ResMax: Detecting Voice Spoofing Attacks with Residual Network and Max Feature Map

Il-Youp Kwak, Sungsu Kwag, Junhee Lee, Jun Ho Huh, Choong-Hoon Lee, Youngbae Jeon, Jeonghwan Hwang, Ji Won Yoon

Chung-Ang Univ., Samsung Research, Korea Univ.

Why Voice Spoofing Detection?

6-year-old orders $170 dollhouse, cookies with Amazon's Alexa

TV anchor says live on-air 'Alexa, order me a dollhouse' – guess what happens next

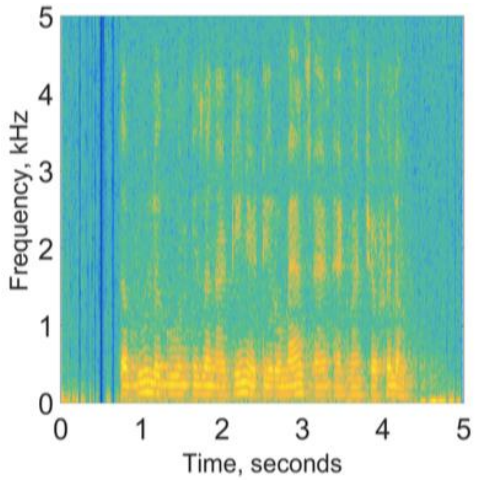
Story on accidental order begets story on accidental order begets accidental order

By Shaun Nichols in San Francisco 7 Jan 2017 at 00:58 244 SHARE ▼

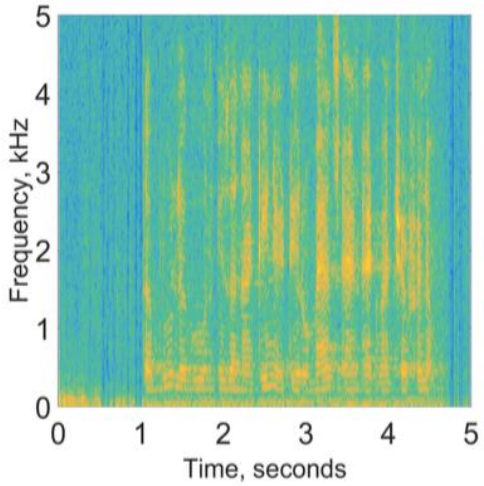
A San Diego TV station sparked complaints this week – after an on-air report about a girl who ordered a dollhouse via her parents' Amazon Echo caused Echoes in viewers' homes to also attempt to order dollhouses.

Voice Spoofing Detection, Physical Access (PA)

Genuine



Replayed

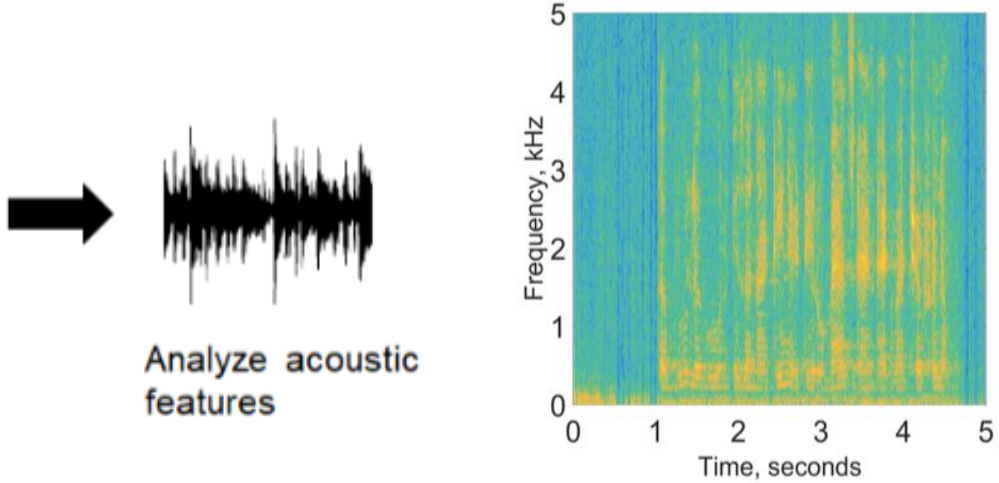


Voice Spoofing Detection, Logical Access (LA)

