

VoxID – Voice Identification and Attribution

Lead CI: Prof. Sanjay Jha (UNSW), Prof. Justin Lipman (UTS)

Co-CI: A/Prof. Yang Song (UNSW), A/Prof. Michael Richardson (UNSW),
A/Prof. Mukesh Prasad (UTS), Wenbin Wang (UNSW), Xing Zi (UTS)

Introduction

AI voice cloning and text-to-speech (TTS) systems can now generate speech that sounds convincingly human using only a few seconds of audio. This has unlocked real benefits in accessibility, media production, education, and automation. But it also introduces serious risks. Voices from public recordings are being used to train AI models without consent, and cloned voices are already being used in scams, impersonation, and misinformation.



Unconsented Voice Harvesting

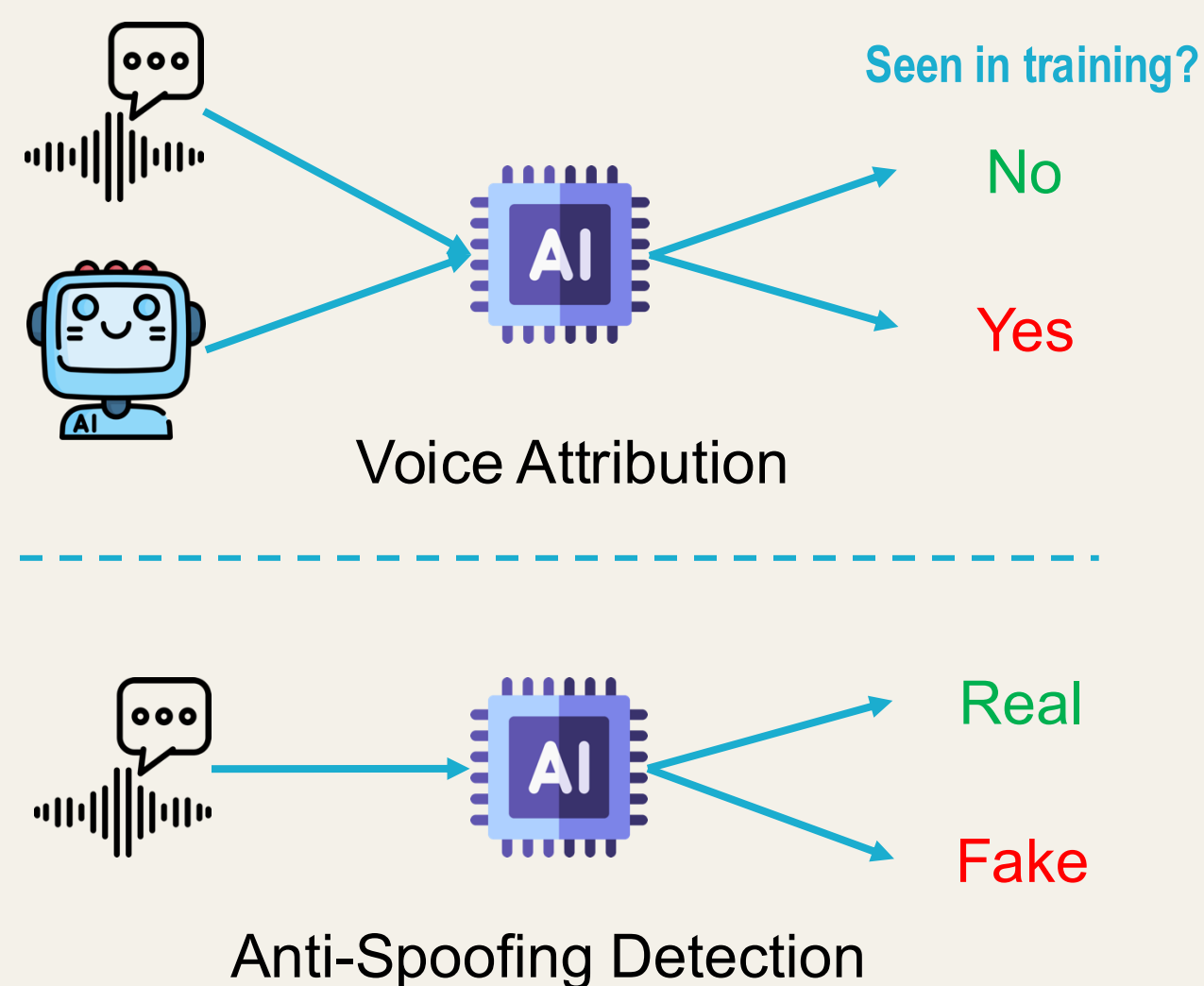


Malicious Voice Cloning

Today, we have almost no way to check if someone's voice was taken, or to prove misuse. This focuses on voice identification and attribution, not just "Is this audio fake?", but "Whose voice was cloned, and can we prove it?"

