SLIM: Style-LInguistics Mismatch Model for Generalized Audio Deepfake Detection

Y. Zhu[1,2] S. Koppisetti[1] T. Tran[1] G. Bharaj[1]

[1]Reality Defender

[2]Institut National de la Recherche Scientifique (INRS)

NeurIPS 2024

Table of Contents

1. Background
2. System description
3. Experiment results and discussions

State-of-the-art detection methods

Typical pipeline

Waveform → Upstream encoder (e.g., w2v) → Downstream classifier

- Data augmentations
  - RawBoost (Tak et al. 2022)
  - Vocoded data (Wang and Yamagishi 2023)

Reality Defender

State-of-the-art detection methods

Typical pipeline

Waveform → Upstream encoder (e.g., w2v) → Downstream classifier

- Data augmentations
  - RawBoost (Tak et al. 2022)
  - Vocoded data (Wang and Yamagishi 2023)
- Improving classifiers
  - AASIST (Jung et al. 2022)
  - MFA (Guo et al. 2024)

Reality Defender

State-of-the-art detection methods

Typical pipeline

Waveform → Upstream encoder (e.g., w2v) → Downstream classifier

- Data augmentations
  - RawBoost (Tak et al. 2022)
  - Vocoded data (Wang and Yamagishi 2023)
- Improving classifiers
  - AASIST (Jung et al. 2022)
  - MFA (Guo et al. 2024)
- Others
  - Model distillation (Lu et al. 2024)
  - Ensemble multiple upstream representations (Yang et al. 2024)
  - Full finetune

Reality Defender

Interpretation methods

- Using knowledge-based features
  - Vocal tract (Blue et al. 2022)
  - Breathing (Layton et al. 2024)
  - Emotions (Conti et al. 2022)

🤔 Performance-interpretability tradeoff

Reality Defender

Interpretation methods

- Using knowledge-based features
  - Vocal tract (Blue et al. 2022)
  - Breathing (Layton et al. 2024)
  - Emotions (Conti et al. 2022)

  🤔 Performance-interpretability tradeoff
- Post-hoc XAI methods
  - Saliency map

  🤔 Sensitive to hyperparam setup

Goal

Typical pipeline

Waveform $\rightarrow$ Upstream encoder (e.g., w2v) $\rightarrow$ Downstream classifier

Learn an upstream representation that

- **generalizes** better across different attacks
- can be **interpreted**

Motivation

Style-Linguistics decomposition

- Style: prosody, accent, ethnicity, gender, age, emotion
- Linguistics: syllables, words, sentences

Examples

- gender difference in languages (Xia 2013)
- pronunciation of non-native speakers (Pullen 2011)
- prosody and language understanding (Cutler, Dahan, and Van Donselaar 1997)
- age effects in conversational speech (Pereira et al. 2019)
- emotional states and word choices (Lindquist, MacCormack, and Shablack 2015)

Reality Defender

SLIM: **S**tyle-**LI**nguistic **M**ismatch Model

Hypothesis

The underlying style-linguistics dependency of real speech in hard to model perfectly by existing TTS and VC models

Proposed modelling strategy:

- Learn dependency via self-supervised learning (SSL) using real speech
- Use the dependency embeddings for downstream supervised training