Genuine

Text to Speech (TTS): synthetic speech

Voice Conversion (VC): converted voice

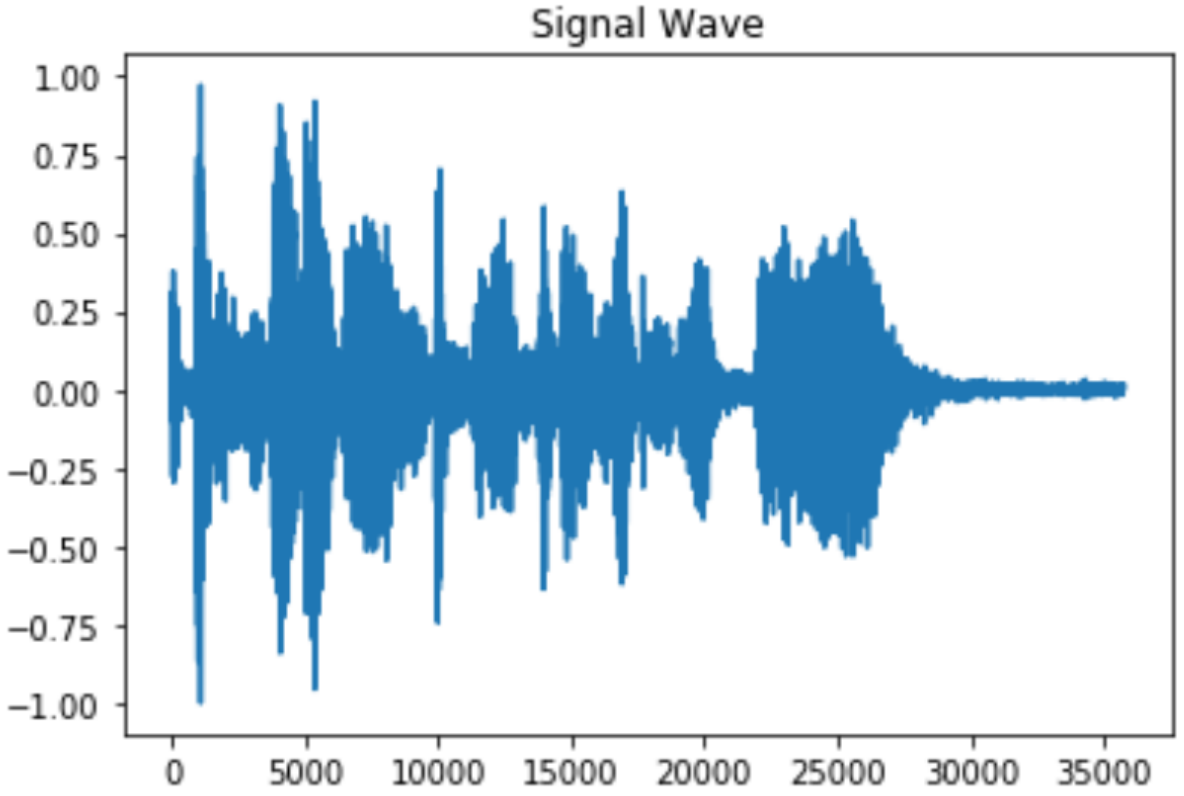


What is voice feature?

Each pixel is sum of RGB

3D array (24,24,3)

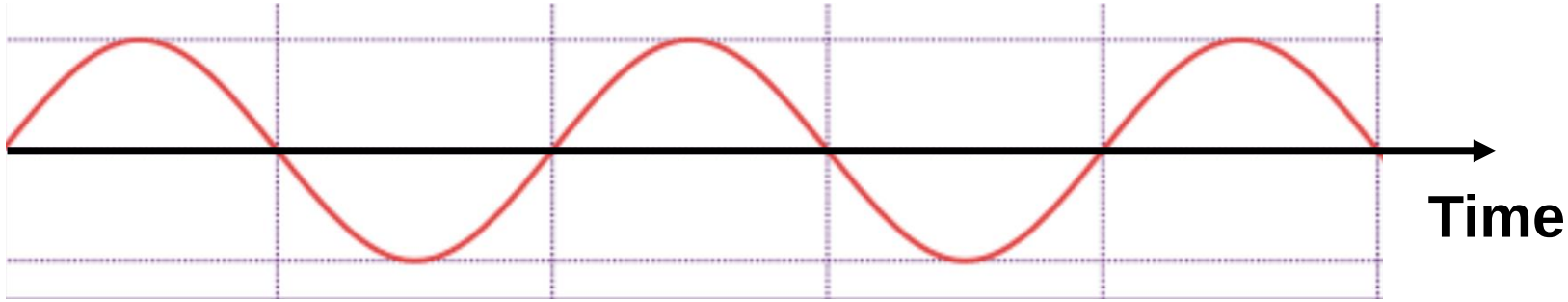
Each time interval is sum of frequency waves