Our Objectives

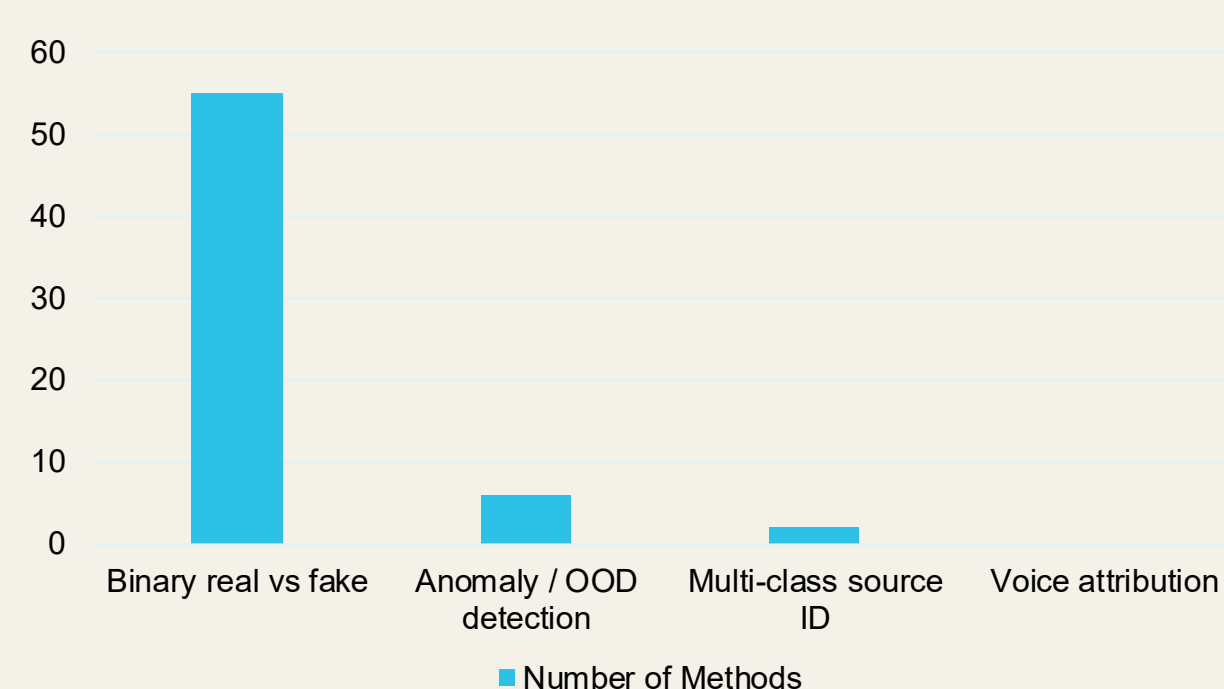
Our goal is twofold:



Literature Review

We reviewed 200+ publications since 2020 and built a matrix of 67 representative methods. We found:

- Nearly all existing work frames deepfake audio as a binary "real vs fake" problem. None of these systems can answer: "Whose voice was cloned?" — there is no standard way to prove voice misuse.



- Public benchmark datasets test only this question: "Can you tell real vs fake audio?"

