Exploiting Acoustic and Content-Oriented Speaker Verification Attacks Against Multilingual Voice Anonymization

Ridwan Arefeen*, Ze Li†, Rong Tong*, Ming Li‡, Xiaoxiao Miao§
*Singapore Institute of Technology, Singapore
†Wuhan University, China
‡The Chinese University of Hong Kong, China
§Duke Kunshan University, China

***Abstract*—Attacker ASV systems for voice anonymization have been studied primarily in English, leaving their behavior in multilingual settings largely unexplored. Conventional ASV has shown that both acoustic and contextual information are important for multilingual speaker verification. Inspired by this, we investigate whether the same holds for attacker ASV on anonymized speech. We evaluate both acoustic- and content-oriented attackers on multilingual anonymized speech and construct a multilingual voice-converted dataset to improve cross-lingual generalization. Our results show that attacker effectiveness depends on the linguistic utility of the anonymized speech. Overall, acoustic-oriented attackers achieve better performance. However, when linguistic information is well preserved, the performance gap between content- and acoustic-oriented attackers narrows compared with conditions involving stronger speech distortion. The multilingual voice-converted dataset further improves performance and partially reduces the cross-lingual gap. These findings highlight the need for more comprehensive attacker modeling and evaluation protocols that consider both privacy and utility, rather than relying on a attacker strategy[1].**

***Index Terms*—speaker anonymization, automatic speaker verification, cross-lingual generalization, voice conversion**

I. Introduction

Speech anonymization has emerged as an important research direction in privacy-preserving speech processing, as speech recordings reveal not only linguistic content but also speaker identity and other sensitive attributes. A central goal is to remove or obfuscate identity-related information without destroying linguistic intelligibility or other utility cues [1], [2]. The privacy protection of an anonymization system is conventionally measured by how well an attacker's automatic speaker verification (ASV) system can re-identify the original speaker, typically reported as the equal error rate (EER). However, existing work [3], [4] has shown that evaluating against an ASV trained on real speech is insufficient. A stronger, *semi-informed* attacker instead trains its ASV model on speech anonymized by the same or similar anonymization system, so that it can learn to exploit the residual source-speaker information left by that anonymization. Beyond this, several works further strengthen the attacker, mainly from three distinct perspectives: (i) training on large-scale anonymized or voice-converted data to trace the source speaker [5], [6]; (ii) exploiting residual source-speaker information from other features, such as linguistic content, duration, and prosody [7]–[10]; and (iii) improving the ASV architecture, including self-supervised models [11], [12].

Among these developments, most ASV attacker systems have been developed and assessed largely on English anonymized speech, with little extension to multilingual settings. This is mainly because existing anonymization studies and challenge benchmarks have emphasized English [2], [3], [13]–[15]. The latest VoicePrivacy 2026 Challenge (VPC 2026) [16] introduces multilingual anonymization (English, German, French, and Spanish), evaluated with a single attacker model shared across all languages. Two main questions, however, remain unexplored. First, the evaluated languages are all Indo-European, so the reported robustness may not reflect how the attacker behaves on structurally distant languages. Second, the challenge results reveal a striking but unexplained pattern: English consistently yields the lowest EER, and the word error rate (WER) of the anonymized speech varies considerably across languages, yet what underlying factor drives this gap remains unclear.

This work thus takes a step further to investigate the bottleneck of the shared attacker across languages. To broaden linguistic diversity beyond the Indo-European family, we deliberately add two structurally distant East Asian languages, Japanese and Chinese, and extend our evaluation to the cross-lingual setting of Track 2 of the VPC 2026 challenge. To probe what actually drives the cross-lingual gap, we compare two state-of-the-art but fundamentally different attacker paradigms: a WavLM–ECAPA-based acoustic-oriented attacker [5], the same model used in the VPC 2026 challenge, and a w2v-BERT-based content-oriented attacker with multi-scale feature aggregation (MFA) [17][2].

2403754@sit.singaporetech.edu.sg, tong.rong@singaporetech.edu.sg, xiaoxiao.miao@dukekunshan.edu.cn.
Corresponding author: Xiaoxiao Miao.

[1]Full code and pretrained models and MultiVC Dataset link are available at `https://github.com/monkeyDarefeen/DAST`.