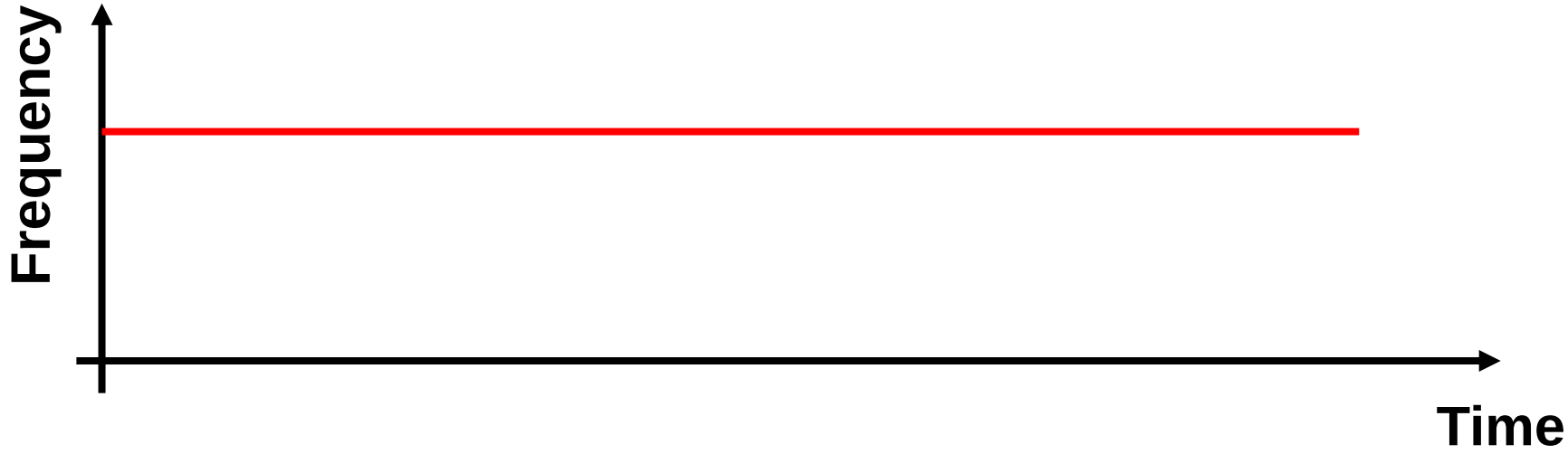
1D array (36000)

Better data representation with 2D array

1D array



2D array



Unlike Image data, we need feature engineering!

Binary Classification

- **Classifying Human or Speaker** $f : \mathbf{x} \xrightarrow{f_\theta} \mathbb{R}_{[0,1]}$
- **Dimension for x is (n_freq, n_time)**
- **Data:** $(\mathbf{x}, y) \quad \mathbf{x} \in \mathbb{R}^{(n_f, n_t)}, y \in \{0, 1\}$

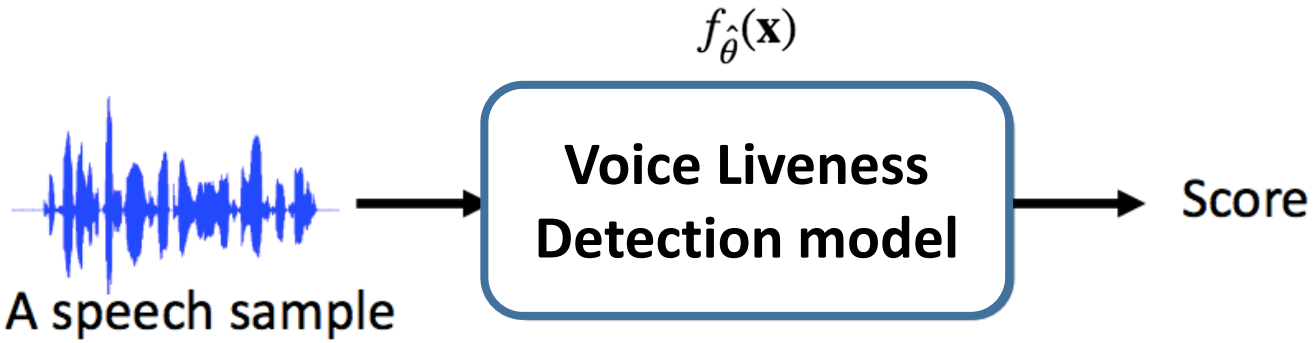
$$m \text{ training examples } \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)})\}$$

$$X = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}] \quad Y = [y^{(1)}, \ldots, y^{(m)}]$$

- Given $\mathbf{x} \in \mathbb{R}^{(n_f, n_t)}$, want $\hat{y} = P(y = 1|\mathbf{x}) = f_{\hat{\theta}}(\mathbf{x}) \in \mathbb{R}^{[0,1]}$
- **Objective is to minimize Cost, C(θ), w.r.t θ :**

$$C(\theta) = \sum_{i=1}^{m} L(f_\theta(\mathbf{x}_i), y_i)$$

What would be a good model($f_{\hat{\theta}}(\mathbf{x})$) ?



