# SVeritas: Benchmark for Robust Speaker Verification under Diverse Conditions

Anonymous ACL submission

## Abstract

Speaker verification (SV) models are increasingly integrated into security, personalization, and access control systems, yet their robustness to many real-world challenges remains inadequately benchmarked. These include a variety of natural and maliciously created conditions causing signal degradations or mismatches between enrollment and test data, impacting performance. Existing benchmarks evaluate only subsets of these conditions, missing others entirely. We introduce SVeritas, a comprehensive Speaker Verification tasks benchmark suite, assessing SV systems under stressors like recording duration, spontaneity, content, noise, microphone distance, reverberation, channel mismatches, audio bandwidth, codecs, speaker age, and susceptibility to spoofing and adversarial attacks. While several benchmarks do exist that each cover some of these issues, SVeritas is the first comprehensive evaluation that not only includes all of these, but also several other entirely new, but nonetheless important real-life conditions that have not previously been benchmarked. We use SVeritas to evaluate several state-of-the-art SV models and observe that while some architectures maintain stability under common distortions, they suffer substantial performance degradation in scenarios involving cross-language trials, age mismatches, and codec-induced compression. Extending our analysis across demographic subgroups, we further identify disparities in robustness across age groups, gender, and linguistic backgrounds. By standardizing evaluation under realistic and synthetic stress conditions, SVeritas enables precise diagnosis of model weaknesses and establishes a foundation for advancing equitable and reliable speaker verification systems.

## 1 Introduction

Speaker verification technology has achieved remarkable accuracy under controlled conditions, driven by advances in deep neural embeddings, margin-based losses, and self-supervised pretraining. However, real-world deployments – from secure access control and telephony authentication, to personalized assistants, and law enforcement – confront a broad spectrum of challenges that degrade performance, including degradations to the signal itself, and mismatches between the test utterances and the enrollment recordings they are compared to. These mismatches arise from natural variability (e.g., spontaneous versus read speech, cross-language trials, or temporal drift), environmental distortions (e.g., reverberation, background noise, far- versus near-field capture), device and codec artifacts, demographic factors (age, health or physical condition), and even malicious manipulations such as spoofing or adversarial attacks. Without comprehensive, standardized evaluation across these diverse stressors, it remains unclear which aspects of SV systems are robust in practice and where critical vulnerabilities lie.

Existing benchmarks each target a narrow subset of these challenges. For example, CommonBench (Hintz and Siegert) offers large-scale multilingual text-independent trials, yet it relies on an ECAPA-based outlier filter that may prune precisely the hardest cases (e.g., heavy accents or noisy recordings), and omits deliberate distortions such as codec compression or spoofed audio. IndicSUPERB (Javed et al., 2023) highlights performance on twelve Indian languages but focuses exclusively on scripted, read speech in clean or synthetic noise conditions, neglecting cross-language scenarios, far-field capture, or adversarial manipulations. Other benchmarks examine specific dimensions in isolation – far-field effects in MultiSV (Mošner et al., 2022), age variation in time-varying SV (Doddington, 2012), or spoofing attacks in ASVspoof (Wu et al., 2017). To the best of our knowledge, no prior suite spans the full gamut of natural, environmental, demographic, codec, and adversarial factors under a unified framework. Fur-

thermore, many datasets lack sufficient metadata to analyze fairness across age, gender, or linguistic subgroups, rely on relatively large enrollment durations, or assume static, text-independent protocols that do not reflect modern low-resource, multi-file, or cross-domain requirements.

To address these gaps, we introduce SVeritas, the first comprehensive speaker verification benchmark suite that systematically evaluates state-of-the-art models across an extensive set of real-world and synthetic stressors. SVeritas assembles trials spanning (i) content and style variations (read vs. spontaneous, same vs. different sentences, multi-language), (ii) acoustic and channel mismatches (noise types and levels, far- vs. near-field, codec and bandwidth variations), (iii) demographic and physical factors (age-group mismatches, health or emotional states), (iv) enrollment/test duration and multi-file enrollment, and (v) security threats (spoofing via TTS/VC pipelines, universal and adversarial perturbations). By unifying these dimensions, SVeritas not only measures aggregate metrics such as equal error rate and detection cost function, but also facilitates fine-grained analyses of performance disparities across demographic subgroups and operating conditions. Through extensive evaluation of leading architectures, we uncover systemic weaknesses – particularly in cross-language, age-mismatch, and codec-compressed trials – and expose fairness gaps that vary nontrivially by gender and language background.

In summary, SVeritas establishes a rigorous, reproducible foundation for diagnosing robustness and equity in speaker verification. By revealing hitherto uncharacterized vulnerabilities and enabling targeted stress-testing, our benchmark paves the way for developing more reliable, inclusive, and secure SV systems suitable for deployment in the complex acoustic and demographic landscapes of real-world applications. Our code is publicly available with documentation, fostering straightforward reproducibility and extensibility. https://anonymous.4open.science/r/SVeritas-3217/.

## 2 Background and Related Work

### 2.1 Speaker Verification Systems and their Vulnerabilities

Speaker Verification systems attempt to verify the identity of a speaker by comparing (embeddings) derived from) their voice recordings to a template such as a statistical model (Reynolds et al., 2000), or other embeddings derived from "enrollment" data (Dehak et al., 2010).

Besides the models themselves, the performance of the SV system depends on many other factors, primary being the native quality of the signal itself. The best performances are generally obtained with studio-quality broadband signals (Villalba et al., 2020), which degrades when the bandwidth of the signal is restricted, such as over telephony or cellphone channels (Kenny, 2010). The application of various codecs also degrades performance (Njegovec, 2025). External influences such as background noise, recording room responses and reverberation also degrade performance (Ko et al., 2017) (Nandwana et al., 2018). Perhaps most concerningly, innate biases within the system too result in reduced performances for some categories of subjects (Hajavi and Etemad, 2023). Performance is also dependent on the duration of the recording (longer recordings are better (Poddar et al., 2018)), and on whether the speech is spontaneous or recited, e.g. by reading (Nakamura et al., 2008).

A second, and equally important source of degradation is *mismatches* between the conditions of the test and enrollment data. Signal differences in bandwidth, channel condition, duration *etc.* can result in degraded performance. *Content* variations, such as language and dialectal differences (Abdullah et al., 2025), as well as exactly what is spoken (Dey, 2018) can cause degradations. *Biological* influences, such as changes in the age or health status of the speaker too can cause degradations (Kelly and Harte, 2011).

A third and increasingly important source of degradation is *active* misdirection, such as through voice mimicry (Hautamäki et al., 2013), synthetic voice recordings (Zuo et al., 2024) or adversarial modifications (Alzantot et al., 2018)(Zhou et al., 2023)(Jati et al., 2021) which can make an SV system fraudulently accept an imposter or reject a genuine match.

