

The Real Outlier: Improving Cross-Lingual Synthetic Speech Detection

Stefano Dell'Anna, Kratika Bhagtani, Amit K. S. Yadav, Giulia Boato, Edward J. Delp



Abstract: The rapid proliferation of highly realistic speech synthesizers poses a growing threat to information integrity and public trust. Existing synthetic speech detectors are mainly trained for English or Chinese, with large performance penalties when applied to other languages. Our contributions are twofold: First, to foster the development of multilingual detectors, we introduce **ITASpoof**, a **large-scale dataset of over one million real and synthetic Italian speech samples**. Synthetic speech is generated using seven recent zero-shot Text-to-Speech models, while real data is collected from 15 different cities, ensuring gender balance while minimizing speaker characteristic bias. Second, we introduce a novel **synthetic speech detector: ReSOD**, adopting an **Omnilingual-ASR backbone**, and trained with a **real-speech Out-of-distribution paradigm**. Our evaluation demonstrates improved detection abilities over existing detectors, achieving SoTA performances over three multilingual benchmarks.

Problem formulation

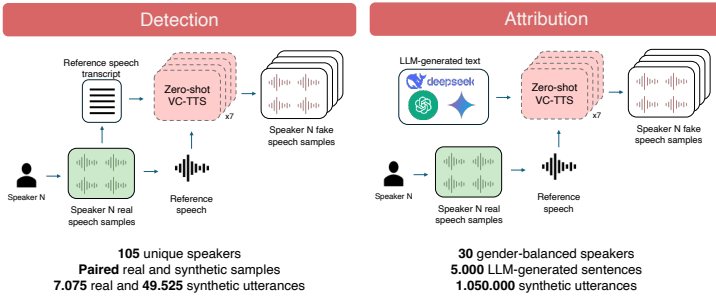
- Existing synthetic speech datasets focus primarily for English and Chinese, limiting the cross-language abilities of existing detectors
- A significant portion of the global population is left without reliable forensic tools against voice synthesis misuse

Language-specific datasets are necessary to evaluate language-agnostic detectors

ITASpoof dataset

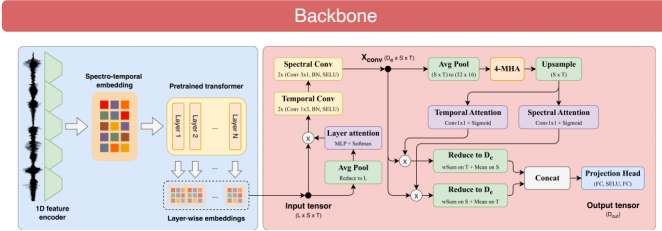
1M+ dataset of paired real and synthetic Italian speech

- 7,075 transcribed utterances
- 9.91 s average duration
- 105 unique speakers
- 7 SoA VC-TTS models: Elevenlabs multilingual v2 — F5-TTS — Fish-Speech XTTS-v1 — XTTS-v2 — Your-TTS — Chatterbox

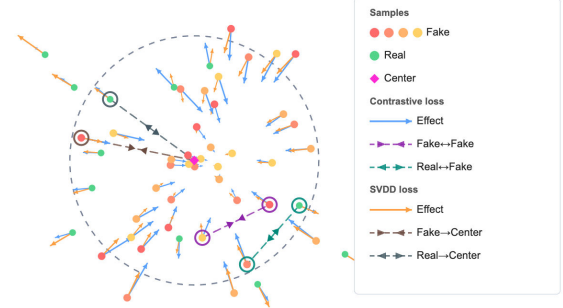


Proposed detector

- Custom Spectro-Temporal adaptation network over pretrained Wav2Vec-style encoder.
- Real-Speech as Out-of-Distribution (ReSOD) strategy for better cross-language generalization.