Automatic Speaker Verification Spoofing And Countermeasures Challenge (ASVspoof 2015, 2017 and 2019)

Wu et al. (2015) ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge

Kinnunen et al. (2017) The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection

Todisco et al (2019) ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection

Deep-learning based methods and ensemble solutions are dominating in voice liveness detection challenge

Top 2017 PA scenario

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

Using baseline CQCC features

DNN-based classifier
Other classifier

T: training
T+D: training + development

Top 2019 LA scenario

#	ID	t-DCF	EER
1	**T05**	0.0069	0.22
2	**T45**	0.0510	1.86
3	**T60**	0.0755	2.64
4	**T24**	0.0953	3.45
5	**T50**	0.1118	3.56
6	**T41**	0.1131	4.50
7	**T39**	0.1203	7.42
8	**T32**	0.1239	4.92
9	**T58**	0.1333	6.14
10	T04	0.1404	5.74

Top 2019 PA scenario

#	ID	t-DCF	EER
1	**T28**	0.0096	0.39
2	**T45**	0.0122	0.54
3	**T44**	0.0161	0.59
4	**T10**	0.0168	0.66
5	**T24**	0.0215	0.77
6	**T53**	0.0219	0.88
7	**T17**	0.0266	0.96
8	**T50**	0.0350	1.16
9	**T42**	0.0372	1.51
10	**T07**	0.0570	2.45

Grey — **Used Neural Networks**

Bold — **Used Ensemble**

Kinnunen et al. (2017) The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection

Todisco et al (2019) ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection

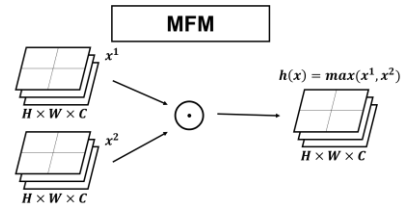
How to develop well performing light-weighted model?

TABLE I
ENSEMBLE SOLUTIONS FROM ASVSPOOF 2019 AND THE LIST OF MODELS USED.

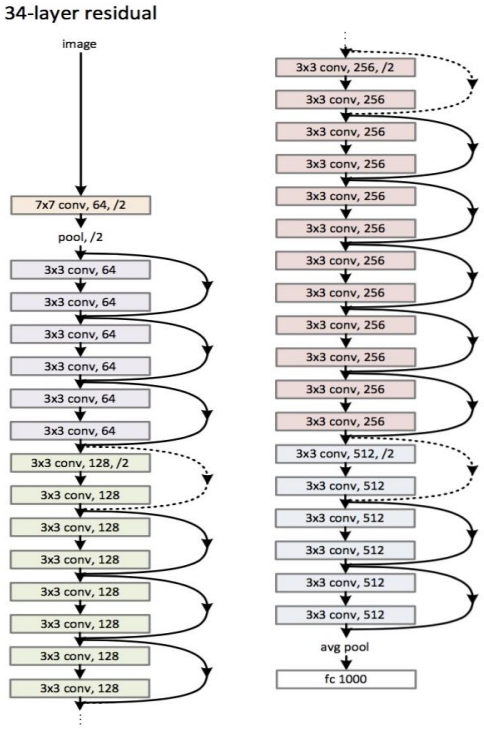
Model	Data	All models used
T10 [13]	PA	LFCC-ResNet, GD gram-ResNet .. Joint gram-ResNet
T44 [12]	PA	logspec-SENet34, CQCC-ResNet .. logspec-SENet50
T45 [6]	LA	LFCC-LCNN, LFCC-CMVN-LCNN .. CQT-LCNN
T50 [14]	PA	CQT-LCNN, LFCC-LCNN, DCT-LCNN
	LA	CQT-CGCNN, CQT-ResNet18 .. CQT-ResNet18IVec
T60 [15]	PA	FFT-CNN, FFT-CRNN, IMFCC-GMM, SVMs-IVec

