MDRT: MULTI-DOMAIN SYNTHETIC SPEECH LOCALIZATION

Amit Kumar Singh Yadav†, Kratika Bhagtani†, Sriram Baireddy†
Paolo Bestagini‡, Stefano Tubaro‡, Edward J. Delp†

†School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA
‡Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy

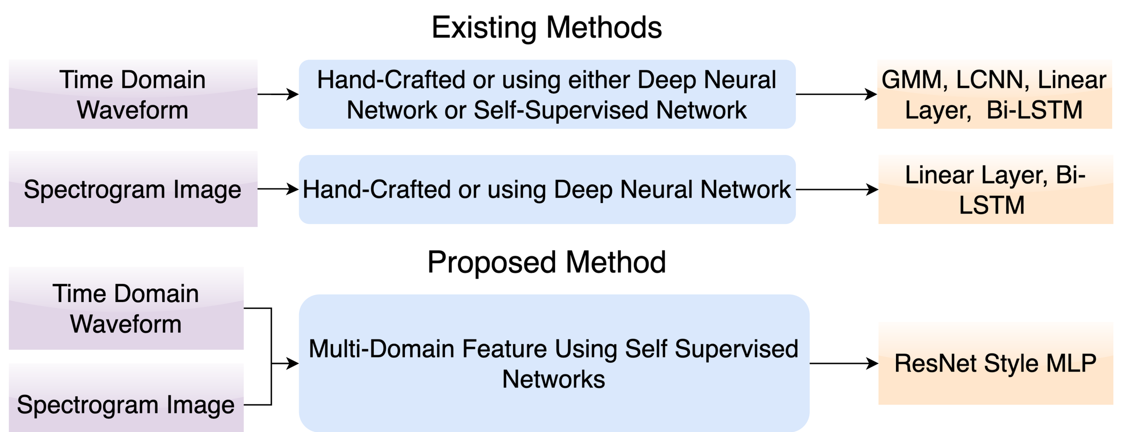
1 - Abstract

Context

- High quality synthetic speech impersonating human speaker is easily available and often misused in supporting fraud
- Limited work on localizing the synthetic segments within the speech signal

Goal

- To localize the synthetic speech segments in a partially synthetic speech signal
- Existing methods use single domain features, proposed Multi-Domain ResNet Transformer (MDRT) obtains multi-domain features to improve localization



2 - Introduction

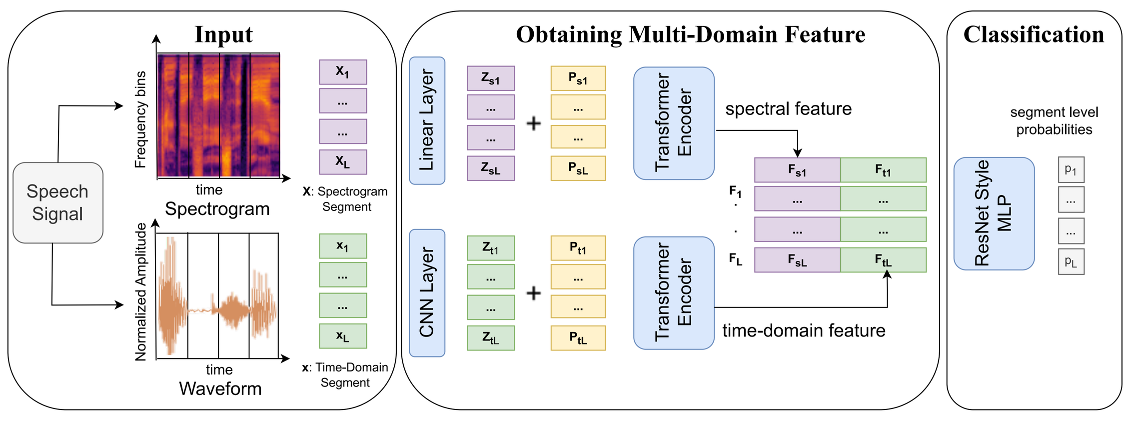
Problem formulation

- Consider $\mathbf{x}$ as the time domain speech signal and $\mathbf{X}$ as its corresponding mel-scale spectrogram representation
- Both can be divided into L non-overlapping segments, each of duration 20ms *i.e.*, $\mathbf{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L\}$ and the corresponding spectrogram $\mathbf{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_L\}$
- $\mathbf{x}$ and corresponding $\mathbf{X}$ have ground truth label $\mathbf{y} = \{y_1, y_2, \ldots, y_L\}$, s.t. $y_i \in \{0, 1\}$, where 0 and 1 indicate bona fide and synthetic speech segment, respectively
- Goal is to develop a localization model that classify each speech segment as bona fide or synthetic *i.e.*, provides the probability vector for the entire speech signal, $\mathbf{p} = \{p_1, p_2, \ldots, p_L\}$

