

Beyond EER: Multi-Dimensional Evaluation of Information Leakage in Speaker De-Identification

Seungmin Seo

Contractor Associate, National Institute of Standards and Technology, Gaithersburg, MD, USA
Chakra Consulting Inc., Clarksburg, MD, USA

Oleg Aulov P. Jonathon Phillips Kevin Mangold Jonathan Eskin

National Institute of Standards and Technology, Gaithersburg, MD, USA

{seungmin.seo, oleg.aulov, jonathon.phillips, kevin.mangold, jonathan.eskin}@nist.gov

Abstract

Speaker de-identification (SDID) aims to preserve privacy by concealing speaker identity while maintaining speech utility. However, current evaluations often reduce privacy to a single dimension—biometric verification performance—typically measured by Equal Error Rate (EER). This narrow focus ignores critical leakage channels, such as soft biometric inference, embedding-level re-identification, and structural template similarity, which threaten the unlinkability and irreversibility of biometric references. We propose a holistic evaluation framework across five complementary metrics: (i) EER, (ii) soft biometric leakage score, (iii) cumulative match characteristic re-identification analysis, (iv) canonical correlation analysis and Procrustes embedding alignment, and (v) intelligibility via word error rate and semantic similarity. Evaluating five SDID systems from the IARPA ARTS program, we demonstrate that these metrics capture independent dimensions of information leakage. Our results indicate that reliance on a single metric can misrepresent the privacy properties of an SDID system.

1. Introduction

Speech signals encode biometric and behavioral attributes beyond linguistic content, implicitly revealing a speaker’s sex, age, and accent [16, 9]. Speaker de-identification (SDID) systems aim to transform these signals to remain intelligible while concealing the speaker’s identity [24, 12, 21, 26]. However, the growing availability of high-accuracy, publicly accessible attribute classifiers [27] has fundamentally changed the threat landscape: modern speech representation models enable accurate infer-

ence of biometric attributes even from anonymized speech. This raises a central question: **How much identity information remains detectable in speech processed by current SDID systems, and along which dimensions can it be measured?**

This pattern echoes findings in other biometric modalities: in face recognition, embeddings have been shown to inadvertently store demographic, paralinguistic, and extrinsic factors beyond identity [13, 5, 3, 14]. Yet while the VoicePrivacy Challenges [24, 21, 12, 22] have established a benchmark for voice anonymization, evaluation of these residual signals in speech remains fragmented.

Most current evaluations rely on a solitary metric such as Equal Error Rate (EER) for speaker verification [24, 12, 8, 10, 11, 28, 21], which captures resistance to direct re-identification but not whether an adversary can recover demographic attributes or exploit systematic patterns across a population. Privacy metrics alone are also insufficient without understanding their cost in utility: a system that renders speech unintelligible achieves perfect privacy at the expense of speech communication’s fundamental purpose.

This paper makes the following contributions:

1. Evaluation across complementary privacy metrics.

We evaluate SDID systems using various privacy metrics spanning verification resistance (EER), attribute leakage (SBLs), search-based re-identification using cumulative match characteristic (CMC), and

Disclaimer: Certain equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement of any product or service by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose. These opinions, recommendations, findings, and conclusions do not necessarily reflect the views or policies of NIST or the United States Government.

embedding-space similarity measured by canonical correlation analysis (CCA) and Procrustes alignment. These metrics capture distinct aspects of residual speaker information and can yield different assessments of privacy.

2. **Biometric leakage analysis on anonymized speech.** We extend SBLS[19] to include accent prediction and introduce subgroup protection analysis to measure demographic disparities in attribute leakage.
3. **Large-scale cross-corpus evaluation.** We evaluate SDID systems across four test sets from three corpora (Mixer 3, 6, and 7), covering multiple accents and sampling conditions with more than 22,000 segments per system and over 3.4 million verification trials.
4. **Privacy-utility quantification.** We measure the impact of anonymization on intelligibility using word error rate (WER) and semantic similarity, revealing a measurable trade-off between identity suppression and speech utility.

Comparing system performance across these dimensions, we find that no system achieves top performance across all of them: the metrics capture fundamentally different aspects of privacy that can trade off against one another, and reliance on any single metric yields an incomplete—and potentially misleading—assessment.

2. Related Work

2.1. Speaker De-Identification Systems and Evaluation

SDID spans signal-processing transforms and neural voice conversion. The VoicePrivacy Challenge [24, 12, 21, 22] established a common benchmark, primarily evaluating systems using EER and log-likelihood ratio cost. Subsequent work explores diverse approaches, including speaker attribute perturbation [1], voice conversion with alternative distance metrics and kNN conversion in self-supervised spaces [26]. Despite architectural diversity, evaluation remains largely centered on speaker verification, offering limited insight into demographic leakage or privacy-utility trade-offs.

2.2. Identity Information Leakage in Anonymized Speech

Recent work shows that anonymized speech can still contain residual identity information detectable through modern speaker representation models [8, 11, 28, 23, 10]. These studies primarily evaluate speaker-level resistance using speaker verification metrics.

Beyond speaker identity, speech embeddings also encode soft biometric attributes such as sex, age, and accent [27]. VoxProfile [4] demonstrates large-scale extraction of speaker traits from foundation-model embeddings, while prior studies on benchmark dataset usage dynamics [16] highlight privacy challenges.

Together, these findings suggest that resistance to speaker re-identification alone does not guarantee broader privacy protection. Anonymized speech may still reveal demographic attributes or structural embedding similarities. This motivates evaluating SDID systems across multiple complementary metrics of privacy and utility.