### 2.2 Remediations

The most common approach to remediation of natural and mismatch-based degradations is through inclusive training – adding data with the variations that one must be robust to in the training data of the model (Ko et al., 2017), *e.g.* far-field and noisy conditions (often through simulated room responses and digital addition of noise) (Ko et al., 2017; Yakovlev et al., 2024; Thienpondt and De-

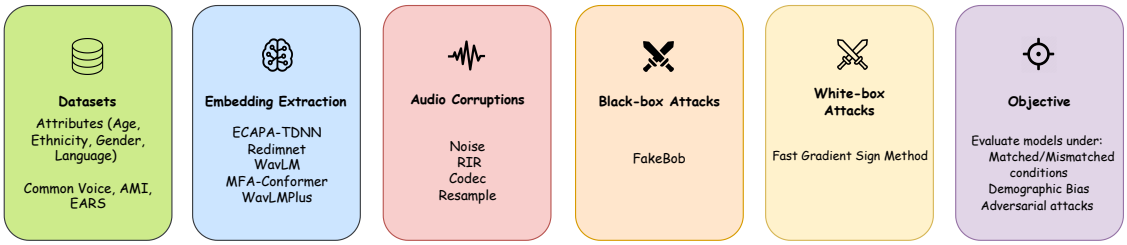


Figure 1: Overview of our benchmark SVERITAS.

muynck, 2023; Al-Karawi, 2021), codec and bandwidth variations (Polacky et al., 2016), *etc.*. Explicit modeling of, and compensation for effects such as noise and reverberation has also been found to be effective in some settings (Al-Karawi, 2021).

Another popular approach to mitigate the effect of variations is through contrastive losses that attempt to neutralize variations by minimizing the distances of (embeddings from) recordings and their mismatched counterparts (Inoue and Goto, 2020). Alternate methods attempt to disentangle confounding sources of variation (Nam et al., 2024).

Defenses against misdirection attacks include explicit attempts at detecting mimicry (Hautamäki et al., 2013), or through adversarial defenses such as adversarial training, which can protect to some extent against adversarial attacks (Zhou et al., 2023).

### 2.3 Benchmarking

Speaker verification systems are often used in critical settings, such as user authentication or law enforcement. Consequently, benchmarking their performance under these various challenges becomes necessary.

Indeed, benchmarking has been central to the development of SV systems, guiding progress through standardized evaluation protocols. Traditional efforts such as the NIST Speaker Recognition Evaluations (SREs) (Sadjadi et al., 2022, 2017) have driven advances in SV for over two decades, though their design primarily targets constrained settings involving telephone and microphone speech. The VoxCeleb Speaker Recognition Challenge (VoxSRC) (Nagrani et al., 2020) was introduced to evaluate the ability of modern speaker recognition systems to identify speakers from speech captured 'in the wild', To address isolated robustness factors, several specialized benchmarks have been introduced. The Short-Duration Speaker Verification Challenge (SDSVC) (Zeinali et al., 2019) and Far-Field SVC (Qin et al., 2020) focus on duration and spatial variability. SUPERB (Yang et al., 2021) offers a comprehensive suite for evaluating speech representation learning across multiple tasks, including speaker verification, but the SV component remains relatively coarse-grained and lacks detailed stress testing. More recently, VoxBlink (Lin et al., 2024) emphasized robustness to device mismatch and short-duration utterances, uncovering substantial performance degradation under realistic deployment scenarios. Other efforts at benchmarking have been mentioned in Section 1.

Notably, while each of these benchmarks evaluates the system under subsets of the various challenges a real-life deployment may face, unlike other speech pattern recognition tasks, such as speech recognition (Shah et al., 2025), there is as yet no single benchmark suite that integrates all of the broader robustness dimensions such as recording condition variations, demographic variation, adversarial perturbations, and codec-induced compression into a unified diagnostic framework. Some factors related to demographics and content are not addressed by any existing benchmark. SVeritas addresses this gap.

## 3 SVeritas Benchmark

SVeritas aims to provide a thorough benchmarking of SV systems, evaluating its performance under various degradations, mismatches, sources of bias, and attacks, providing both detailed and summarized evaluations, along with statistical significance reports where appropriate. The tests are not only intended to evaluate the performance of the system under various conditions and threats that may be expected in real-life deployments, but to also provide a diagnostic tool to identify weaknesses, and detect any systematic biases or vulnerabilities.

SVeritas evaluates the robustness of SV models

| Condition | WavLM-Base | WavLM-Base+ | RedimNet | ECAPA-TDNN | MFA-Conformer |
|---|---|---|---|---|---|
| Real vs. Synthetic | 25.74% | 23.76% | 5.94% | 5.94% | 0.00% |
| FGSM | 48.38% | 48.39% | 37.63% | 52.63% | 45.26% |
| FakeBob | 25.81% | 18.28% | 10.75% | 62.36% | 35.48% |

Table 1: EARS: EER for SV models under clean, spoofing, and adversarial attack conditions.

| Category | Subgroup | WavLM-Base | WavLM-Base+ | RedimNet | ECAPA-TDNN | MFA-Conformer |
|---|---|---|---|---|---|---|
| Gender | Female (59 spks) | 13.74% | 10.86% | 1.50% | 3.72% | 2.44% |
| | Male (43 spks) | 17.26% | 12.77% | 1.79% | 4.41% | 2.76% |
| Age | F (18–25), 13 spks | 13.01% | 10.97% | 2.34% | 7.00% | 4.62% |
| | F (26–35), 13 spks | 15.26% | 12.71% | 1.80% | 4.24% | 3.11% |
| | F (36–45), 7 spks | 10.91% | 8.30% | 0.27% | 1.41% | 1.07% |
| | F (46–55), 14 spks | 14.25% | 11.99% | 1.47% | 3.31% | 2.49% |
| | F (56–65), 10 spks | 16.84% | 15.04% | 1.52% | 3.07% | 1.83% |
| | F (66–75), 2 spks | 26.28% | 18.63% | 0.73% | 3.67% | 1.61% |
| | M (18–25), 14 spks | 23.35% | 16.85% | 3.61% | 7.81% | 4.99% |
| | M (26–35), 10 spks | 16.16% | 13.75% | 2.02% | 3.72% | 2.78% |
| | M (36–45), 10 spks | 14.22% | 10.79% | 1.78% | 3.43% | 1.88% |
| | M (46–55), 4 spks | 23.40% | 18.07% | 2.50% | 7.89% | 4.04% |
| | M (56–65), 5 spks | 26.21% | 19.52% | 2.16% | 6.43% | 4.46% |
| Ethnicity | F, White (40 spks) | 14.67% | 11.99% | 1.44% | 3.90% | 2.44% |
| | F, Hispanic (4 spks) | 8.12% | 5.88% | 0.43% | 2.02% | 1.24% |
| | F, Black (13 spks) | 15.70% | 13.63% | 2.34% | 5.18% | 3.68% |
| | F, Asian (2 spks) | 6.51% | 2.24% | 0.95% | 6.59% | 2.09% |
| | M, White (31 spks) | 19.18% | 14.30% | 1.94% | 4.87% | 2.92% |
| | M, Hispanic (5 spks) | 20.74% | 15.86% | 1.30% | 5.97% | 3.28% |
| | M, Black (5 spks) | 16.23% | 15.26% | 1.47% | 3.10% | 1.50% |
| | M, Asian (2 spks) | 17.58% | 9.15% | 0.06% | 0.30% | 0.06% |