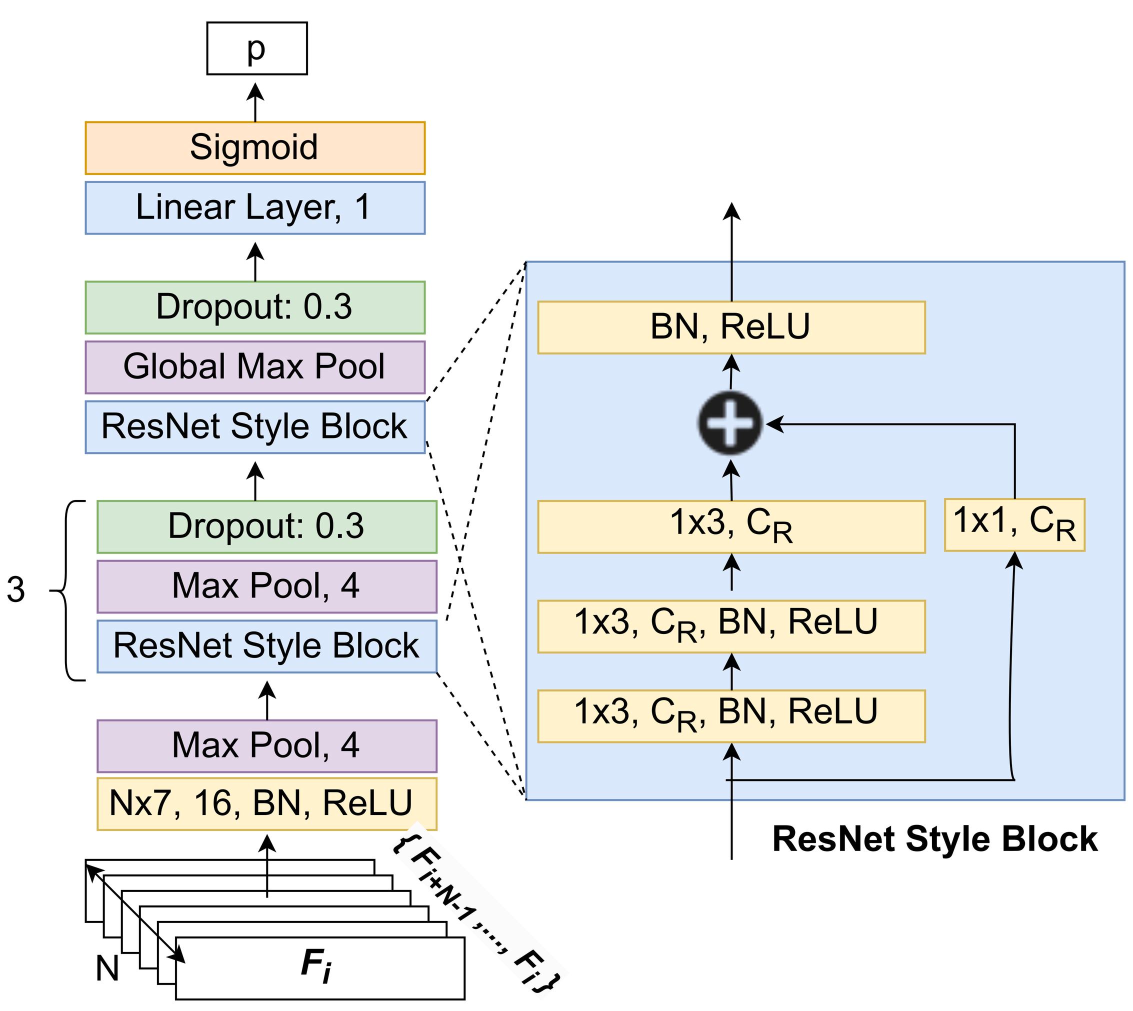
Evaluation Metric and Dataset

- Used Equal Error Rate (EER) as the performance metric, computed from Receiver Operating Characteristic curve
- EER is the rate where False Negative Rate and False Positive Rate are equal
- EER of 0% means perfect performance and EER of 50% corresponds to random performance
- Used PartialSpoof dataset, contains 25.4K training, 24.8K validation, and 71.2K evaluation speech signals