3. Evaluation Setup

3.1. Corpora and Test Conditions

Evaluation data were developed by NIST from selected segments of the Mixer 3, 6, and 7 corpora collected by the Linguistic Data Consortium (LDC). Segments of 10, 30, and 60 seconds were generated using the LDC Broad Phonetic Class Speech Activity Detector [17]. The first 60 seconds of each recording were discarded to exclude greetings and channel stabilization effects. Test and initialization recordings were disjoint.

The test sets were designed to address complementary evaluation goals. Test 1 enables controlled sample-rate comparison (8 kHz vs. 16 kHz) on the same speakers using Mixer 6 CHiME interview recordings with available human transcriptions. Test 2 uses Mixer 3 8 kHz sampled English data. Test 3 and 4 introduce linguistic diversity: Mixer 7 Spanish speakers and Mixer 3 Hindi speakers producing English speech, enabling analysis of non-native accent handling. The statistics of test sets are shown in Table 1.

The Mixer/ARTS corpora were selected over alternatives such as the VoicePrivacy Challenge data because they provide the necessary speaker metadata, including annotations for sex, age group, and native language that are not available in existing privacy-evaluation benchmarks.

Demographic distributions are imbalanced across test sets. Test 3 contains no Young speakers and is 73% Female; Test 4 is sex-balanced but skewed Adult; These conditions reflect realistic deployment scenarios and expose potential subgroup vulnerabilities.

3.2. Trials

We employ five trial types: three designed to assess resistance to speaker identification (SID) attacks, one to evaluate pseudo-identity consistency, and one to measure pseudo-identity distinctness. Here, a pseudo-identity refers to the anonymized speaker profile produced by the SDID system, where each original speaker is mapped to a consistent synthetic identity across utterances (e.g., Alice → pseudo-Alice). Table 2 summarizes the trial types.

	Test 1 (16k)	Test 1 (8k)	Test 2 (8k)	Test 3 (16k)	Test 4 (8k)
Original corpus	Mixer 6	Mixer 6	Mixer 3	Mixer 7	Mixer 3
Total trials	277,634	277,634	951,025	991,474	1,244,014
Target trials	4,838	4,838	10,458	52,300	77,434
Non-target trials	272,796	272,796	940,567	939,174	1,166,580
Unique speakers	76	76	223	60	76
Unique segments	1,058	1,058	2,983	2,778	3,355
Male speakers	43	43	81	13	37
Female speakers	33	33	142	47	39
Adult speakers (25–54)	47	47	161	40	63
Senior speakers (55–85)	4	4	30	20	13
Young speakers (18–24)	25	25	32	–	–
Native language	English	English	English	Spanish	Hindi

Table 1. Statistics of evaluation dataset.

Trial type	Composition (target vs. non-target)	Target EER	Purpose
<i>oaoa</i>	orig-anon vs. orig-anon	50%	SID attack resistance
<i>oao</i>	orig-anon vs. orig-orig	50%	SID attack resistance
<i>oaaa</i>	orig-anon vs. anon-anon	50%	SID attack resistance
<i>aaaa</i>	anon-anon vs. anon-anon	0%	Self-consistency
<i>cross-profile</i>	p1-p1 vs. p1-p2 (cross pseudo-profile)	0%	Pseudo-identity distinctness

Table 2. Five different trial types used in evaluation.

A well-performing anonymization system should produce voices that cannot be traced back to the original speaker. Pseudo-speaker output should functionally behave like real speaker output in that SID systems consistently match the same pseudo-speaker to itself, and not to other pseudo-speakers or real speakers.

The three SID attack trials assess different aspects of this goal. In *oaoa*, original segments are compared against anonymized segments (e.g., Alice vs. pseudo-Bob), directly testing whether an attacker can match a known voice to its anonymized counterpart. In *oao*, original segments are compared against other original recordings (e.g., Alice vs. Bob), testing whether the presence of anonymized enrollment data degrades the SID backend’s ability to distinguish real speakers. In *oaaa*, anonymized segments are compared against other anonymized segments (e.g., pseudo-Alice vs. pseudo-Bob), testing whether different anonymized speakers remain distinguishable from one another. EER near 50% indicates strong performance on each. However, an inflated *oaaa* EER (well above 50%) can indicate that anonymized voices from different speakers have become acoustically similar, compressing the speaker identity space rather than successfully concealing individual identities.

The *aaaa* trial evaluates self-consistency with a target EER of 0%. Pseudo-Alice should consistently sound like pseudo-Alice, regardless of which utterance is anonymized. Higher EER indicates that different segments from the same speaker may sound like different people after processing. The *cross-profile* trial evaluates pseudo-identity distinctness, also targeting 0% EER. Pseudo-Alice should be

clearly distinguishable from pseudo-Bob; an EER near 50% would indicate that different pseudo-identities are acoustically interchangeable.

3.3. Speaker De-Identification Systems

Five SDID systems developed under the IARPA ARTS¹ program were evaluated: four performer submissions and one baseline system. The systems span diverse architectural approaches:

VOXLET maps audio to wav2vec 2.0 latent representations, applies differential privacy noise in latent space, and reconstructs speech using HiFiGAN 2.0.

RASP employs a disentangled autoencoder separating content (HuBERT), speaker identity, and pitch/energy. Speaker identity is replaced with a pseudo-speaker embedding selected via cosine similarity.

SHADOW uses an autoregressive language model over EnCodec tokens conditioned on Wav2Vec2 features. Pseudo-speaker embeddings are generated via PLDA and converted with FreeVC.