Table 2: EARS: EER for SV models across gender, age, and ethnicity subgroups.

through a structured three-stage pipeline: (1) scenario simulation, (2) embedding extraction, and (3) performance evaluation. As illustrated in Figure 1, the first stage introduces a wide range of real-world and synthetic perturbations to both enrollment and test audio. These include natural variations (e.g., speaking style, duration, and linguistic content), environmental conditions (e.g., noise, reverberation, and microphone distance), and recording artifacts (e.g., codec compression and bandwidth limitations). SVeritas also incorporates physical and demographic variability, such as speaker age, health, and accent, as well as adversarial factors including spoofing attempts and both black-box and white-box attacks. The second stage applies multiple state-of-the-art embedding models to extract speaker representations. Finally, performance is evaluated using metrics such as Equal Error Rate (EER) across matched/mismatched scenarios and demographic subgroups, enabling a comprehensive assessment of model robustness and fairness.

### 3.1 Scenario Simulation

SVeritas evaluates speaker verification systems across a range of real-world and synthetic scenarios. These simulations are organized into six broad categories, each capturing a unique aspect of deployment variability or robustness challenge.

The data themselves were obtained by simulating the various effects on a number of public corpora such as EARS (Richter et al., 2024), AMI Meeting Corpus (Kraaij et al., 2005) and Mozilla CommonVoice 21 (Ardila et al., 2020). In order to maximally ensure fair implementation of the benchmark we only employed the *test* portions of the corpora, under the assumption that developers of SV systems are unlikely to have used these to train the model.

#### 3.1.1 Audio Capture

The *audio capture* benchmark evaluates the performance of SV systems under various audio capture conditions that may be encountered in real life.

1. *Broadband clean:* These are 16-bit resolution linear PCM 16khz sampled studio-quality data. In the context of speech processing tasks, this

| Age | Gender | WavLM-Base | WavLM-Base+ | RedimNet | ECAPA | MFA-Conformer |
|---|---|---|---|---|---|---|
| Teens | F 112 spks | 31.15% | 31.61% | 13.82% | 14.08% | 15.01% |
| | M 112 spks | 23.12% | 19.98% | 4.53% | 5.02% | 6.89% |
| Twenties | F 582 spks | 23.27% | 19.34% | 3.67% | 5.41% | 5.93% |
| | M 582 spks | 22.95% | 20.87% | 6.62% | 8.16% | 8.27% |
| Thirties | F 240 spks | 22.70% | 22.34% | 2.94% | 4.67% | 5.52% |
| | M 240 spks | 20.14% | 17.05% | 2.69% | 3.80% | 3.99% |
| Fourties | F 140 spks | 20.75% | 19.30% | 2.47% | 4.08% | 4.84% |
| | M 140 spks | 17.70% | 17.33% | 1.79% | 3.09% | 3.00% |
| Fifties | F 110 spks | 24.49% | 23.09% | 2.88% | 5.54% | 6.43% |
| | M 126 spks | 18.99% | 16.55% | 1.39% | 2.68% | 3.78% |
| Sixties | F 49 spks | 27.95% | 26.11% | 5.60% | 11.11% | 10.63% |
| | M 57 spks | 20.09% | 18.27% | 1.66% | 4.15% | 3.71% |
| Seventy+ | F 17 spks | 19.61% | 16.37% | 1.64% | 6.56% | 1.86% |
| | M 69 spks | 21.09% | 25.23% | 8.89% | 11.96% | 9.95% |

Table 3: CommonVoice: EER variation over age for both genders.

has been the long-standing standard for "ideal" recordings.

2. *G-711:* The G-711 standard data are captured with 8-bit mu law quantization, sampled at 8khz. These remain common in telephony applications. The G-711 achieves a low bitrate of 64kbps (CCITT, 1988).

3. *GSM 06.10:* The GSM 06.10 is a legacy codec, standardized for 2G GSM mobile communications, operates on 8 kHz sampled signals and uses Regular Pulse Excitation with Long Term Prediction (RPE-LTP) to compress speech to approximately 13 kbps, introducing characteristic bandlimited and quantization distortions (TC-SMG, 1993).

4. *Opus:* Opus is a dynamic-bitrate codec employed in applications such as WhatsApp (Kumar et al., 2024), Zoom (Zoom Video Communications, Inc., n.d.) and WebRTC (Valin et al., 2012) . It dynamically adjusts the compression of the signal according to current network conditions. SVeritas uses Opus in two modes, narrowband 8khz and wideband 16khz and randomly selects one of the two to apply to any signal, to simulate the unpredictable nature of the compression.

5. *AMR:* The "Adaptive MultiRate" (AMR) codec is a legacy codec prevalent particularly in 2G and 3G cellular networks. It operates on 8khz mu-law sampled data, and employs variants of CELP coding, but the dynamic switching enables higher-quality audio. SVeritas chooses randomly between AMR-Narrowband (4.75–12.2 kbps)) and AMR-wideband (12.6 kbps or higher) (Sjoberg et al., 2007) to emulate the dynamic nature of the codec.

We evaluate systems both under conditions of *match*, where the same codec is used for both test and enrollment data, and *mismatch*, where the two are different. Note that while the results reported in this paper only consider the codecs mentioned above, our actual package implements and tests against a wider set of popular codecs.

#### 3.1.2 Noise and Channel

Real-world recordings are often affected by the room responses of the space they are recorded in and any noise sources present in them. These introduce distortions which may be further exacerbated by coding schemes that are part of the data capture and transmission. The *noise and channel* benchmark evaluates the robustness of the SV system under these conditions.

1. *Noise:* We evaluate the performance under three varities of noise, namely gaussian noise, environmental noise and crosstalk, at three different signal to noise ratios of 5, 15 and 25dB SNR.

2. *Real Room Response:* We also evaluate the influence of the room response. To implement these, we consider room impulse responses (RIRs) of three different severity levels (in terms of T60 times) drawn randomly from the Room Impulse Response and Noise corpus (Ko et al., 2017).