Reality Defender

SLIM: Model architecture

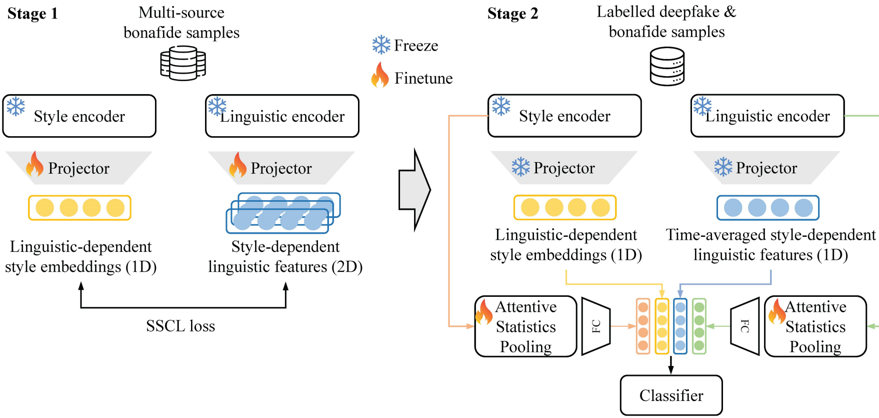


Figure: *Two-stage training framework of SLIM.*

SLIM: Training details

- Subspace representations
  - Style: first 11 layers from Wav2vec-XLSR-53-SER
  - Linguistics: 14-21 layers from Wav2vec-XLSR-53-ASR
  - Note that these were changed to WavLM-base for the ASVspoof5 challenge due to restrictions on backbones

Reality Defender

SLIM: Training details

- Subspace representations
  - Style: first 11 layers from Wav2vec-XLSR-53-SER
  - Linguistics: 14-21 layers from Wav2vec-XLSR-53-ASR
  - Note that these were changed to WavLM-base for the ASVspoof5 challenge due to restrictions on backbones
- Stage-1 objective:

$$\mathcal{L}_{SSC} = \mathcal{L}_D + \lambda \mathcal{L}_R \tag{1}$$

$$\mathcal{L}_D = \frac{1}{T} \sum_{t=0}^{T} \|\mathbf{S}_{f,t} - \mathbf{L}_{f,t}\|_{\mathbf{F}}^2, \tag{2}$$

$$\mathcal{L}_R = \|\mathbf{S}_f \mathbf{S}_f^\intercal - \mathbb{I}\|_{\mathbf{F}}^2 + \|\mathbf{L}_f \mathbf{L}_f^\intercal - \mathbb{I}\|_{\mathbf{F}}^2 \tag{3}$$

Reality Defender

SLIM: Training details

- Subspace representations
  - Style: first 11 layers from Wav2vec-XLSR-53-SER
  - Linguistics: 14-21 layers from Wav2vec-XLSR-53-ASR
  - Note that these were changed to WavLM-base for the ASVspoof5 challenge due to restrictions on backbones
- Stage-1 objective:

$$\mathcal{L}_{SSC} = \mathcal{L}_D + \lambda \mathcal{L}_R \tag{1}$$

$$\mathcal{L}_D = \frac{1}{T} \sum_{t=0}^{T} \|\mathbf{S}_{f,t} - \mathbf{L}_{f,t}\|_{\mathbf{F}}^2, \tag{2}$$

$$\mathcal{L}_R = \|\mathbf{S}_f \mathbf{S}_f^{\intercal} - \mathbb{I}\|_{\mathbf{F}}^2 + \|\mathbf{L}_f \mathbf{L}_f^{\intercal} - \mathbb{I}\|_{\mathbf{F}}^2 \tag{3}$$

- Stage-2: Standard supervised training with binary labels

Reality Defender

SLIM: Training details

- Subspace representations
  - Style: first 11 layers from Wav2vec-XLSR-53-SER
  - Linguistics: 14-21 layers from Wav2vec-XLSR-53-ASR
  - Note that these were changed to WavLM-base for the ASVspoof5 challenge due to restrictions on backbones
- Stage-1 objective:

$$\mathcal{L}_{SSC} = \mathcal{L}_D + \lambda \mathcal{L}_R \tag{1}$$

$$\mathcal{L}_D = \frac{1}{T} \sum_{t=0}^{T} \|\mathbf{S}_{f,t} - \mathbf{L}_{f,t}\|_{\mathbf{F}}^2, \tag{2}$$

$$\mathcal{L}_R = \|\mathbf{S}_f \mathbf{S}_f^{\intercal} - \mathbb{I}\|_{\mathbf{F}}^2 + \|\mathbf{L}_f \mathbf{L}_f^{\intercal} - \mathbb{I}\|_{\mathbf{F}}^2 \tag{3}$$