PHORTRESS decomposes speech using the SPARC articulatory coding framework. The speaker embedding is replaced with a fabricated identity, and a DDSP vocoder resynthesizes speech from articulatory features.

Baseline system performs k-nearest-neighbor regression on WavLM features, averaging matched pseudo-speaker embeddings and synthesizing with HiFiGAN.

¹<https://www.iarpa.gov/research-programs/arts>

3.4. Speaker Identification Backends

Four independent SID backends assess anonymization effectiveness:

- **NeMo TitaNet Large:** depth-wise separable convolutions with SE layers and channel-attention pooling, trained on VoxCeleb 1&2, Fisher, SwitchBoard, LibriSpeech, and augmented data [6].
- **NeMo ECAPA-TDNN:** TDNN with SE and Res2Block layers and multi-scale attention [2].
- **Hyperion:** ResNet-based x-vector extractor with PLDA backend trained on NIST SRE CTS Superset. [25]
- **OLIVE:** TDNN x-vector extractor with PNCC features and PLDA backend trained on NIST SRE 2004–2012, Mixer6, and VoxCeleb 1&2 [7].

3.5. Attribute Classifier and Automated Speech Recognition Systems

VoxProfile [4] serves as a strong attribute inference adversary for evaluating soft biometric leakage. While greater attacker diversity would offer a more complete assessment, using a single threshold-free attacker in our analysis reduces sensitivity to classifier calibration. Evaluated attributes include sex (binary), age group (Young 17–24, Adult 25–54, Senior 55+), and accent family (North America, Romance, South Asia). Accent is evaluated only on aggregated cross-test results.

For intelligibility assessment, OpenAI Whisper [15] and NVIDIA NeMo Canary-1B [18] transcribe original and anonymized speech. Canary-1B was additionally used because it exhibits fewer hallucinations during silent segments. The Whisper English text normalizer (lowercasing, contraction expansion, punctuation removal, numeric normalization) is applied for WER scoring.

4. Evaluation Metrics

We evaluate SDID systems along five complementary dimensions capturing speaker-level privacy using EER, soft biometric leakage using SBLS, re-identification risk using CMC, embedding structure using CCA, and utility preservation using WER.

4.1. Speaker Verification EER

EER measures the operating point at which the false acceptance rate equals the false rejection rate in a speaker verification task. EER results are aggregated across SID backends and pseudo-identity profiles. Higher EER in SID attack trials (target 50%) indicates stronger privacy; lower EER in consistency trials (target 0%) indicates reliable pseudo-identity behavior.

4.2. Soft Biometric Leakage Score (SBLS)

SBLS [19] quantifies soft biometric leakage via three components:

$$\text{SBLS} = \alpha P_{\text{attr}} + \beta P_{\text{assoc}} + \gamma P_{\text{subgroup}}, \quad (1)$$

where $\alpha, \beta, \gamma \geq 0$, $\alpha + \beta + \gamma = 1$, and each component lies in $[0, 1]$ (1 = maximal privacy). In this work, we set $\alpha = \beta = 0.4$ and $\gamma = 0.2$ to slightly downweight subgroup robustness, though the choice is heuristic. The impact of parameter is reported in section 5.2.3

1. **Zero-Shot Attribute Privacy (P_{attr}):** Let A be the set of attributes (e.g., $A = \{\text{Male/Female label, age group}\}$). For each attribute $a \in A$ with K_a classes, the frozen attacker outputs class scores on the anonymized dataset. We compute one-vs-rest AUC for each class from scores.

$$P_{\text{attr}} = 1 - \frac{1}{|A|} \sum_{a \in A} \frac{\max\{0, \text{mAUC}_a^* - 0.5\}}{0.5}. \quad (2)$$

Here $P_{\text{attr}}=1$ indicates chance-level zero-shot inference (low leakage) and $P_{\text{attr}}=0$ indicates near-perfect attribute recoverability (high leakage). When only hard predictions are available, we substitute macro balanced accuracy for mAUC_a^* and normalize by the chance-level baseline $1/K_a$.

2. **Systematic Association (P_{assoc}):** Let $\hat{A}_a^*$ denote hard predictions obtained via $\arg\max$ over the permutation-aligned scores from the previous step. Residual statistical dependence between true and predicted labels is measured using normalized mutual information:

$$P_{\text{assoc}} = 1 - \frac{1}{|A|} \sum_{a \in A} \frac{I(A_a; \hat{A}_a^*)}{\log K_a}. \quad (3)$$

where $I(A_a; \hat{A}_a^*)$ is the mutual information between true and predicted attributes. In practice, we estimate $\tilde{I}_a$ from the confusion matrix between $(A_a, \hat{A}_a^*)$ using standard entropy calculations. High $\tilde{I}_a$ indicates strong systematic dependence (higher leakage), while $\tilde{I}_a \approx 0$ suggests that predictions contain little information about true attributes beyond random chance.

3. **Subgroup Protection (P_{subgroup}):** Equity across demographic subgroups is evaluated as

$$P_{\text{subgroup}} = \omega(1 - \max_g L_g) + (1 - \omega) \frac{\min_g (1 - L_g)}{\max_g (1 - L_g)}, \quad (4)$$

where L_g denotes subgroup leakage and $\omega = 0.7$. Varying $\omega \in \{0.3, 0.5, 0.7, 0.9\}$ changes SBLs by at most 0.004 and does not affect rankings. When accent is included, demographic and accent subgroup protection are averaged equally.