The actual benchmarks considers both the room responses and the noises in isolation, and their

compounded effect (with RIRs applied on top of the noise). All data are generated through digital simulation of these effects on CommonVoice data. Finally, since codecs too will cause additional distortion of noisy speech, we also consider the effect of codec compression on signals corrupted by noise and room response. For the results reported in this we have only considered Opus, G-729 and AMR codecs applied to signals corrupted by medium severity levels of room response and noise; our actual package reports results on the comprehensive set of combinations and their summary statistics.

#### 3.1.3 Demographic Variations

To assess fairness and generalization, we evaluate speaker verification models across demographic groups, including variations in gender, age, ethnicity, and native language. A key focus of this category is cross-lingual robustness: we use CommonVoice (Ardila et al., 2020) to test whether models trained primarily on English can correctly verify speakers when they speak in other languages. Since speaker identity is grounded in vocal acoustics, a robust SV model should recognize the same speaker regardless of the spoken language. This evaluation reveals whether models rely too heavily on language-specific cues and whether they generalize across linguistic boundaries. We also include TTS-generated speech conditioned on demographic traits to further probe model behavior under controlled variation. This setup allows us to measure demographic robustness and detect possible bias in model predictions.

#### 3.1.4 Synthetic and Adversarial

Real-life deployments are also vulnerable to a variety of attacks. The *synthetic and adversarial* benchmark quantifies this vulnerability. We consider the following attacks.

1. *Synthetic speech:* Here we evaluate the vulnerability of the system to synthetic speech. In all test pairs, both recordings are from the same speaker. In one case both recordings are real, whereas in the other one of the two is synthetic. Ideally the system must accept the former and reject the latter. In this paper we employ CosyVoiceTTS (Du et al., 2024) for the synthetic speech; the full benchmark also evaluates other TTS systems.

2. *Adversarial attack:* We consider adversarial attacks where an imposter attempts to mislead the system. The test is similar to the synthetic speech attack, except that instead of synthetic speech, we have adversarially modified speech. In this paper we consider two adversarial attacks: a *white-box* (full access to model weights) attack, namely the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014), and one *black-box* attack (access restricted to output label only), namely the Fakebob attack (Chen et al., 2021). The full SVeritas benchmark package also considers other popular attacks.

### 3.2 Metrics

We evaluate the performance of speaker verification systems using three standard metrics: Equal Error Rate (EER), minimum Detection Cost Function (minDCF), and Area Under the Curve (AUC). EER is defined as the point at which the false acceptance rate (FAR) equals the false rejection rate (FRR), providing a balanced indicator of accuracy across operating points. It is used as the primary metric due to its intuitive interpretability. minDCF measures the minimum cost achievable when accounting for application-specific penalties (e.g., a higher cost for FAR in high-security contexts) and thus reflects performance under asymmetric decision costs. AUC, a threshold-independent metric, quantifies the separability between genuine and impostor trials and is particularly sensitive to systemic errors in low-FAR regimes, such as those required in forensic applications.

## 4 Evaluation

We evaluate several state-of-the-art SV models using SVeritas and analyze their robustness across a broad range of challenging scenarios. We further extend this analysis to examine model behavior across various demographic subgroups, including speaker age, language background, ethnicity, and gender. Prior work (Hutiri and Ding, 2022) has noted the presence of biases in SV systems, and our findings corroborate and expand upon these observations by revealing that disparities in robustness can emerge across subgroups. These results highlight the importance of standardized evaluation under real-world conditions and underscore the utility of SVeritas in advancing fair and reliable speaker verification.

To further quantify fairness and robustness, we conduct a series of pairwise statistical tests across demographic groups using EER as the primary metric. While each group yields a single EER

| Codec | Condition | WavLM-Base | WavLM-Base+ | RedimNet | ECAPA | MFA-Conformer |
|---|---|---|---|---|---|---|
| | Clean | 23.05% | 20.23% | 4.69% | 6.13% | 6.65% |
| GSM | No Noise | 23.09% | 20.24% | 4.65% | 6.15% | 6.72% |
| | GaussNoise+RIR | 40.64% | 36.67% | 22.46% | 17.89% | 27.30% |
| | EnvNoise+RIR | 40.32% | 38.63% | 17.24% | 16.47% | 20.12% |
| | CrossTalk+RIR | 40.47% | 38.29% | 24.63% | 24.68% | 24.95% |
| AMR | No Noise | 23.09% | 20.24% | 4.65% | 6.15% | 6.72% |
| | GaussNoise+RIR | 40.64% | 36.67% | 22.46% | 17.89% | 27.30% |
| | EnvNoise+RIR | 40.32% | 38.63% | 17.24% | 16.47% | 20.12% |
| | CrossTalk+RIR | 40.47% | 38.29% | 24.63% | 24.68% | 24.95% |
| Opus | No Noise | 23.09% | 20.24% | 4.65% | 6.15% | 6.72% |
| | GaussNoise+RIR | 40.64% | 36.67% | 22.46% | 17.89% | 27.30% |
| | EnvNoise+RIR | 40.32% | 38.63% | 17.24% | 16.47% | 20.12% |
| | CrossTalk+RIR | 40.47% | 38.29% | 24.63% | 24.68% | 24.95% |
| AMI | NearField (F) | 39.64% | 34.60% | 12.12% | 21.69% | 20.63% |
| | NearField (M) | 37.92% | 38.69% | 17.27% | 20.27% | 22.31% |
| | FarField (F) | 47.06% | 47.63% | 34.96% | 36.04% | 36.67% |
| | FarField (M) | 46.63% | 45.39% | 34.65% | 35.00% | 37.65% |

Table 4: CommonVoice: EERs under audio degradation from codecs and noise conditions. AMI results reflect real-world variability in near-field and far-field social environments.

value per model, we leverage the diversity of five models to enable paired comparisons between groups. This design allows us to assess whether performance disparities are consistent across architectures. Full statistical test tables, including $t$-statistics, $p$-values, and significance levels, are provided in Appendix A.

### 4.1 Models

We evaluate a range of SOTA SV models which are publicly available, including WavLM-Base and WavLM-Base+ (Chen et al., 2022), ECAPA-TDNN (Desplanques et al., 2020), and RedimNet (Yakovlev et al., 2024). In addition, we include MFA-Conformer(Zhang et al., 2022), which we train ourselves due to the lack of publicly available checkpoints. All publicly available models are sourced from official repositories or HuggingFace implementations, where applicable.

### 4.2 Robustness in Noise Environment

We evaluate robustness to noise and channel variability using two benchmarks: synthetic distortions applied to CommonVoice and real-world social conditions captured in the AMI corpus. For synthetic testing, we simulate Gaussian, environmental, and cross-talk noise at varying SNRs (5, 15, 25 dB) with and without room impulse response (RIR) of severity levels 2, 3, and 4. These are evaluated under three codecs (GSM, AMR, Opus), with results shown in Table 4. For real-world testing, we use AMI recordings captured in near-field and far-field microphone setups to assess model performance in natural interactive environments.