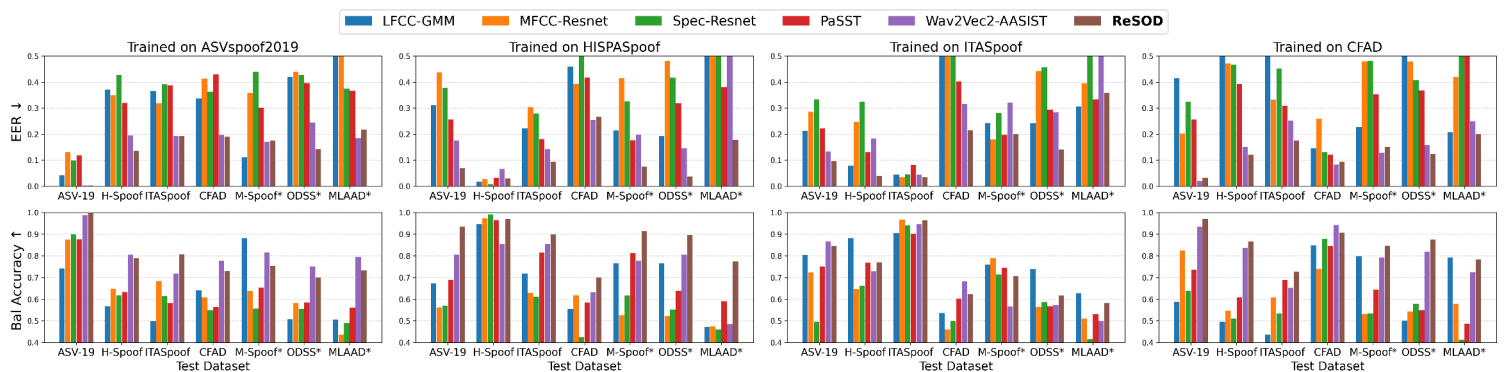
Training strategy



Dual training objective:

- Contrastive loss with real sample mask, encouraging compact representation of synthetic samples and a sparse representation of real ones.
- SVDD loss, enforcing radial separation between synthetic and real samples, further encouraging a compact inlier region and a sparse outlier one.

Results



Average EER in cross-language scenario (Trained on column-specified dataset, tested on all others)

EER [%] ↓	ASVspooft2019 [4]	HISPASpoof [5]	ITASpoof	CFAD [6]
LFCC-GMM [4]	41.78	38.41	26.62	42.57
MFCC-Resnet [1]	43.82	42.97	34.93	39.72
Spec-Resnet [1]	40.42	42.98	40.51	45.55
PaSST [2]	36.70	28.83	26.34	36.52
W2V-AASIST [3]	19.74	24.06	29.43	15.97
ReSOD	17.57	11.96	17.49	13.30

Takeaways:

- Cross-language generalization remains a major bottleneck for synthetic speech detectors, which rely on costly fine-tuning and struggle when linguistic variability is disentangled from authenticity cues.
- ITASpoof, a large-scale Italian dataset for synthetic speech detection and attribution, improving inclusivity and allowing rigorous analysis of detector robustness under linguistic shifts.
- ReSOD, trained with a real-as-OOD strategy, promotes sparse latent representations, significantly improving cross-dataset and cross-language generalization while preserving strong in-domain detection performance.

[1] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," in Interspeech 2019, 2019, pp. 1078–1082.
 [2] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in Interspeech 2022, 2022, pp. 2753–2757.
 [3] H. Tak, M. Todisco, X. Wang, J. Weon Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in The Speaker and Language Recognition Workshop (Odyssey 2022), 2022, pp. 112–119.
 [4] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," in Interspeech 2019, 2019, pp. 1008–1012.
 [5] M. Risques, K. Bhagtani, A. K. S. Yadav, and E. J. Delp, "Hispoof: A new dataset for Spanish speech forensics," in 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2026.
 [6] H. Ma, J. Yi, C. Wang, X. Yan, J. Tao, T. Wang, S. Wang, and R. Fu, "Clad: A chinese dataset for fake audio detection," Speech Commun., vol. 164, no. C, Oct. 2024.