4.3. CMC

CMC simulates a closed-set identification attack. For each anonymized embedding, cosine distances to all original embeddings are computed, and the rank of the first correct match is recorded:

$$\text{CMC}@k = \frac{1}{N} |\{i : \text{rank}_i \leq k\}|. \quad (5)$$

Lower CMC values indicate stronger anonymization. Both NeMo embeddings are used. A permutation bootstrap establishes chance-level performance.

4.4. CCA and Procrustes Alignment

CCA measures linear relationships between original and anonymized embedding subspaces. CCA is fit on matched pairs (80/20 train/test split), and the mean of the top-10 canonical correlations is reported on held-out data.

Procrustes alignment learns an orthogonal rotation R minimizing alignment error and evaluates alignment via mean cosine similarity.

A random permutation baseline breaks original-anonymized pairing, yielding reference levels (CCA Top-10 ≈ 0.734 , Procrustes cosine ≈ 0.29 – 0.31). Values near baseline indicate decorrelation; values approaching 1 indicate linear predictability.

4.5. Speech Intelligibility and Semantic Preservation

Utility cost is measured using two ASR systems: OpenAI Whisper and NVIDIA NeMo Canary-1B. WER is computed as

$$\text{WER} = \frac{S + D + I}{N}, \quad (6)$$

where S , D , and I denote substitutions, deletions, and insertions, and N is the number of reference words.

Because WER penalizes all errors equally regardless of semantic impact, we complement it with cosine similarity of sentence embeddings [20]. Cosine similarity ranges from 0 to 1, with higher values indicating stronger semantic preservation.

5. Evaluation Results

5.1. EER-Based Speaker Verification Results

We first report EER-based evaluation to establish the speaker-level privacy landscape before examining soft biometric and embedding-structure dimensions.

SID Backend	Test 1 (16k)	Test 1 (8k)	Test 2	Test 3	Test 4
NeMo TitaNet	4.75	29.70	3.82	4.51	6.49
NeMo ECAPA	4.81	22.14	3.64	5.21	5.06
Hyperion	4.70	7.55	4.86	5.50	4.48
SRI OLIVE	2.60	6.06	4.38	4.19	3.45

Table 3. Reference EER (%) on original speech for each SID backend across different test dataset.

System	oaoa	oaoo	oaaa	aaaa	cross-profile
Baseline	39.62 $\pm$ 1.19	55.42 $\pm$ 3.11	60.94 $\pm$ 3.99	4.90 $\pm$ 1.52	30.89 $\pm$ 1.27
VOXLET	27.75 $\pm$ 1.72	36.50 $\pm$ 4.53	66.79 $\pm$ 3.13	5.87 $\pm$ 1.15	50.01 $\pm$ 0.02
RASP	38.92 $\pm$ 3.07	48.82 $\pm$ 7.31	88.61 $\pm$ 2.95	20.04 $\pm$ 3.53	27.57 $\pm$ 2.93
SHADOW	45.40 $\pm$ 1.21	55.54 $\pm$ 4.42	80.49 $\pm$ 3.76	5.45 $\pm$ 0.86	3.47 $\pm$ 0.57
PHORTRESS	49.79 $\pm$ 0.60	64.08 $\pm$ 3.87	86.81 $\pm$ 2.02	24.62 $\pm$ 2.08	23.28 $\pm$ 3.10

Table 4. Mean EER (%) for trials, aggregated across all test sets and SID backends. Values are shown as *mean $\pm$ half-width of 95% CI*, computed by non-parametric bootstrap ($B = 10000$) over the per-condition EER values.

Table 3 shows the reference performance on original speech with different SID backends. All four SID backends achieve EER below 7% on most conditions, except Test 1 at 8 kHz, where TitaNet (29.7%) and ECAPA-TDNN (22.1%) show degraded performance. Hyperion (7.6%) and OLIVE (6.1%) remain robust.

5.1.1 Overall EER Landscape

Aggregating across all test sets and SID backends, the EER landscape varies substantially by system and trial type. For SID attack trials, mean EERs are shown in Table 4.

PHORTRESS is closest to the 50% target on *oaoa* trial, indicating the strongest resistance under the primary attack scenario. A consistent pattern appears in the *oaaa* trial: all systems exceed 50%, often substantially. This elevation suggests anonymization-induced compression of the speaker identity space, making non-target pairs harder to distinguish and inflating EER above chance.

For verification trials (target: 0% EER), Baseline, VOXLET, and SHADOW achieve low EER on *aaaa* trial types, indicating that their pseudo-speakers consistently sound like themselves across utterances. PHORTRESS shows the highest EER, suggesting that its anonymized segments for the same speaker can sound noticeably different from one another. Cross-profile results show a sharp contrast: SHADOW achieves low EER, meaning its pseudo-identities are readily distinguishable from one another (pseudo-Alice sounds different from pseudo-Bob). VOXLET is near chance (50%), meaning its pseudo-identities sound so similar that SID backends cannot tell them apart.