As shown in 4 and 13, we observe a consistent trend: WavLM-based models degrade rapidly in the presence of noise and reverberation, especially when combined with low-bitrate codecs such as GSM or AMR. For instance, WavLM-Base shows a sharp EER increase from 23.05% in clean conditions to over 40% across nearly all noisy+RIR combinations. In contrast, RedimNet, ECAPA-TDNN, and MFA-Conformer exhibit significantly stronger robustness, maintaining substantially lower EERs in both synthetic and real-world conditions. These results emphasize the need for channel-aware model development and highlight the importance of including realistic acoustic variation during training.

### 4.3 Robustness in Adversarial Scenarios

We run different tests for adversarial attacks and TTS spoofing. For the TTS spoofing, we use the EARS dataset, which contains 109 speakers (59 female, 50 male). For Real vs. Synthetic evaluations, both utterances in a pair originate from the same speaker. Positive pairs consist of two real utterances, while negative pairs include one real and one synthetic sample generated by a TTS system. For adversarial attacks, we adopt a targeted verification setup in which the first utterance is adversarially perturbed using either FGSM or FakeBob,

and the second is a clean utterance from the same speaker. This design ensures speaker consistency while isolating the effect of the perturbation.

As shown in Table 1, MFA-Conformer demonstrates the strongest robustness across all tested conditions, achieving 0% EER under TTS spoofing and the lowest error rates under both FGSM (45.26%) and FakeBob (35.48%) attacks. RedimNet also performs well under TTS spoofing (5.94%), though it is more susceptible to adversarial attacks. In contrast, ECAPA-TDNN is highly vulnerable to FakeBob, reaching an EER of 62.36%. The WavLM-based models (Base and Base+) show consistent vulnerability under both spoofing and adversarial conditions. These findings highlight substantial variability in model robustness and underscore the need to develop verification systems that are resilient to both synthetic speech and adversarial manipulation.

### 4.4 Robustness in Demographic Variations

To investigate demographic bias in speaker verification, we generate verification pairs by first splitting the dataset by gender, and then further dividing each gender group based on the desired demographic category such as age, ethnicity, or language. Within each subgroup, we form same-speaker pairs and compute the EER independently for each model. This stratified evaluation allows us to analyze whether models exhibit bias or performance disparities across demographic dimensions, especially among underrepresented groups. We run these experiments on both the EARS and Mozilla CommonVoice datasets, leveraging their detailed metadata.

As seen in Table 2, certain language-gender or ethnicity-gender combinations (e.g., male-Hispanic, female-Asian) have significantly fewer speakers and exhibit elevated EERs, suggesting weaker generalization. Our analysis reveals signs of demographic bias in several models. For instance, WavLM-Base shows degradation in older age groups. In contrast, Redimnet maintains the most consistent performance regarding all demographic splits, with minimal variation across age, gender, and ethnicity.

Using paired t-tests across five models in the EARS dataset (Richter et al., 2024), we assess consistency of group-level EER differences. Males show higher EERs than females (17.3% vs. 13.7%), but the gap is not statistically significant ($p = 0.095$). Younger males (18–25) outperform older males ($p < 0.05$). Females aged 36–45 significantly outperform other female age groups ($p < 0.01$). Black females show significantly higher EERs than white females ($p < 0.001$). Asian and Hispanic males also perform worse, but sample sizes are small ($n \leq 5$) (see Appendix A.1).

CommonVoice results (Appendix A.2) confirm these trends. Gender identity groups show no significant EER gap ($p = 0.779$), but age remains a major factor. Older male and female speakers (60+) consistently underperform compared to younger groups ($p < 0.01$).

These results collectively suggest that demographic imbalance in training data may contribute to uneven generalization and reduced fairness, particularly in age and ethnicity subgroups with limited representation.

## 5 Conclusion

We present *SVeritas*, a comprehensive and extensible benchmark for evaluating speaker verification models under diverse real-world and synthetic stressors. Unlike prior work, it covers environmental noise, channel mismatches, codecs, cross-lingual variation, demographic shifts, adversarial attacks, and importantly, TTS-based spoofing—an often overlooked but growing threat. SVeritas enables fine-grained robustness and fairness analysis across gender, age, ethnicity, and language, while offering a modular framework that supports easy integration of new models and evaluation settings. It provides a unified, reproducible foundation for building SV systems that are not only accurate, but also resilient, equitable, and ready for real-world deployment.

## Limitations

While *SVeritas* provides a broad and extensible evaluation framework, it currently applies stress conditions at fixed levels of severity. This design simplifies benchmarking and ensures consistency across models, but may not fully capture how systems degrade under progressively harder conditions. In real-world scenarios, distortions such as noise, reverberation, or compression vary in intensity and interact in complex ways. Future work could extend SVeritas with parameterized or continuous stress levels, enabling finer-grained robustness analysis and stress-adaptive training strategies

## References


Abdulhady Abas Abdullah, Soran Badawi, Dana A Abdullah, Dana Rasul Hamad, Hanan Abdulrahman Taher, Sabat Salih Muhamad, Aram Mahmood Ahmed, Bryar A Hassan, Sirwan Abdolwahed Aula, and Tarik A Rashid. 2025. From dialect gaps to identity maps: Tackling variability in speaker verification. *arXiv preprint arXiv:2505.04629*.

Khamis A Al-Karawi. 2021. Mitigate the reverberation effect on the speaker verification performance using different methods. *International Journal of Speech Technology*, 24(1):143–153.

Moustafa Alzantot, Bharathan Balaji, and Mani Srivastava. 2018. Did you hear that? adversarial examples against automatic speech recognition. *arXiv preprint arXiv:1801.00554*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222.

CCITT. 1988. Pulse code modulation (pcm) of voice frequencies. In *ITU*.

Guangke Chen, Sen Chenb, Lingling Fan, Xiaoning Du, Zhe Zhao, Fu Song, and Yang Liu. 2021. Who is real bob? adversarial attacks on speaker recognition systems. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 694–711. IEEE.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2010. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *Interspeech*.

Subhadeep Dey. 2018. Phonetic aware techniques for speaker verification. Technical report, EPFL.

George R Doddington. 2012. The effect of target/non-target age difference on speaker recognition performance. In *Odyssey*, pages 263–267.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, and 1 others. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Amirhossein Hajavi and Ali Etemad. 2023. A study on bias and fairness in deep speaker recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Rosa González Hautamäki, Tomi Kinnunen, Ville Hautamäki, Timo Leino, and Anne-Maria Laukkanen. 2013. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In *Interspeech*, pages 930–934. Citeseer.

Jan Hintz and Ingo Siegert. Commonbench: A larger scale speaker verification benchmark.