LCNN architecture

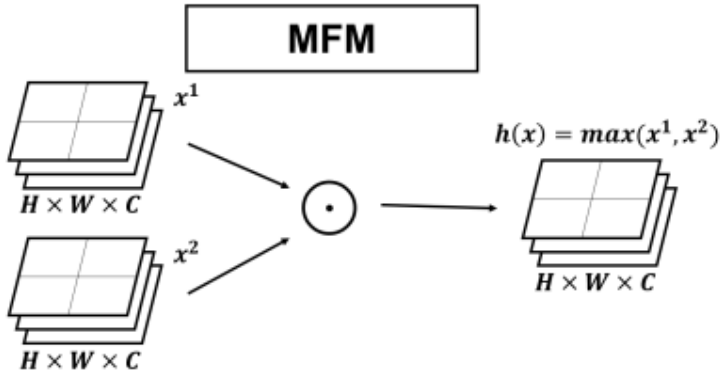
Type	Filter / Stride	Output	Params
Conv_1	5 × 5 / 1 × 1	863 × 600 × 64	1.6K
MFM_2	–	864 × 600 × 32	–
MaxPool_3	2 × 2 / 2 × 2	431 × 300 × 32	–
Conv_4	1 × 1 / 1 × 1	431 × 300 × 64	2.1K
MFM_5	–	431 × 300 × 32	–
BatchNorm_6	–	431 × 300 × 32	–
Conv_7	3 × 3 / 1 × 1	431 × 300 × 96	27.7K
MFM_8	–	431 × 300 × 48	–
MaxPool_9	2 × 2 / 2 × 2	215 × 150 × 48	–
BatchNorm_10	–	215 × 150 × 48	–
Conv_11	1 × 1 / 1 × 1	215 × 150 × 96	4.7K
MFM_12	–	215 × 150 × 48	–
BatchNorm_13	–	215 × 150 × 48	–
Conv_14	3 × 3 / 1 × 1	215 × 150 × 128	55.4K
MFM_15	–	215 × 150 × 64	–
MaxPool_16	2 × 2 / 2 × 2	107 × 75 × 64	–
Conv_17	1 × 1 / 1 × 1	107 × 75 × 128	8.3K
MFM_18	–	107 × 75 × 64	–
BatchNorm_19	–	107 × 75 × 64	–
Conv_20	3 × 3 / 1 × 1	107 × 75 × 64	36.9K
MFM_21	–	107 × 75 × 32	–
BatchNorm_22	–	107 × 75 × 32	–
Conv_23	1 × 1 / 1 × 1	107 × 75 × 64	2.1K
MFM_24	–	107 × 75 × 32	–
BatchNorm_25	–	107 × 75 × 32	–
Conv_26	3 × 3 / 1 × 1	107 × 75 × 64	18.5K
MFM_27	–	107 × 75 × 32	–
MaxPool_28	2 × 2 / 2 × 2	53 × 37 × 32	–
FC_29	–	160	10.2 MM
MFM_30	–	80	–
BatchNorm_31	–	80	–
FC_32	–	2	64
Total			10.2MM



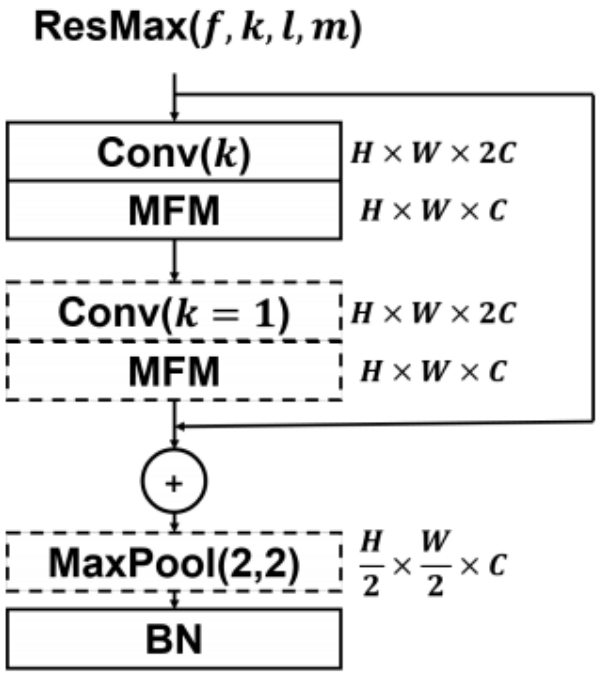
ResNet architecture



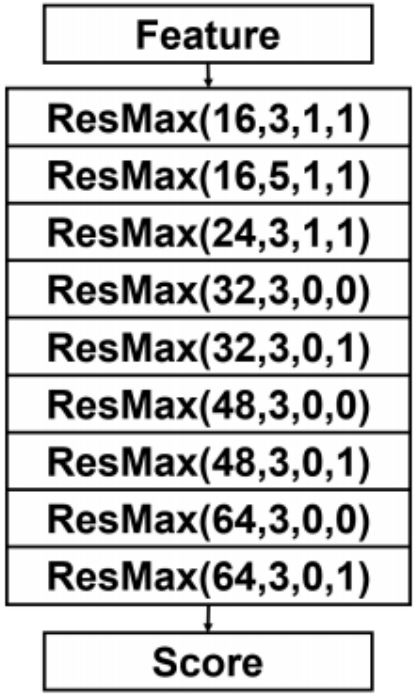
High performance of our light-weighted ResMax