5.1.2 Privacy–Consistency Trade-off

Figure 1 illustrates that systems occupy distinct operating points in the privacy vs. consistency plane. PHORTRESS is the hardest system to trace back to the original speaker, but

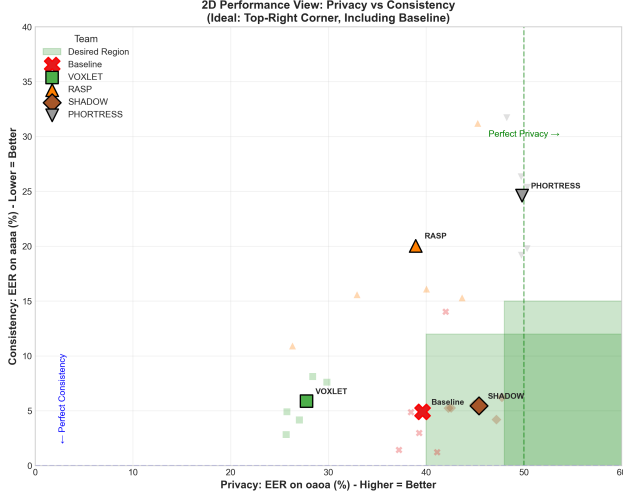


Figure 1. Privacy–consistency trade-off across systems. The ideal region combines high EER on *oaoa* trial (better privacy) with low EER on *aaaa* trial (better consistency).

its anonymized segments for the same speaker are less reliably recognized as belonging to the same pseudo-identity. SHADOW offers a more balanced profile, maintaining both reasonable privacy and consistent pseudo-speaker output. VOXLET exhibits strong consistency but weaker privacy.

5.1.3 Accent Effects

Evaluation on Test 3 and 4 assess robustness to non-native accents. Figure 2 shows the results. Differences in *oaoa* trial EER across accent conditions are generally modest relative to inter-team differences. In contrast, the results on *oaoa* and *oaaa* trials exhibit larger shifts for some teams (e.g., Test 4 Hindi increasing *oaoa* EER for every system), suggesting interactions between accent-dependent acoustics and trial-type composition.

5.1.4 Sample Rate Effects

Evaluation results on Test 1 dataset shown in Figure 3 provide a controlled comparison using the same 76 English speakers at both 16 kHz and 8 kHz. For *oaoa* trial, most systems show stable performance across sample rates, except RASP, which moves from 26.3% (16 kHz) to 45.3% (8 kHz), toward the 50% target. However, RASP’s *aaaa* EER degrades (10.9% to 31.2%), indicating that reduced bandwidth affects both attackers and within-system consistency.

5.2. SBLS Results: Soft Biometric Leakage

5.2.1 Overall Results (Sex + Age)

Table 5 summarizes SBLS using sex and age. All SDID systems improve over the unprocessed baseline, indicat-

ing meaningful but incomplete protection. Across systems, P_{assoc} is consistently high, suggesting errors are not driven by simple, exploitable mappings. The dominant differentiator is subgroup protection P_{subgroup} , where residual vulnerability remains.

Sex is consistently harder to mask than age: several systems yield near-perfect sex recovery ($\text{AUC} \approx 0.95$) even when age leakage is moderate. This asymmetry is consistent with deeply embedded sex-correlated acoustic cues (e.g., fundamental frequency and formant structure).

SBLS rankings are stable across test sets (Table 6). PHORTRESS ranks first and Baseline second in every condition; other systems swap only in isolated cases.

5.2.2 Accent as a Third Attribute

Adding accent family preserves top and bottom rankings while narrowing mid-tier differences (Table 7). PHORTRESS remains near-chance on accent ($\text{AUC} \approx 0.51$), whereas VOXLET exhibits substantial accent leakage ($\text{AUC} \approx 0.78$), producing the largest SBLS drop. RASP and SHADOW improve due to stronger accent subgroup protection.

5.2.3 Weight Sensitivity

We recompute SBLS under seven weighting schemes that vary (α, β, γ) for attribute predictability, systematic association, and subgroup protection: default (0.4,0.4,0.2), equal (0.33,0.33,0.34), α -heavy (0.6,0.2,0.2), β -heavy (0.2,0.6,0.2), γ -heavy (0.2,0.2,0.6), no-sub (0.5,0.5,0.0), and min-sub (0.45,0.45,0.1).

Results remain stable (Table 8): the top-3 and bottom-1 are invariant, and only RASP and SHADOW swap under equal and γ -heavy weighting. This reflects SHADOW’s stronger subgroup protection versus RASP’s higher P_{attr} and P_{assoc} . Removing or downweighting subgroup protection does not alter overall conclusions.

5.2.4 Component Correlation

With sex and age only, SBLS components are strongly positively correlated (Pearson $r > 0.80$), with perfect rank correlation between P_{attr} and P_{assoc} (Spearman $\rho = 1.0$). Including accent yields weaker correlations for accent subgroup protection: $\rho = 0.20$ with P_{attr} , $\rho = 0.31$ with P_{assoc} , and $\rho = 0.43$ with $P_{\text{subgroup.noaccent}}$, supporting accent as a distinct privacy dimension.

5.3. CMC Analysis: Speaker Re-Identification

Table 9 reports CMC re-identification rates (lower is better). PHORTRESS achieves the lowest rates, near the random-permutation baseline. VOXLET exhibits substantially elevated rates across ranks, indicating strong residual