Wiebke Toussaint Hutiri and Aaron Yi Ding. 2022. Bias in automated speaker recognition. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 230–247.

Nakamasa Inoue and Keita Goto. 2020. Semi-supervised contrastive learning with generalized contrastive loss and its application to speaker recognition. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1641–1646. IEEE.

Arindam Jati, Chin-Cheng Hsu, Monisankha Pal, Raghuveer Peri, Wael AbdAlmageed, and Shrikanth Narayanan. 2021. Adversarial attack and defense strategies for deep speaker recognition systems. *Computer Speech & Language*, 68:101199.

Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M Khapra. 2023. Indicsuperb: a speech processing universal performance benchmark for indian languages. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, pages 12942–12950.

Finnian Kelly and Naomi Harte. 2011. Effects of long-term ageing on speaker verification. In *European Workshop on Biometrics and Identity Management*, pages 113–124. Springer.

Patrick Kenny. 2010. Bayesian speaker verification with heavy-tailed priors. In *Proc. Odyssey 2010*, pages paper–14.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5220–5224. IEEE.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*, pages 1–4.

Jatin Kumar, Bikash Agarwalla, Sriram Srinivasan, King Wei Hor, and Tim Wong. 2024. MLow: Meta's low bitrate audio codec. Engineering at Meta. Accessed: 2025-05-20.

Yuke Lin, Xiaoyi Qin, Guoqing Zhao, Ming Cheng, Ning Jiang, Haiying Wu, and Ming Li. 2024. Voxblink: A large scale speaker verification dataset on camera. In *ICASSP*, pages 10271–10275. IEEE.

Ladislav Mošner, Oldřich Plchot, Lukáš Burget, and Jan Honza Černockỳ. 2022. Multisv: Dataset for far-field multi-channel speaker verification. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 7977–7981. IEEE.

Arsha Nagrani, Joon Son Chung, Jaesung Huh, Andrew Brown, Ernesto Coto, Weidi Xie, Mitchell McLaren, Douglas A Reynolds, and Andrew Zisserman. 2020. Voxsrc 2020: The second voxceleb speaker recognition challenge. *arXiv preprint arXiv:2012.06867*.

Masanobu Nakamura, Koji Iwano, and Sadaoki Furui. 2008. Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, 22(2):171–184.

KiHyun Nam, Hee-Soo Heo, Jee-weon Jung, and Joon Son Chung. 2024. Disentangled representation learning for environment-agnostic speaker recognition. *arXiv preprint arXiv:2406.14559*.

Mahesh Kumar Nandwana, Julien van Hout, Mitchell McLaren, Allen R Stauffer, Colleen Richey, Aaron Lawson, and Martin Graciarena. 2018. Robust speaker recognition from distant speech under real reverberant environments using speaker embeddings. In *Interspeech*, pages 1106–1110.

Vendela Njegovec. 2025. Forensic automatic speaker recognition: Analyzing codecs for calibration and their impact on system performance. Master's thesis, University of Twente.

Arnab Poddar, Md Sahidullah, and Goutam Saha. 2018. Speaker verification with short utterances: a review of challenges, trends and opportunities. *IET Biometrics*, 7(2):91–101.

Jozef Polacky, Peter Pocta, and Roman Jarina. 2016. An impact of wideband speech codec mismatch on a performance of gmm-ubm speaker verification over telecommunication channel. In *2016 ELEKTRO*, pages 77–82. IEEE.

Xiaoyi Qin, Ming Li, Hui Bu, Wei Rao, Rohan Kumar Das, Shrikanth Narayanan, and Haizhou Li. 2020. The interspeech 2020 far-field speaker verification challenge.

Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. 2000. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41.

Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinjii Watanabe, Alexander Richard, and Timo Gerkmann. 2024. EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. In *Interspeech*.

Seyed Omid Sadjadi, Craig Greenberg, Elliot Singer, Lisa Mason, and Douglas Reynolds. 2022. The 2021 nist speaker recognition evaluation. *arXiv preprint arXiv:2204.10242*.

Seyed Omid Sadjadi, Timothée Kheyrkhah, Audrey Tong, Craig Greenberg, Douglas Reynolds, Elliot Singer, Lisa Mason, and Jaime Hernandez-Cordero. 2017. The 2016 nist speaker recognition evaluation. *Interspeech 2017*.

Muhammad A Shah, David Solans Noguero, Mikko A Heikkila, Bhiksha Raj, and Nicolas Kourtellis. 2025. Speech robust bench: a robustness benchmark for speech recognition. *ICLR*.

Johan Sjoberg, Magnus Westerlund, Ari Lakaniemi, and Qiaobing Xie. 2007. Rtp payload format and file storage format for the adaptive multi-rate (amr) and adaptive multi-rate wideband (amr-wb) audio codecs. Technical report.

ETSI TC-SMG. 1993. European digital cellular telecommunications system (phase 2); full rate speech transcoding (gsm 06.10). *European Telecommunication Standard ETS*, 300:580–2.

Jenthe Thienpondt and Kris Demuynck. 2023. Ecapa2: A hybrid neural network architecture and training strategy for robust speaker embeddings. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Jean-Marc Valin, Koen Vos, and Timothy Terriberry. 2012. Definition of the opus audio codec. Technical report.

Jesús Villalba, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Jonas Borgstrom, Leibny Paola García-Perera, Fred Richardson, Réda Dehak, and 1 others. 2020. State-of-the-art speaker recognition with neural network embeddings in nist sre18 and speakers in the wild evaluations. *Computer Speech & Language*, 60:101026.

Zhizheng Wu, Junichi Yamagishi, Tomi Kinnunen, Cemal Hanilçi, Mohammed Sahidullah, Aleksandr Sizov, Nicholas Evans, Massimiliano Todisco, and Hector Delgado. 2017. Asvspoof: the automatic speaker verification spoofing and countermeasures challenge. *IEEE Journal of Selected Topics in Signal Processing*, 11(4):588–604.

Ivan Yakovlev, Rostislav Makarov, Andrei Balykin, Pavel Malov, Anton Okhotnikov, and Nikita Torgashov. 2024. Reshape dimensions network for speaker recognition. In *Proc. Interspeech 2024*, pages 3235–3239.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, and 1 others. 2021. Superb: Speech processing universal performance benchmark. *Interspeech*.

Hossein Zeinali, Kong Aik Lee, Jahangir Alam, and Lukas Burget. 2019. Short-duration speaker verification (sdsv) challenge 2021: the challenge evaluation plan. *arXiv preprint arXiv:1912.06311*.

Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung-yi Lee, and Helen Meng. 2022. Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification. In *Proc. Interspeech 2022*, pages 306–310.

Zhenyu Zhou, Junhui Chen, Namin Wang, Lantian Li, and Dong Wang. 2023. Adversarial data augmentation for robust speaker verification. In *Proceedings of the 2023 9th International Conference on Communication and Information Processing*, pages 226–230.