(a) MFM



(b) ResMax Block



(c) Model Architecture

LA

#	*Model*	*t-DCF (Dev)*	*EER (Dev)*	*t-DCF (Eval)*	*EER (Eval)*	*#Mo*	*# Params*
1	T05	-	-	0.0069	0.22	-	-
2	T45	0.0000	0.000	0.0510	1.86	5	1484K
3	***CQT-1_100-ResMax***	0.0179	0.56	0.0600	2.19	1	262K
4	T60	0.0	0.0	0.0755	2.64	4	-
5	T24	-	-	0.0953	3.45	-	-
6	T50	0.027	0.90	0.1118	3.56	-	-
	T45 (FFT-LCNN)	0.0009	0.040	0.1028	4.53	1	371K
	T45 (LFCC-LCNN)	0.0043	0.157	0.1000	5.06	1	371K

PA

#	*Model*	*t-DCF (Dev)*	*EER (Dev)*	*t-DCF (Eval)*	*EER (Eval)*	*#Mo*	*# Params*
1	***CQT-1_120-ResMax***	0.0066	0.31	0.0091	0.37	1	262K
2	T28	-	-	0.0096	0.39	-	-
3	T45	0.0001	0.0154	0.0122	0.54	3	1113K
4	T44	0.003	0.129	0.0161	0.59	5	5811K
5	T10	0.0064	0.24	0.0168	0.66	6	1330K
6	T24	-	-	0.0215	0.77	-	-
	T28	-	-	-	0.50	1	-
	T45 (CQT-LCNN)	0.0197	0.800	0.0295	1.23	1	371K
	T44 (logspec-SENet)	0.015	0.575	0.0360	1.29	1	1344K

TABLE I
ENSEMBLE SOLUTIONS FROM ASVSPOOF 2019 AND THE LIST OF MODELS USED.

Model	Data	All models used
T10 [13]	PA	LFCC-ResNet, GD gram-ResNet .. Joint gram-ResNet
T44 [12]	PA	logspec-SENet34, CQCC-ResNet .. logspec-SENet50
T45 [6]	LA	LFCC-LCNN, LFCC-CMVN-LCNN .. CQT-LCNN
T50 [14]	PA LA	CQT-LCNN, LFCC-LCNN, DCT-LCNN CQT-CGCNN, CQT-ResNet18 .. CQT-ResNet18IVec
T60 [15]	PA	FFT-CNN, FFT-CRNN, IMFCC-GMM, SVMs-IVec

Non-speech part have information?