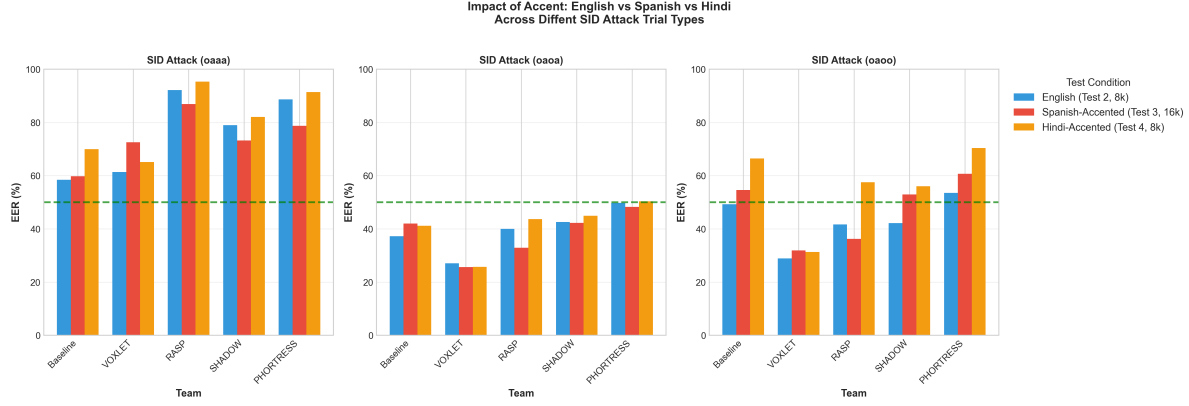


Figure 2. Impact of speaker accent on SID attack performance. Accent effects on the primary *oaoa* trial are modest, while *oaoo* and *oaoo* show larger system-specific interactions.

System	SBLS	P_{attr}	P_{assoc}	$P_{subgroup}$	Sex AUC	Age AUC	Max Reident	Most Vulnerable
Baseline	0.877±0.039	0.921±0.063	0.997±0.005	0.552±0.089	0.569±0.048	0.510±0.044	45.1±8.9%	Adult_Male
VOXLET	0.728±0.027	0.645±0.034	0.903±0.030	0.542±0.055	0.855±0.033	0.482±0.034	46.4±5.5%	Adult_Female
RASP	0.617±0.021	0.515±0.037	0.895±0.017	0.264±0.041	0.922±0.016	0.564±0.034	73.7±4.1%	Adult_Male
SHADOW	0.593±0.031	0.505±0.040	0.776±0.030	0.403±0.080	0.955±0.014	0.541±0.037	60.3±7.8%	Adult_Male
PHORTRESS	0.920±0.011	1.000±0.012	0.997±0.004	0.606±0.050	0.464±0.039	0.496±0.027	39.4±5.0%	Adult_Male
Original	0.456±0.023	0.368±0.039	0.676±0.028	0.194±0.031	0.973±0.008	0.659±0.039	81.4±3.0%	Adult_Female

Table 5. SBLS results (sex + age). Higher SBLS indicates better privacy. AUC of 0.5 corresponds to chance-level attribute prediction. Values are shown as *mean ± half-width of 95% CI*, computed by a speaker-clustered non-parametric bootstrap ($B = 1000$).

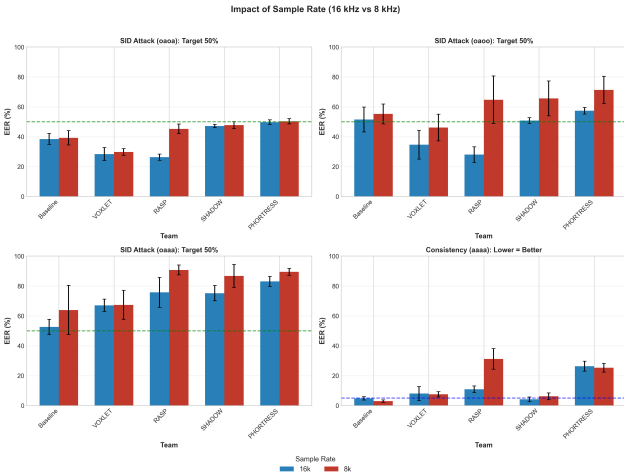


Figure 3. Impact of sample rate on speaker de-identification performance for Test 1 dataset (76 English speakers, Mixer 6).

	test1-16k	test1-8k	test2	test3	test4
Baseline	0.816±0.080	0.812±0.086	0.863±0.057	0.847±0.092	0.846±0.108
VOXLET	0.665±0.072	0.659±0.074	0.724±0.047	0.634±0.096	0.698±0.065
SHADOW	0.547±0.045	0.552±0.065	0.542±0.049	0.629±0.053	0.522±0.100
RASP	0.605±0.045	0.611±0.054	0.634±0.046	0.566±0.057	0.588±0.040
PHORTRESS	0.863±0.042	0.879±0.037	0.908±0.024	0.893±0.070	0.931±0.024

Table 6. SBLS with sex and age attributes per dataset. Values are shown as *mean ± half-width of 95% CI* from a speaker-clustered non-parametric bootstrap ($B = 1000$).

System	SBLS (accent)	SBLS (no accent)	Δ	Accent AUC
Baseline	0.857±0.029	0.877±0.039	-0.020±0.020	0.610±0.034
VOXLET	0.674±0.024	0.728±0.027	-0.054±0.018	0.775±0.026
RASP	0.652±0.021	0.617±0.021	+0.035±0.014	0.692±0.026
SHADOW	0.649±0.020	0.593±0.031	+0.056±0.016	0.710±0.019
PHORTRESS	0.911±0.009	0.920±0.011	-0.009±0.010	0.509±0.020
Original	0.453±0.022	0.456±0.023	-0.003±0.016	0.857±0.020

Table 7. SBLS with and without accent as a third attribute. Values are shown as *mean ± half-width of 95% CI* from a speaker-clustered non-parametric bootstrap ($B = 1000$); the Δ CI is paired (accent and no-accent SBLS are evaluated on the same resample each iteration).

System	default	equal	α -heavy	β -heavy	γ -heavy	no-sub	min-sub
Baseline	2	2	2	2	2	2	2
VOXLET	3	3	3	3	3	3	3
RASP	4	5	4	4	5	4	4
SHADOW	5	4	5	5	4	5	5
PHORTRESS	1	1	1	1	1	1	1
Original	6	6	6	6	6	6	6

Table 8. SBLS rankings across weight configurations.