3 - Proposed Method



- MDRT processes the i-th time domain waveform segment $\boldsymbol{x}_i$ and corresponding i-th spectrogram segment $\boldsymbol{X}_i$ using a convolutional layer and linear layer to get latent representation vectors $\boldsymbol{Z}_{ti}$ and $\boldsymbol{Z}_{si}$, respectively, where $i = \{1, 2, \ldots, L\}$
- Positional encoding $\boldsymbol{P}_{ti}$ and $\boldsymbol{P}_{si}$ are added to $\boldsymbol{Z}_{ti}$ and $\boldsymbol{Z}_{si}$, respectively and the two latent representations obtained *i.e.*, $(\boldsymbol{Z}_{ti} + \boldsymbol{P}_{ti})$ and $(\boldsymbol{Z}_{si} + \boldsymbol{P}_{si})$ are processed by self-supervised pre-trained transformer neural networks: Wav2Vec2-Base and M2D4Speech
- The time domain feature $\boldsymbol{F}_{ti}$ obtained by Wav2Vec2-Base and the spectral feature $\boldsymbol{F}_{si}$ obtained by M2D4Speech are concatenated to obtain multi-domain feature vector $\boldsymbol{F}_i$
- Therefore, for each $\boldsymbol{x}_i \in \mathbf{x}$ and $\boldsymbol{X}_i \in \mathbf{X}$, we obtain a multi-domain feature vector $\boldsymbol{F}_i \in \mathbf{F}$, where $\mathbf{F} = \{\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_L\}$
- MDRT processes obtained multi-domain features using a novel ResNet-style Multi Layer Perceptron (MLP)
- Contrary to existing ResNet-style MLP that perform single channel convolution, the ResNet-style MLP used in MDRT performs convolution on multi-channel features
- Each channel in ResNet-style MLP corresponds to multi-domain feature $\boldsymbol{F}_i$ obtained from a different speech segment. So, multi-channel convolution helps to capture temporal artifacts from consecutive speech segments and the dropout layer improves generalization



- In the above figure, N=1 for 20ms segments, N=2 for 40ms and so on. MDRT is tested on localizing 20ms, 40ms, 80ms and 160ms synthetic segments in a partially synthetic speech

4 - Experimental Results

Improved Localization Performance

- MDRT has best performance w.r.t 12 existing methods on PartialSpoof dataset for localizing synthetic speech segments of duration 160ms. It uses approximately half the number of parameters (≈182M) as in method $B12$ (≈317M)

Method Name	Training Sample Duration	Feature Used	Classification Network	EER (in %)
$B01$	utt.	LFCC	LCNN+BLSTM	40.20
$B02$	160ms			16.21
$B03$	utt.	LFCC	H-MIL	33.12
$B04$	utt.		LS-H-MIL	31.96
$B05$	utt.	LFCC	SELCNN+BLSTM	44.00
$B06$	160ms			15.93
$B07$	160ms, utt.			20.04
$B08$	160ms, utt.			17.75
$B09$	160ms, utt.			17.55
$B10$	160ms, utt.			17.77
$B11$	160ms, utt.			16.60
$B12$	20ms∼640ms	W2V2-L	5 gMLP	9.24
MDRT	160ms	W2V2-B+M2D	ResNet-style MLP	**8.82**

- MDRT also performs better than existing methods for localizing synthetic speech segments of duration smaller than 160ms

Testing Duration	Method Name	Training Sample Duration	Feature Used	Classification Network	EER (in %)
20ms	$B13$	20ms	CQCC	LCNN	27.17
	$B12$	20ms∼160ms	W2V2-L	5 gMLP	12.84
		20ms	W2V2-B+M2D	ResNet-style MLP	**11.14**
40ms	$B12$	20ms∼160ms	W2V2-L	5 gMLP	11.94
		40ms	W2V2-B+M2D	ResNet-style MLP	**10.18**
80ms	$B12$	20ms∼160ms	W2V2-L	5 gMLP	10.92
		80ms	W2V2-B+M2D	ResNet-style MLP	**9.82**

Ablation Study

Indicates:
- effectiveness of proposed ResNet-style MLP classification network than existing networks
- better performance by processing features from multi-domain than single-domain
- benefit of using augmentation and dropout

Hyperparameters/ Configuration		EER (in %)
Classification Network	1 FC Layer	2.89
	1 BLSTM+1FC Layer	2.84
	2 BLSTM+1FC Layer	2.49
	1 gMLP	4.28
	5 gMLP	2.22
	ResNet-style MLP	**1.97**
Feature Choice	all-hidden layer feature	1.97
	last hidden layer feature	**1.93**
Domain	Single Domain (only time-domain)	1.93
	Single Domain (only spectrogram)	2.20
	Multi Domain (both)	**1.66**
Augmentation & Dropout	w/o aug. and dropout	1.66
	w/ aug.	1.64
	w/ aug. and dropout	**1.54**

5 - Conclusion

- Proposed a novel Multi-Domain ResNet Transformer (MDRT) for localizing synthetic speech segments
- MDRT performs better than several existing methods that use single-domain features
- MDRT uses half the number of parameters than the most promising existing method
- Future research will investigate performance of MDRT on synthetic speech from recent diffusion-based generators, and robustness to compressed and noisy speech signals