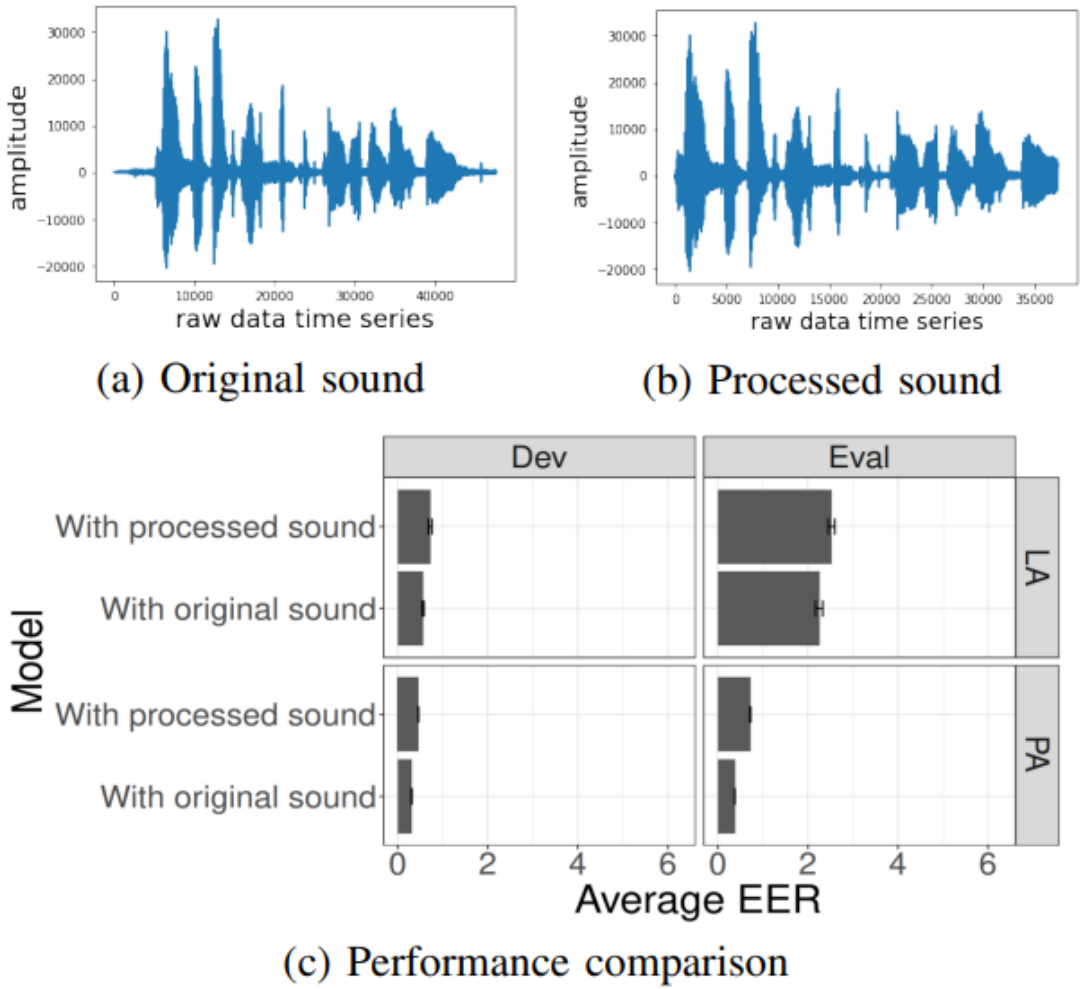


Fig. 3. The non-speech part remover suggested and tested. The ResMax model worked better without the non-speech part remover. The barplot indicates averaged EER with one standard deviation error bar.

The longer you listen the better the performance

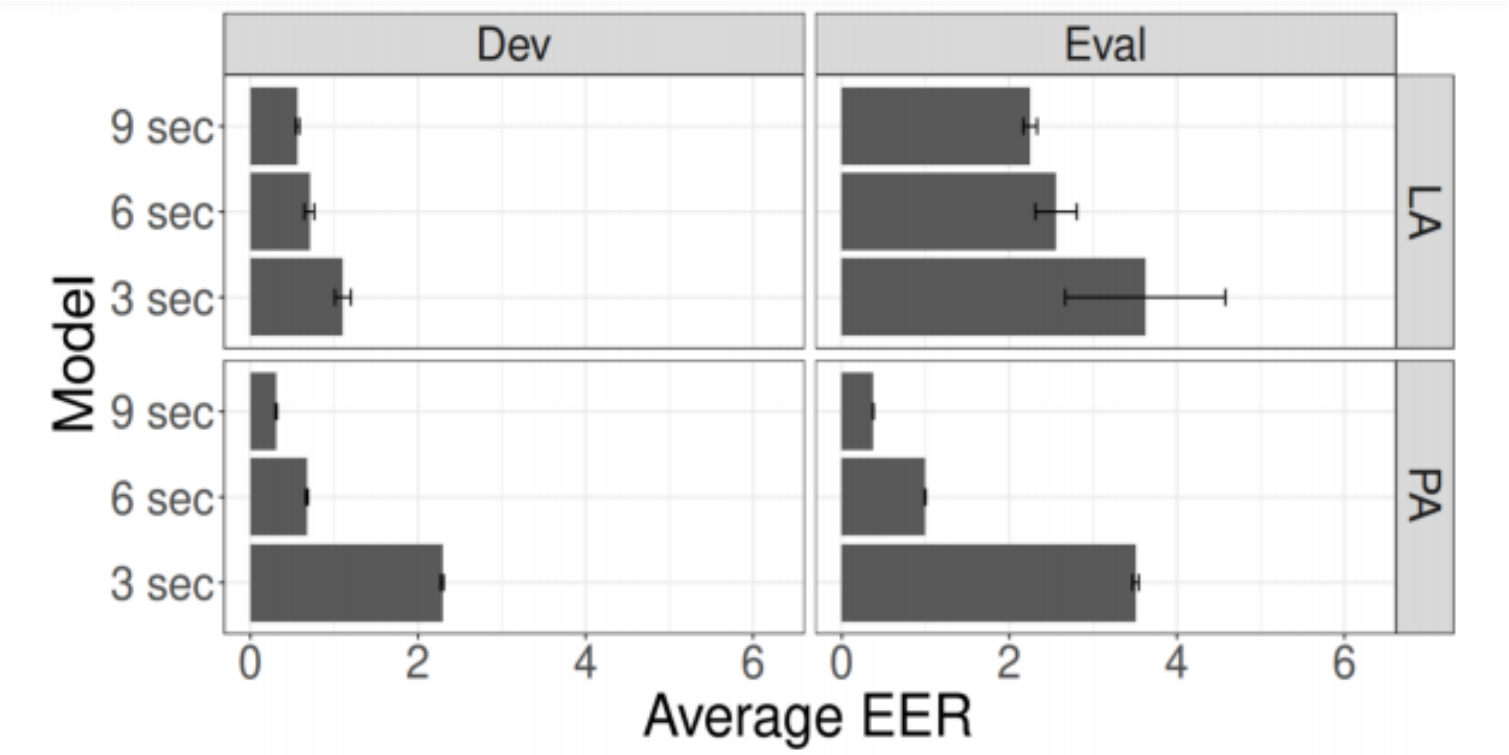


Fig. 4. The 9-second model performed best for both development and evaluation sets in LA and PA data. The barplot indicates averaged EER with one standard deviation error bar.