- Stage-2: Standard supervised training with binary labels
- Trainable components
  - Subspace representations remain frozen for both stages
  - Projectors trained at stage-1 (2M); frozen at stage-2
  - Classification head trained at stage-2 (5M)

Reality Defender

NeurIPS main results

Category	Model	ASVspoof19 EER↓	ASVspoof19 F1↑	ASVspoof21 EER↓	ASVspoof21 F1↑	In-the-wild EER↓	In-the-wild F1↑	MLAAD-EN EER↓	MLAAD-EN F1↑	#Param (million)
	LCNN	3.7	.834	25.5	.197	65.6	.373	37.2	.654	4
	RawNet2	3.0	.875	22.3	.213	37.8	.602	33.9	.676	4
	PS3DT	4.5	—	—	—	29.7	—	—	—	N/A
Frozen frontend	W2V-ASP	3.3	.858	19.6	.233	30.2	.705	29.1	.715	9
	WLM-ASP	**0.3**	.983	9.0	.426	25.4	.751	30.3	.709	9
	HUB-ASP	0.5	.975	15.4	.289	29.9	.718	31.0	.702	9
	W2V-LLGF	2.3	.936	9.4	.402	25.1	.756	27.8	.731	10
	W2V-LCNN	0.6	—	**8.1**	—	24.5	—	—	—	N/A
	W2V+WLM	1.8	.916	22.5	.203	30.3	.704	27.0	.739	9
	W2V+HUB	0.9	.956	14.2	.310	27.9	.737	27.6	.732	9
	WLM+HUB	0.8	.963	16.7	.269	29.2	.724	28.5	.720	9
	SSL-Fusion	**0.3**	.981	8.9	.419	24.2	.765	26.5	.739	10
	SLIM variants	**(ours)**								
	Enc_{sty}	6.7	.740	8.6	.438	29.2	.724	25.4	.756	9
	Enc_{ling}	5.9	.764	9.3	.407	30.4	.708	25.0	.760	9
	$\text{Enc}_{style+ling}$	3.5	.834	9.0	.429	25.1	.757	23.9	.772	10
	Dependency	2.8	.897	20.5	.234	25.8	.750	19.8	.811	9
	Full	0.6	.969	**8.3**	.451	**12.9**	.895	**13.5**	.865	11
Finetuned frontend	W2V-ASP	0.3	.984	4.5	.646	18.6	.836	19.2	.817	317
	W2V-AASIST	**0.2**	.991	**3.6**	.707	17.5	.847	14.5	.856	317
	SLIM (ours)	**0.2**	.989	4.4	.651	**12.5**	.898	**10.7**	.892	253

SLIM: ASVspoof5 Results

	SSL Model	DA	Loss	B_{train}	ASV5 *dev*	ASV2019	ITW	ASV5 *prog*	ASV5 *eval*
Cross-model comparison	WavLM-Base	-	BCE	8	9.6	15.2	20.0	13.4	NA
	WavLM-Base (fft)	-	BCE	8	7.4	18.7	29.6	13.6	NA
	Data2vec-Base	-	BCE	8	9.6	13.7	22.8	NA	NA
	Data2vec-Base (fft)	-	BCE	8	14.6	31.1	37.6	NA	NA
	Wav2vec-Large	-	BCE	8	7.7	15.4	22.1	NA	NA
	Wav2vec-Large (fft)	-	BCE	8	18.0	25.9	35.4	NA	NA
	SLIM (WavLM)	-	BCE	8	5.2	11.1	25.7	7.1	NA
SLIM ablation	SLIM (Wav2vec)	-	BCE	8	7.7	12.9	19.2	NA	NA
	SLIM (WavLM)	RawBoost	BCE	8	**2.9**	9.5	**10.8**	3.6	NA
	SLIM (WavLM)	RawBoost+Noise+RIR	BCE	8	3.3	10.4	12.4	NA	NA
	SLIM (WavLM)	RawBoost	Focal	8	3.8	10.7	14.5	2.7	NA
	SLIM (WavLM)	RawBoost	BCE	4	3.0	**7.4**	**10.8**	**2.4**	**5.5**

