce an new feature map $\mathbf{S}^*$ for the DRN. Mathematically,

$$\mathbf{S}^* = \mathbf{A_s} \circ \mathbf{S} + \overline{\mathbf{S}}$$

We set $\overline{\mathbf{S}}$ as the residual $\mathbf{S}$. Attention heatmap $\mathbf{A_s}$ is described as,

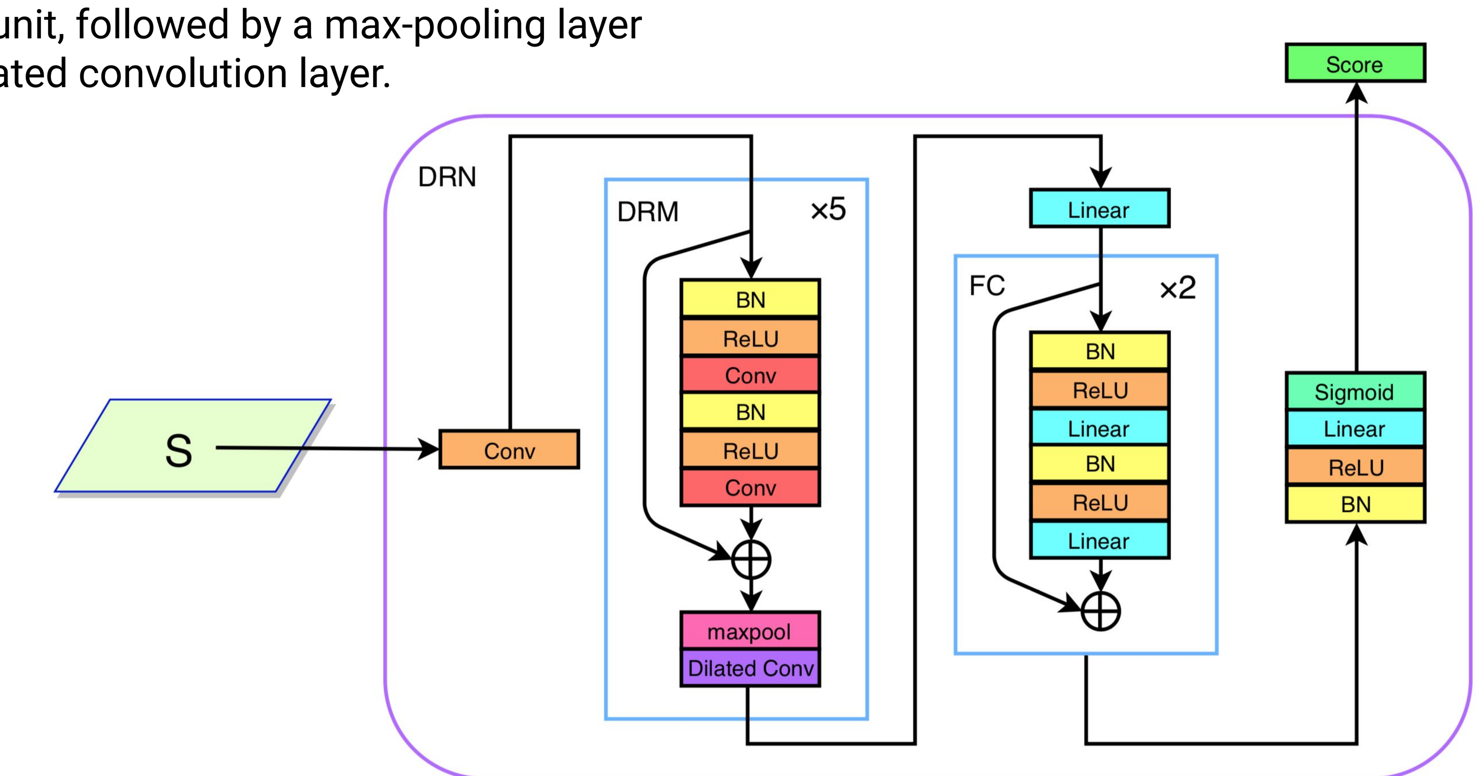
$$\mathbf{A_s} = \phi(U(\mathbf{S}))$$

ϕ is a nonlinear transform such as *sigmoid* or *softmax*, U is a U-net like structure, composed of a series of downsampling (bottom-up) and upsampling (top-down) operations. In addition, skip connections between the corresponding bottom-up and top-down components are added to help learn $\mathbf{A_s}$.