Zoom Video Communications, Inc. n.d. *Premium Audio*. Zoom. PDF.

Chu-Xiao Zuo, Zhi-Jun Jia, and Wu-Jun Li. 2024. Advtts: Adversarial text-to-speech synthesis attack on speaker identification systems. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4840–4844. IEEE.

# A Pairwise Statistical Tests

Each demographic group has a single EER value per model. While speaker counts are sometimes small (as few as 2–3), the number of utterances per speaker is high, resulting in relatively stable EER estimates per group. Traditional t-tests require multiple observations per group; instead, we leverage the fact that we have five models and treat their EERs as paired samples.

We apply a paired t-test to assess whether models consistently yield higher EERs for one group than another. This test evaluates the consistency of performance differences across models, not population-level differences. We report the number of speakers per group and disregard comparisons involving statistically underpowered cases.

Each table below compares pairwise group performance:

- **Rows:** Reference groups (with speaker count).
- **Columns:** Comparison groups (with speaker count).
- **Cells:** `t-stat / p-value / significance`
  - $t$-stat $< 0$: comparison group has *higher* EER than reference.
  - $t$-stat $> 0$: comparison group has *lower* EER than reference.
  - Significance levels: *** for $p < 0.001$ (very highly significant); ** for $p < 0.01$ (highly significant); * for $p < 0.05$ (statistically significant); No star for $p \geq 0.05$ (not significant).

## A.1 EARS Dataset (Richter et al., 2024)

| | F (n=59) | M (n=43) |
|---|---|---|
| F (n=59) | — | -2.18 / 0.095 / |
| M (n=43) | 2.18 / 0.095 / | — |

Table 5: EARS: Pairwise t-tests for gender groups (t / p / sig).

| | M_18-25 (n=14) | M_26-35 (n=10) | M_36-45 (n=10) | M_46-55 (n=4) | M_56-65 (n=5) |
|---|---|---|---|---|---|
| M_18-25 (n=14) | — | 3.70 / 0.021 / * | 3.87 / 0.018 / * | 0.34 / 0.751 / | -0.45 / 0.678 / |
| M_26-35 (n=10) | -3.70 / 0.021 / * | — | 2.43 / 0.072 / | -2.89 / 0.044 / * | -2.32 / 0.081 / |
| M_36-45 (n=10) | -3.87 / 0.018 / * | -2.43 / 0.072 / | — | -3.04 / 0.038 / * | -2.47 / 0.069 / |
| M_46-55 (n=4) | -0.34 / 0.751 / | 2.89 / 0.044 / * | 3.04 / 0.038 / * | — | -0.78 / 0.478 / |
| M_56-65 (n=5) | 0.45 / 0.678 / | 2.32 / 0.081 / | 2.47 / 0.069 / | 0.78 / 0.478 / | — |

Table 6: EARS: Pairwise t-tests for male age groups (t / p / sig).

| | F_18-25 (n=13) | F_26-35 (n=13) | F_36-45 (n=7) | F_46-55 (n=14) | F_56-65 (n=10) | F_66-75 (n=2) |
|---|---|---|---|---|---|---|
| F_18-25 (n=13) | — | 0.17 / 0.871 / | 4.88 / 0.008 / ** | 0.94 / 0.398 / | -0.04 / 0.967 / | -0.78 / 0.481 / |
| F_26-35 (n=13) | -0.17 / 0.871 / | — | 5.15 / 0.007 / ** | 6.02 / 0.004 / ** | -0.32 / 0.762 / | -1.12 / 0.327 / |
| F_36-45 (n=7) | -4.88 / 0.008 / ** | -5.15 / 0.007 / ** | — | -4.54 / 0.010 / * | -2.58 / 0.061 / | -1.93 / 0.126 / |
| F_46-55 (n=14) | -0.94 / 0.398 / | -6.02 / 0.004 / ** | 4.54 / 0.010 / * | — | -1.24 / 0.283 / | -1.37 / 0.243 / |
| F_56-65 (n=10) | 0.04 / 0.967 / | 0.32 / 0.762 / | 2.58 / 0.061 / | 1.24 / 0.283 / | — | -1.34 / 0.252 / |
| F_66-75 (n=2) | 0.78 / 0.481 / | 1.12 / 0.327 / | 1.93 / 0.126 / | 1.37 / 0.243 / | 1.34 / 0.252 / | — |

Table 7: EARS: Pairwise t-tests for female age groups (t / p / sig).

| | F_white (n=40) | F_black (n=13) | F_asian (n=2) | F_hispanic (n=4) | M_white (n=31) | M_black (n=5) | M_asian (n=2) | M_hispanic (n=5) |
|---|---|---|---|---|---|---|---|---|
| F_white (n=40) | — | -9.76 / 0.001 / *** | 1.32 / 0.257 / | 2.73 / 0.053 / | -2.29 / 0.084 / | -0.78 / 0.477 / | 1.27 / 0.274 / | -2.30 / 0.083 / |
| F_black (n=13) | 9.76 / 0.001 / *** | — | 1.79 / 0.148 / | 3.57 / 0.023 / * | -0.69 / 0.527 / | 0.80 / 0.466 / | 2.19 / 0.094 / | -1.22 / 0.289 / |
| F_asian (n=2) | -1.32 / 0.257 / | -1.79 / 0.148 / | — | 0.10 / 0.923 / | -1.62 / 0.180 / | -1.20 / 0.295 / | -0.56 / 0.608 / | -1.72 / 0.161 / |
| F_hispanic (n=4) | -2.73 / 0.053 / | -3.57 / 0.023 / * | -0.10 / 0.923 / | — | -2.62 / 0.059 / | -2.02 / 0.113 / | -0.91 / 0.415 / | -2.56 / 0.062 / |
| M_white (n=31) | 2.29 / 0.084 / | 0.69 / 0.527 / | 1.62 / 0.180 / | 2.62 / 0.059 / | — | 1.73 / 0.159 / | 4.52 / 0.011 / * | -1.88 / 0.133 / |
| M_black (n=5) | 0.78 / 0.477 / | -0.80 / 0.466 / | 1.20 / 0.295 / | 2.02 / 0.113 / | -1.73 / 0.159 / | — | 1.72 / 0.161 / | -2.32 / 0.081 / |
| M_asian (n=2) | -1.27 / 0.274 / | -2.19 / 0.094 / | 0.56 / 0.608 / | 0.91 / 0.415 / | -4.52 / 0.011 / * | -1.72 / 0.161 / | — | -4.09 / 0.015 / * |
| M_hispanic (n=5) | 2.30 / 0.083 / | 1.22 / 0.289 / | 1.72 / 0.161 / | 2.56 / 0.062 / | 1.88 / 0.133 / | 2.32 / 0.081 / | 4.09 / 0.015 / * | — |

Table 8: EARS: Pairwise t-tests for ethnicity groups (t / p / sig).