System	k=1 (ECAPA)	k=1 (TitaNet)	k=20 (ECAPA)	k=20 (TitaNet)
Baseline	0.043	0.055	0.228	0.263
VOXLET	0.231	0.267	0.622	0.686
RASP	0.078	0.084	0.398	0.418
SHADOW	0.027	0.034	0.142	0.165
PHORTRESS	0.016	0.014	0.099	0.101

Table 9. CMC re-identification rates (lower is better privacy). CMC@k with two NeMo SID backends.

identity preservation in embedding space. TitaNet typically yields slightly higher re-identification than ECAPA-TDNN,

suggesting increased sensitivity to residual speaker cues.

System	CCA (ECAPA)	CCA (TitaNet)	Proc (ECAPA)	Proc (TitaNet)
Baseline	0.903	0.910	0.591	0.595
VOXLET	0.900	0.915	0.597	0.624
RASP	0.877	0.899	0.581	0.617
SHADOW	0.870	0.874	0.552	0.557
PHORTRESS	0.694	0.707	0.390	0.407
Random	0.734	0.734	0.291	0.306

Table 10. CCA of top 10 and Procrustes alignment results averaged across datasets and pseudo-profiles. Lower CCA and Procrustes similarity (Proc) indicate stronger decorrelation.

System	Average WER	Average Cosine Similarity
VOXLET	0.24	0.860
RASP	0.44	0.686
SHADOW	0.37	0.729
PHORTRESS	0.70	0.482

Table 11. Overall intelligibility of anonymized segments from SDID systems (lower WER and higher cosine similarity indicate better utility).

System	Test1-16k	Test1-8k	Test2	Test3	Test4
VOXLET	0.26	0.23	0.22	0.27	0.24
RASP	0.24	0.51	0.47	0.29	0.61
SHADOW	0.27	0.32	0.32	0.45	0.39
PHORTRESS	0.57	0.59	0.62	0.82	0.75

Table 12. WER by dataset/accent condition (averaged across ASR models).

5.4. CCA and Procrustes: Embedding Subspace Analysis

Table 10 summarizes embedding subspace similarity. PHORTRESS is the only system with CCA below the random baseline, indicating stronger-than-random decorrelation between original and anonymized subspaces. All other systems remain well above baseline, suggesting that substantial linear structure is preserved. Procrustes alignment is consistent with CCA trends.

5.5. Speech Intelligibility: The Privacy–Utility Trade-off

Table 11 reports overall intelligibility aggregated across datasets and ASR configurations. VOXLET achieves the best utility, while PHORTRESS exhibits the largest utility degradation. This highlights a sharp privacy–utility trade-off: the most privacy-protective system is also the least intelligible, whereas the most intelligible system offers weaker privacy on several dimensions.

6. Discussion

This study demonstrates that privacy in speaker de-identification is inherently multi-dimensional. Evaluation practices often rely on a single metric—most commonly speaker verification EER—while other signals of identity leakage remain less frequently measured. Our results show that EER, CMC, embedding subspace similarity, and SBLS each probe a distinct mechanism of identity retention: pair-

wise matching, gallery-based retrieval, global structural preservation, and demographic predictability, respectively. Notably, the weak correlations between certain SBLS components, particularly accent subgroup protection with others, suggest that accent attributes may occupy partially independent representational subspaces.

These findings carry clear methodological implications: reliance on a single metric can mischaracterize privacy risk, since a system may reduce speaker verification accuracy while preserving embedding-level structure or demographic predictability; subgroup-level analyses are essential, as aggregate performance can obscure uneven protection across demographic groups; and privacy must be assessed alongside utility, as identity suppression can degrade linguistic fidelity under distributional shifts such as non-native accents or reduced bandwidth. Notably, WER and semantic similarity capture linguistic content preservation but not perceived naturalness, speaker consistency, signal-level quality, or forensic detectability; human listening studies, speech-quality measures, and audio-forensic detectors would complement the current framework.

Finally, subspace decorrelation below random-pair baselines suggests that some anonymization strategies actively restructure representational geometry rather than merely perturbing it—a distinction future work could formalize using information-theoretic or geometric measures.

7. Conclusion

We evaluated speaker de-identification systems using five complementary metrics, showing that they capture different aspects of identity leakage and can yield divergent privacy assessments. These findings motivate multi-metric evaluation as standard practice and point toward anonymization methods that better balance privacy protection, speech utility, and broader measures of speech quality.

8. Acknowledgements

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), Anonymous Real-Time Speech (ARTS) research program, under Interagency Agreement (IAA) with NIST IARPA-20001-D250300042. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, NIST or the U.S. Government.