Performance depend on replay device quality

TABLE III

DETECTION PERFORMANCE ON THE ASVSPOOF2019 PHYSICAL ACCESS EVALUATION SETS IN VARIOUS ENVIRONMENTS. THE **A**, **B**, **C** REPRESENT THE CLASSES OF EACH FACTOR WHICH IS WELL DESCRIBED IN [5]. ALL NUMERICAL VALUES REPRESENT THE AVERAGE OF EER.

	Factors	**A**	**B**	**C**
Verification Env.	Room size (S)	0.0047	0.0044	0.0041
	T60 (R)	0.0055	0.0029	0.0038
	Talker-to-ASV distance	0.0059	0.0036	0.0042
Recording Env.	Attacker-to-talker distance (D_a)	0.0051	0.0036	0.0041
	Replay Device Quality (Q)	0.0067	0.0036	0.0009

High performance on best performing TTS, VC systems

ID	Type	Description	EER
A07	TTS	vocoder + GAN	0.0022
A08	TTS	neural waveform	0.0388
A09	TTS	vocoder	0.0003
A10	TTS	neural waveform	0.0045
A11	TTS	griffin lim	0.0039
A12	TTS	neural waveform	0.0002
A13	TTS,VC	waveform concatenation & filtering	0.0051
A14	TTS,VC	vocoder	0.0012
A15	TTS,VC	neural waveform	0.0030
A16	TTS	waveform concatenation	0.0039
A17	VC	waveform filtering	0.0561
A18	VC	vocoder	0.0225
A19	VC	spectral filtering	0.0317

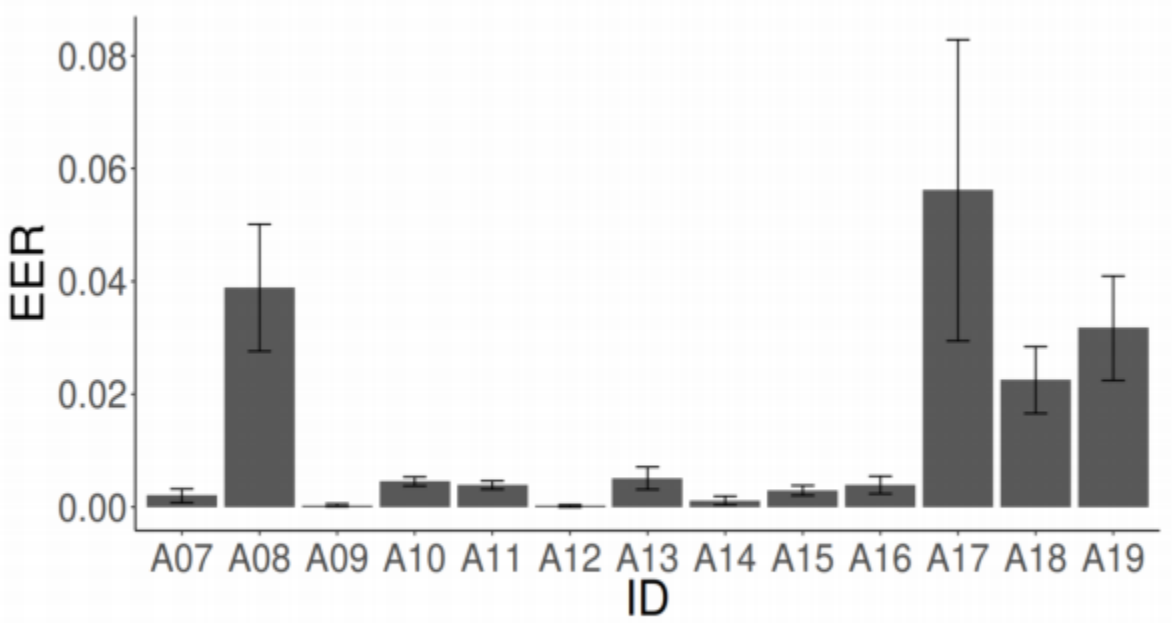


Fig. 5. The averaged EER for 13 attack types in evaluation set. The barplot indicate averaged EER with one standard deviation error bar.

Known attacks are A16,A19 and 4 others

Summary

Existing voice spoofing attack detection solutions have been designed without considering real-world model complexity and detection latency requirements

We combined the notions of skip connection (from ResNet) and max feature map (from Light CNN)

Our proposed model used only a single deep learning model with far fewer model parameters to outperform other models

Thank you!

ikwak2@cau.ac.kr