Key improvement:

- **-6.3%** EER by applying SLIM
- **-3.5%** EER by adding RawBoost augmented samples
- Without SLIM, RawBoost did not work well; Other top systems used neural vocoders to expand training data

Reality Defender

Quantifying style-linguistics mismatch

- Mismatch is quantified by the distance between style-linguistics dependency feature pairs

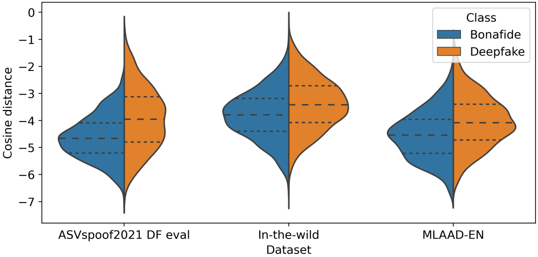
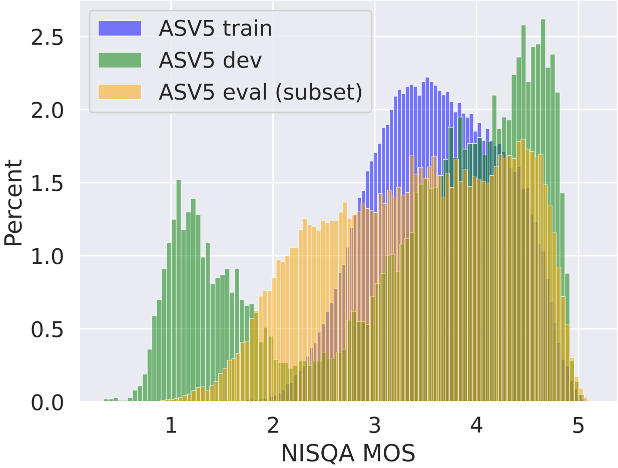


Figure: Cosine distance (log scale) calculated between the style and linguistics dependency features for `ASVspoof2021 DF eval`, `In-the-wild`, and `MLAAD-EN`.

Probing into misclassifications

- Dev and eval data are **less intelligible** (see figure below)
- 10% of eval data has **more than 1 speaker** identified



Choice of layers to represent style and linguistics

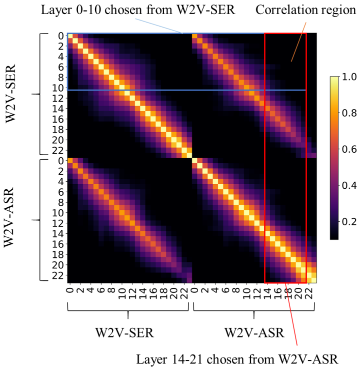


Figure: Spearman correlation coefficients. Blue: layers 0-10 from Wav2vec-SER to represent style information. Red: layers 14-21 from Wav2vec-ASR to represent linguistics information. Figure is from our arxiv paper.

Projector architecture

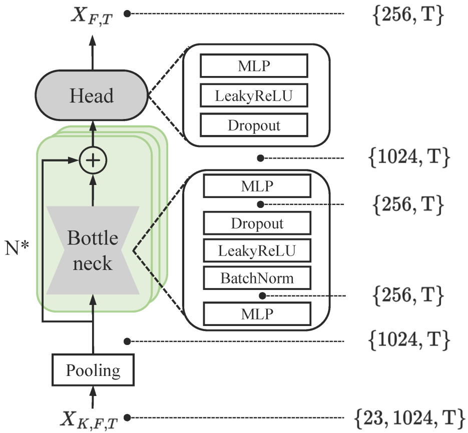


Figure: Architecture of the projector module.