References

- [1] L. Chen, C. Guo, R. Wang, K. Aik Lee, and Z.-H. Ling. Any-to-any speaker attribute perturbation for asynchronous voice

- anonymization. *IEEE Transactions on Information Forensics and Security*, 20:7736–7747, 2025. [2](#)
- [2] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, and H. Na. Ecapa-tdnn embeddings for speaker diarization. In *Proc. Interspeech 2021*, pages 3560–3564, 2021. [4](#)
- [3] P. Dhar, J. Gleason, A. Roy, C. D. Castillo, and R. Chelappa. Pass: Protected attribute suppression system for mitigating bias in face recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15067–15076. IEEE, 2021. [1](#)
- [4] T. Feng, J. Lee, A. Xu, Y. Lee, T. Lertpetchpun, X. Shi, H. Wang, T. Thebaud, L. Moro-Velazquez, D. Byrd, et al. Vox-profile: A speech foundation model benchmark for characterizing diverse speaker and speech traits. *arXiv preprint arXiv:2505.14648*, 2025. [2](#), [4](#)
- [5] M. Q. Hill, C. J. Parde, C. D. Castillo, Y. I. Colón, R. Ranjan, J.-C. Chen, V. Blanz, and A. J. O’Toole. Deep convolutional neural networks in the face of caricature. *Nature Machine Intelligence*, 1:522–529, 2018. [1](#)
- [6] N. R. Koluguri, T. Park, and B. Ginsburg. Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 8102–8106. IEEE, 2022. [4](#)
- [7] A. Lawson, M. McLaren, H. Bratt, M. Graciarena, H. Franco, C. George, A. Stauffer, C. Bartels, and J. VanHout. Open language interface for voice exploitation (olive). In *Proc. Interspeech 2016*, pages 377–378, 2016. [4](#)
- [8] Y. Li, Y. Zheng, Z. Guo, Y. Wang, J. Yin, and H. Fei. Specwav-attack: Leveraging spectrogram resizing and wav2vec 2.0 for attacking anonymized speech. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025. [1](#), [2](#)
- [9] Y.-C. Lin, Y.-S. Tsai, K.-Y. Chen, H.-Y. Huang, H.-C. Chou, and H.-y. Lee. Toward fair speech technologies: A comprehensive survey of bias and fairness in speech ai. *arXiv preprint arXiv:2605.01597*, 2026. [1](#)
- [10] X. Lyu, Y. Wang, T. Zhao, and H. Liu. Fast adaptation of pretrained speaker verification system for source speaker tracking. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025. [1](#), [2](#)
- [11] C. O. Mawalim, A. Adila, and M. Unoki. Fine-tuning titanet-large model for speaker anonymization attacker systems. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025. [1](#), [2](#)
- [12] M. Panariello, N. Tomashenko, X. Wang, X. Miao, P. Champion, H. Nourtel, M. Todisco, N. Evans, E. Vincent, and J. Yamagishi. The voiceprivacy 2022 challenge: Progress and perspectives in voice anonymisation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3477–3491, 2024. [1](#), [2](#)
- [13] C. J. Parde, C. Castillo, M. Q. Hill, Y. I. Colon, S. Sankaranarayanan, J.-C. Chen, and A. J. O’Toole. Face and image representation in deep cnn features. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 673–680, 2017. [1](#)
- [14] P. J. Phillips and D. White. The state of modelling face processing in humans with deep learning. *British Journal of Psychology*, 2025. [1](#)
- [15] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. [4](#)
- [16] C. Rusti, A. Leschanowsky, C. Quinlan, M. Pnacekova, L. Gorce, and W. T. Hutiri. Benchmark dataset dynamics, bias and privacy challenges in voice biometrics research. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2023. [1](#), [2](#)
- [17] N. Ryant. Linguistic Data Consortium Broad Phonetic Class Speech Activity Detector (ldc-bpcsad), 2023. [2](#)
- [18] M. Sekoyan, N. R. Koluguri, N. Tadevosyan, P. Zelasko, T. Bartley, N. Karpov, J. Balam, and B. Ginsburg. Canary-1b-v2 & parakeet-tdt-0.6 b-v3: Efficient and high-performance models for multilingual asr and ast. *arXiv preprint arXiv:2509.14128*, 2025. [4](#)
- [19] S. Seo, O. Aulov, and P. J. Phillips. Measuring soft biometric leakage in speaker de-identification systems. *arXiv preprint arXiv:2509.14469*, 2025. [2](#), [4](#)
- [20] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu. Mpnnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867, 2020. [5](#)
- [21] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and M. Todisco. The voiceprivacy 2024 challenge evaluation plan. *arXiv preprint arXiv:2404.02677*, 2024. [1](#), [2](#)
- [22] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi. The first voiceprivacy attacker challenge evaluation plan. *arXiv preprint arXiv:2410.07428*, 2024. [1](#), [2](#)
- [23] N. Tomashenko, E. Vincent, and M. Tommasi. Analysis of speech temporal dynamics in the context of speaker verification and voice anonymization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. [2](#)
- [24] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O’Brien, et al. The voiceprivacy 2020 challenge: Results and findings. *Computer Speech & Language*, 74:101362, 2022. [1](#), [2](#)
- [25] J. Villalba, B. J. Borgstrom, S. Kataria, M. Rybicka, C. D. Castillo, J. Cho, L. P. García-Perera, P. A. Torres-Carrasquillo, and N. Dehak. Advances in cross-lingual and cross-source audio-visual speaker recognition: The jhu-mit system for nist sre21. In *Odyssey*, pages 213–220, 2022. [4](#)
- [26] H. L. Xinyuan, A. Garg, Z. Cai, K. Duh, L. P. García-Perera, S. Khudanpur, N. Andrews, and M. Wiesner. Hltcoe submission to the voiceprivacy attacker challenge. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025. [1](#), [2](#)
- [27] Y. Yang, T. Thebaud, and N. Dehak. Demographic attributes prediction from speech using wavlm embeddings. In *2025*

59th Annual Conference on Information Sciences and Systems (CISS). IEEE, 2025. [1](#), [2](#)

- [28] Y. Zhang, Z. Bi, F. Xiao, X. Yang, Q. Zhu, and J. Guan. Attacking voice anonymization systems with augmented feature and speaker identity difference. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025. [1](#), [2](#)