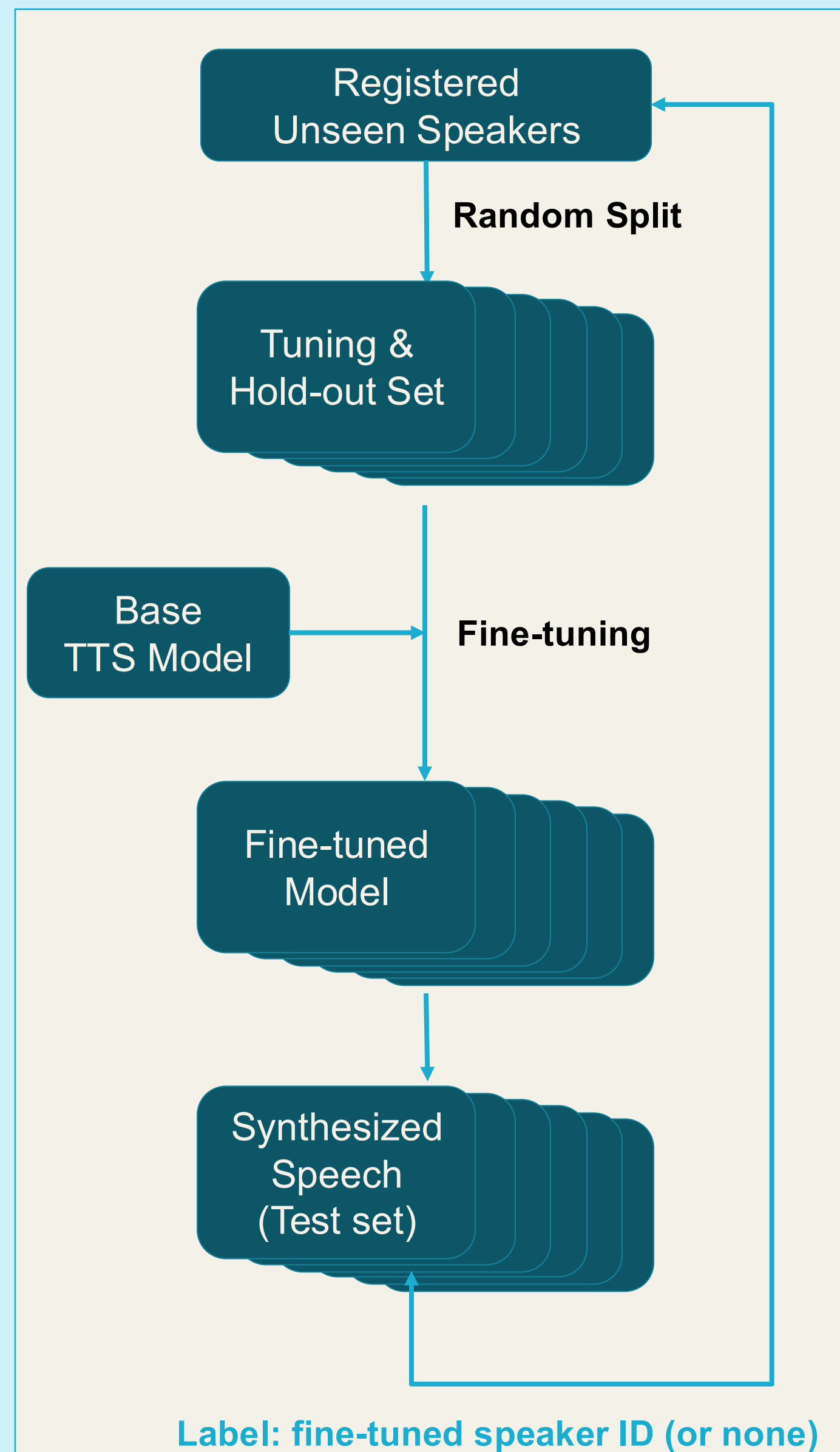
Dataset	Year	Language	Task Focus
ASVspoof 2015	2015	EN	Real vs Fake detection
FoR-original	2019	EN	Real vs Fake detection
ASVspoof 2019-LA	2019	EN	Real vs Fake detection
ASVspoof 2021-LA	2021	EN	Real vs Fake detection
FMFCC-A	2021	CN	Real vs Fake detection
WaveFake	2022	EN, JA	Real vs Fake detection
TIMIT-TTS	2023	EN	Real vs Fake detection
DeepVoice	2024	EN	Real vs Fake detection
MLAAD	2024	Multi	Real vs Fake detection
CodecFake	2024	EN, CN	Real vs Fake detection
CFAD	2024	CN	Real vs Fake detection

Towards Voice Attribution Data

To evaluate Voice Attribution, we need more than "real vs fake." We need an evidence chain:

- Identify the source speaker
- Record whether the model was fine-tuned on that speaker
- Test whether that identity survives in the synthetic speech

Thus, we will construct a controlled evaluation dataset as follows:



Next Steps

- Add real-world channel stress (phone compression, noise, replay)
- Build prototype models
- Benchmark attribution and anti-spoofing under realistic conditions
- Standardize the scoring pipeline (accuracy, false accusation, robustness)
- Release a reusable evaluation harness

Social Impact

- Protects vocal identity as personal data
- Reduces impersonation scams and urgent-payment fraud
- Supports victims instead of only punishing attackers
- Creates an auditable consent trail
- Enables safer AI deployment at policy level
- Builds public trust in synthetic speech