Alethia: a Foundational Encoder for Voice Deepfakes

Yi Zhu[1], Brahmi Dwivedi[2], Jayaram Raghuram, Surya Koppisetti
[1,2] Work conducted at Reality Defender
[1]Correspondence: yzhu00121@gmail.com

Reality Defender

Background

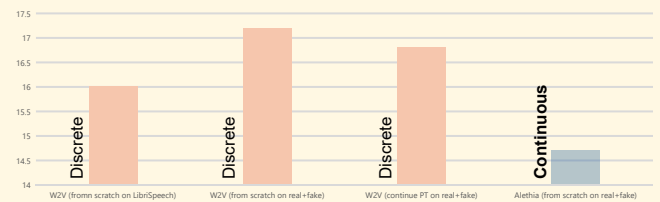
01 **Diminishing returns from post-training** on deepfake detection / localization tasks

02 Can we have a **scalable pre-training recipe** for deepfakes?

03 What are the desired representation **properties** and what **data and pretraining task** should be used?

Type of tokens to predict: **Continuous vs Discrete**

Pre-training with discrete targets led to degraded performance (higher EER), we hypothesize this is due to loss of acoustic details



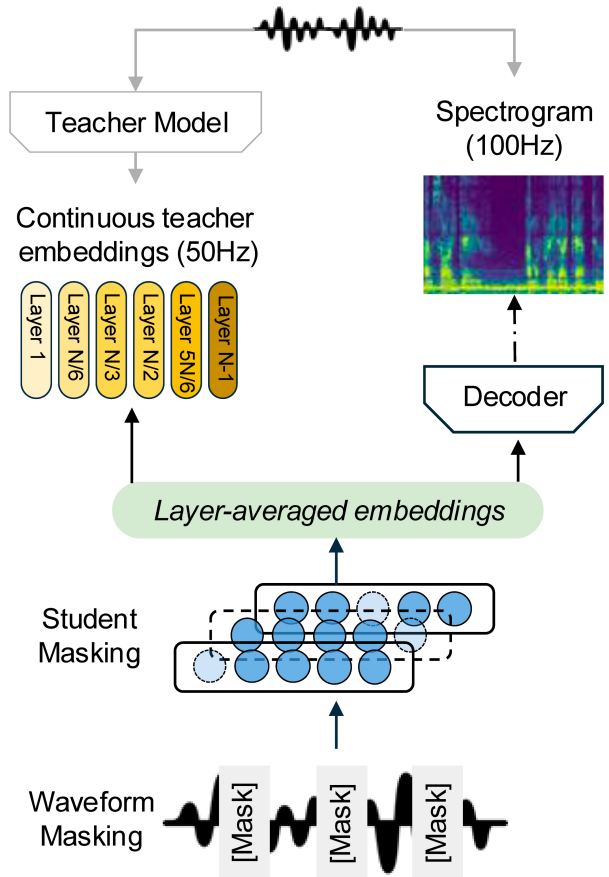
A strong prior for masked latent prediction: **Bottleneck vs Layer-by-layer**

$$\min_{\theta} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}} \left[\mathcal{L}_{MEP}\left(\{f^m_{\text{tea}}(\mathbf{x})\}_{m\in\mathcal{M}}, \Phi(f_{\text{stu}}(\tilde{\mathbf{x}};\theta))\right)\right]$$

weighted cosine+L1 distance

M layers of teacher latent from **unmasked** speech

Φ: Layer pooling -> projection in M layers



Reconstruct **Latent and Spectrogram** Simultaneously

Hypothesis: Acoustic details learned through generation are deepfake "giveaways"

Real and imaginary velocities — Bottleneck latent

$$[\hat{\mathbf{v}}_{real}, \hat{\mathbf{v}}_{imag}] = g_\psi(\mathbf{x}_t, t, \mathbf{z}_{cond})$$

$$\mathcal{L}_{FM} = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{z}}\left[\frac{\mathcal{L}_{real} + \mathcal{L}_{imag}}{\sigma^2_{eps}}\right]$$

Experimental Setup

Baselines: Wav2vec2.0 (0.3b, 1b), WavLM, HuBERT

Pretraining data: 30k hours of real+fake speech (19k after quality control)

Benchmark: 56 datasets, 5 tasks, 10+ languages, 100+ generative models, 10+ perturbations