Dilated Residual Network

Dilated Residual Network (DRN) is composed of five Dilated Residual Modules (DRM) and two fully-connected (FC) modules. Each DRM has a residual unit, followed by a max-pooling layer and a dilated convolution layer.



Dilated Convolution

Dilated convolution operation $*_d$ is defined as,

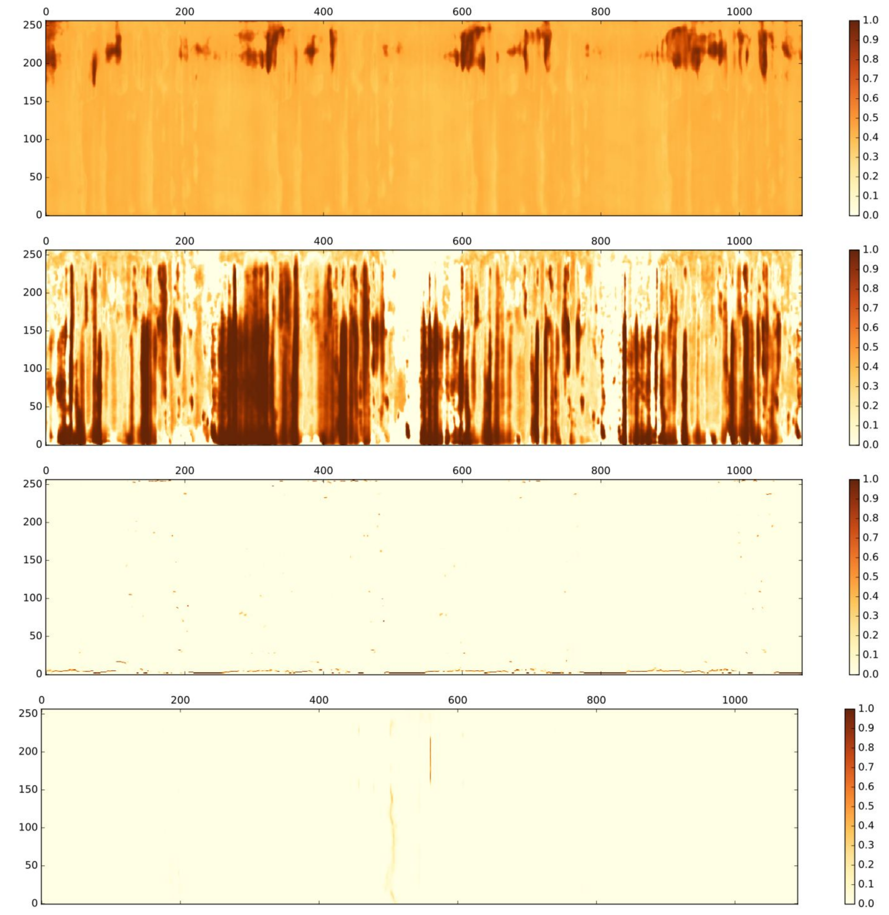
$$(F *_d G)(\mathbf{n}) = \sum_{\mathbf{m_1}+d\mathbf{m_2}=\mathbf{n}} F(\mathbf{m_1})G(\mathbf{m_2}), \forall \mathbf{m_1}, \mathbf{m_2}$$

where F is a feature map, G is a convolution kernel, and d is the dilation rate. With dilated convolution, the DNN's receptive field grows exponentially with layer depth such that it encodes more global knowledge.

Reference

[1] Zhizheng Wu, et al. "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.
[2] Héctor delgado, et al. "Asvspoof 2017 version 2.0: meta-data analysis and baseline enhancements," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 296– 303.

Attention Heatmap Visualization



Effect of ϕ

Visualizations of $\mathbf{A_s}$ with different ϕ (from top to bottom): *Sigmoid*, *Tanh*, *SoftmaxF*, and *SoftmaxT*.