## A.2 CommonVoice Dataset (Ardila et al., 2020)

| | female_feminine (n=45) | male_masculine (n=21) |
|---|---|---|
| female_feminine (n=45) | — | 0.30 / 0.779 / |
| male_masculine (n=21) | -0.30 / 0.779 / | — |

Table 9: CommonVoice: Pairwise t-tests for gender identity groups (t / p / sig).

| | M_teens (n=112) | M_twenties (n=582) | M_thirties (n=240) | M_fourties (n=140) | M_fifties (n=126) | M_sixties (n=57) | M_seventy_plus (n=69) |
|---|---|---|---|---|---|---|---|
| M_teens (n=112) | — | -2.62/0.059/ | 6.61/0.003/** | 5.44/0.006/** | 11.14/0.000/*** | 5.21/0.006/** | -2.31/0.082/ |
| M_twenties (n=582) | 2.62/0.059/ | — | 13.86/0.000/*** | 14.78/0.000/*** | 16.52/0.000/*** | 8.21/0.001/** | -1.88/0.133/ |
| M_thirties (n=240) | -6.61/0.003/** | -13.86/0.000/*** | — | 2.18/0.095/ | 4.03/0.016/* | -0.12/0.909/ | -4.47/0.011/* |
| M_fourties (n=140) | -5.44/0.006/** | -14.78/0.000/*** | -2.18/0.095/ | — | -0.24/0.821/ | -2.45/0.070/ | -7.39/0.002/** |
| M_fifties (n=126) | -11.14/0.000/*** | -16.52/0.000/*** | -4.03/0.016/* | 0.24/0.821/ | — | -2.63/0.058/ | -5.29/0.006/** |
| M_sixties (n=57) | -5.21/0.006/** | -8.21/0.001/** | 0.12/0.909/ | 2.45/0.070/ | 2.63/0.058/ | — | -4.73/0.009/** |
| M_seventy_plus (n=69) | 2.31/0.082/ | 1.88/0.133/ | 4.47/0.011/* | 7.39/0.002/** | 5.29/0.006/** | 4.73/0.009/** | — |

Table 10: CommonVoice: Pairwise t-tests for male age groups (t / p / sig).

| | F_teens (n=112) | F_twenties (n=582) | F_thirties (n=240) | F_fourties (n=140) | F_fifties (n=110) | F_sixties (n=49) | F_seventy_plus (n=17) |
|---|---|---|---|---|---|---|---|
| F_teens (n=112) | — | 12.67/0.000/*** | 24.19/0.000/*** | 25.00/0.000/*** | 12.72/0.000/*** | 5.07/0.007/** | 9.42/0.001/*** |
| F_twenties (n=582) | -12.67/0.000/*** | — | -0.15/0.885/ | 3.13/0.035/* | -1.25/0.279/ | -5.92/0.004/** | 2.48/0.068/ |
| F_thirties (n=240) | -24.19/0.000/*** | 0.15/0.885/ | — | 2.69/0.055/ | -2.89/0.045/* | -7.12/0.002/** | 1.85/0.138/ |
| F_fourties (n=140) | -25.00/0.000/*** | -3.13/0.035/* | -2.69/0.055/ | — | -3.27/0.031/* | -7.92/0.001/** | 1.09/0.338/ |
| F_fifties (n=110) | -12.72/0.000/*** | 1.25/0.279/ | 2.89/0.045/* | 3.27/0.031/* | — | -7.45/0.002/** | 2.36/0.078/ |
| F_sixties (n=49) | -5.07/0.007/** | 5.92/0.004/** | 7.12/0.002/** | 7.92/0.001/** | 7.45/0.002/** | — | 6.02/0.004/** |
| F_seventy_plus (n=17) | -9.42/0.001/*** | -2.48/0.068/ | -1.85/0.138/ | -1.09/0.338/ | -2.36/0.078/ | -6.02/0.004/** | — |

Table 11: CommonVoice: Pairwise t-tests for female age groups (t / p / sig).

| Gender | Speakers | WavLM-Base | WavLM-Base+ | RedimNet | ECAPA | MFA-Conformer |
|---|---|---|---|---|---|---|
| Female | 45 | 28.63% | 20.36% | 3.47% | 5.38% | 6.85% |
| Male | 21 | 17.42% | 17.56% | 5.41% | 9.84% | 10.16% |

Table 12: CommonVoice: EER when target pairs come from different lanuage while non-target pairs come from the same language.

| Noise | SNR | RIR | WavLM-Base | WavLM-Base+ | RedimNet | ECAPA | MFA-Conformer |
|---|---|---|---|---|---|---|---|
| Clean | | | 23.05% | 20.23% | 4.69% | 6.13% | 6.65% |
| GaussNoise | 5 | | 31.86% | 34.53% | 21.75% | 18.95% | 31.51% |
| | 15 | | 32.00% | 31.66% | 16.02% | 12.66% | 18.85% |
| | 25 | | 30.11% | 29.75% | 9.47% | 9.21% | 12.03% |
| GaussNoise w/ RIR | 5 | 2 | 46.57% | 47.67% | 28.19% | 21.89% | 35.36% |
| | 15 | 3 | 40.86% | 36.91% | 22.52% | 18.03% | 27.54% |
| | 25 | 4 | 46.42% | 44.18% | 42.82% | 35.50% | 37.37% |
| EnvNoise | 5 | | 42.01% | 43.05% | 25.04% | 24.46% | 26.70% |
| | 15 | | 34.79% | 32.34% | 10.52% | 13.62% | 15.34% |
| | 25 | | 27.30% | 23.15% | 5.95% | 8.49% | 9.59% |
| EnvNoise w/ RIR | 5 | 2 | 45.17% | 46.32% | 32.75% | 25.82% | 27.26% |
| | 15 | 3 | 40.25% | 38.65% | 19.36% | 15.88% | 20.04% |
| | 25 | 4 | 46.37% | 44.19% | 39.43% | 35.93% | 36.71% |
| CrossTalk | 5 | | 47.54% | 47.35% | 38.97% | 37.97% | 36.15% |
| | 15 | | 43.41% | 41.61% | 20.50% | 26.00% | 24.11% |
| | 25 | | 35.37% | 29.27% | 8.59% | 15.06% | 14.02% |
| CrossTalk w/ RIR | 5 | 2 | 46.19% | 46.75% | 41.23% | 37.73% | 36.13% |
| | 15 | 3 | 40.54% | 38.48% | 26.19% | 24.54% | 24.63% |
| | 25 | 4 | 46.44% | 44.20% | 39.25% | 36.55% | 37.41% |

Table 13: CommonVoice: EER degradation due to noise and room reverberations.