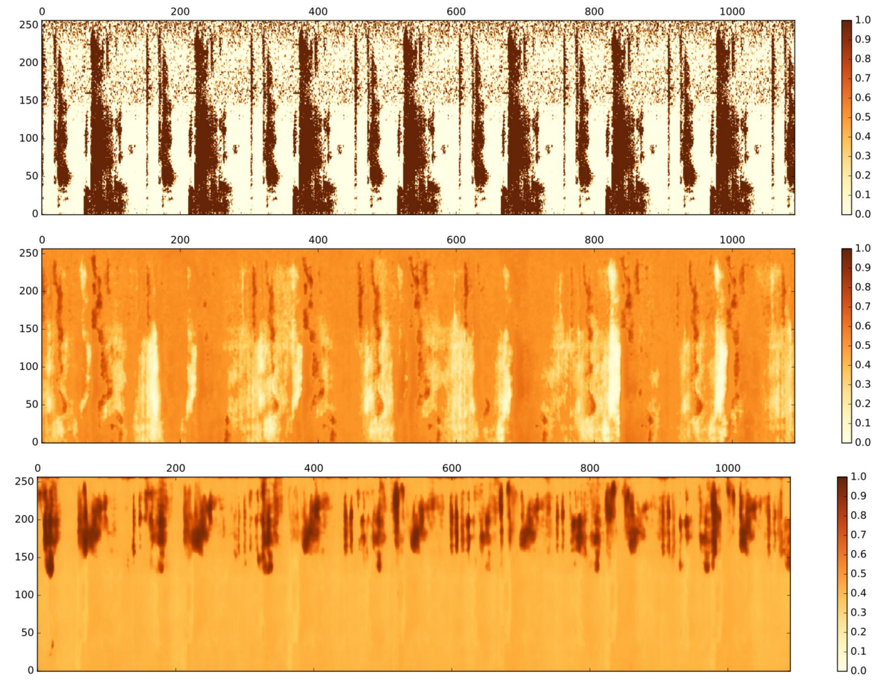
Sigmoid scales each dimension independently to range [0, 1], while *Softmax* scales each dimension dependently between to range [0, 1], implying only a few dimensions (either frequency bins or time frames) are activated and most are suppressed.

On the other hand, *Tanh* outputs in [-1, 1], and potentially losses useful information in $\mathbf{S}$.

Effect of $\overline{\mathbf{S}}$

Visualizations of $\mathbf{A_s}$ with and without $\overline{\mathbf{S}}$: input feature map $\mathbf{S}$ (top), $\mathbf{A_s}$ learned without $\overline{\mathbf{S}}$ (middle) and with $\overline{\mathbf{S}}$ (bottom).

With $\overline{\mathbf{S}}$, $\mathbf{A_s}$ strongly focuses on high frequency components of speech segments; without $\overline{\mathbf{S}}$, $\mathbf{A_s}$ learns a summary of $\mathbf{S}$.



Experimental Results

Systems	*dev* EER	*eval* EER	Diff.
Version 2 dataset			
AF(Sigmoid)+AF(SoftmaxT)	**6.09**	8.54	2.45
AF(Sigmoid)+AF(SoftmaxF)	6.37	8.80	**2.43**
AF(SoftmaxT)+AF(SoftmaxF)	6.39	8.98	2.59
AF(Sigmoid)-DRN(ReLU)	6.55	8.99	2.44
AF(SoftmaxT)-DRN(ReLU)	6.62	9.28	2.66
AF(SoftmaxF)-DRN(ReLU)	6.52	9.34	2.82
DRN(ELU)	7.49	10.16	2.67
AF(Tanh)-DRN(ReLU)	6.87	10.17	3.30
DRN(ReLU)	6.69	10.30	3.61
MDF(fusion)	-	**6.32**	-
qDFTspec	-	11.43	-
CQCC-GMM(CMVN)	9.06	12.24	3.18
i-vectors (Cosine Similarity)	8.99	14.77	5.78
i-vectors (Gaussian)	8.81	15.11	6.30
LCNN (our implementation)	6.47	16.08	9.61
Evolving RNN	18.7	18.20	-0.50
CQCC-GMM	12.08	29.35	17.27
Version 1 dataset			
DLFS(fusion)	3.98	**6.23**	2.25
MDF(fusion)	-	6.54	-
LCNN	4.53	7.34	2.81
ConvRBM(fusion)	**0.82**	8.89	8.07
Multi-task	4.21	9.56	5.35
ResNet	10.95	16.26	5.31

**Our reported numbers are averaged over 8 runs.*

Baselines

- CQCC-GMM
- i-vectors
- Light CNN

Single Systems

EERs of DRN and AF-DRN are reported. We can see that Attentive Filtering Network outperforms almost all previous work.

Fusions

Given visualizations of $\mathbf{A_s}$ we hypothesize that AF with different ϕ could be complementary, and as expected, fusing AF with *Sigmoid* and *Softmax* further reduces the EER.

Our fusion system provided a 30% relative improvement over the enhanced baseline system [2].