[2]We refer to the former as acoustic-oriented and the latter as content-oriented based on whether the ASV architecture contains an explicit content-oriented component, i.e., the BERT component.

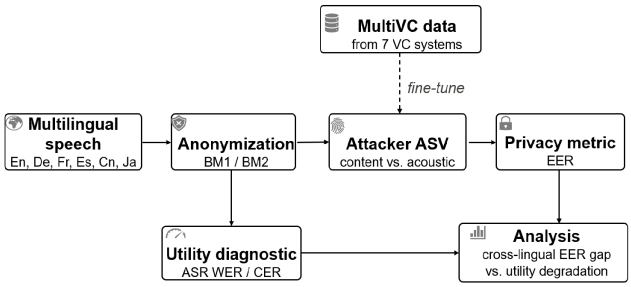


Fig. 1. Overview of the proposed evaluation.

We first find that both attackers re-identify speakers reliably on the European languages but degrade sharply on the East Asian ones, and that this drop tracks rising automatic speech recognition (ASR) WER, revealing that re-identification degrades in step with ASR utility and suggesting a coupling between linguistic intelligibility and identity disclosure. Comparing the two paradigms under these multilingual conditions, we find the acoustic-oriented attackers are generally more robust than content-oriented ones.

These observations lead to our central hypothesis: the residual identity cues that survive anonymization depend on how well acoustic naturalness and linguistic intelligibility are preserved. If so, exposing an attacker ASV system to diverse, multi-speaker VC data during training should broaden its representations, helping it separate speaker-invariant traits from system-specific anonymization artifacts and narrowing any gap caused by data scarcity rather than by language structure itself. To test this, we construct a large-scale multilingual voice-converted corpus using seven distinct voice-conversion (VC) systems and use it to fine-tune both attackers, yielding domain-adapted variants that partially close the cross-lingual gap. This shows that the degradation is only partly data-driven, adaptation narrows the gap but a residual portion remains intrinsic to the language itself, which again highlights the complexity of cross-lingual privacy leakage. These findings highlight the need for more comprehensive attacker modeling that considers both privacy and utility, rather than relying on a single attacker strategy.

II. Attacker ASV Modeling and Acoustic-Linguistic Vulnerability Analysis

Figure 1 gives an overview of our evaluation. Multilingual speech in six languages[3] is first anonymized by one of two VPC 2026 baselines, BM1 [18] or BM2 [19]. The anonymized speech is then presented to two attacker ASV systems built on fundamentally different paradigms: a WavLM–ECAPA-based acoustic-oriented model [5] and a w2v-BERT-MFA-based content-oriented model [17] to quantify re-identification capability, measured by EER. In parallel, we transcribe the anonymized speech and compute ASR WER/CER as a utility diagnostic, capturing how much intelligibility each pipeline preserves. Our analysis then relates these two axes: the cross-lingual EER gap and utility degradation, to determine whether the drop in re-identification across languages is driven by data scarcity or is intrinsic to the language. To probe the data-scarcity side, we build a large-scale multilingual voice-converted corpus (denoted as MultiVC) generated from seven VC systems) and use it as adaptation data to fine-tune both attackers.

A. Training and Evaluation Data

We perform multilingual attacker system analysis on the official training, development, and test sets of the VPC 2026 Track 2 multilingual anonymization task. For brevity, we report only the test set results in this paper.

To cover language diversity, we construct two auxiliary language datasets. For Japanese, we use Mozilla Common Voice [20] for training (1,545 speakers, 19,015 utterances) and JVS [21] for evaluation (48 speakers, 125,360 trials). For Mandarin Chinese, we use AISHELL-3 [22] for both training (340 speakers, 120,418 utterances) and evaluation (30 speakers, 150,648 trials).

B. Anonymization Baselines

We take two multilingual anonymization baselines from VPC 2026:

BM1 (HuBERT-based Framework) [18]: The original speech is decomposed into speaker embeddings, pitch features, and language-independent content representations extracted by a self-supervised pretrained HuBERT model. The speaker embeddings are replaced with those of a target pseudo-speaker, and the pitch and content representations are combined to synthesize the anonymized waveform using the HiFi-GAN neural vocoder.

BM2 (Whisper/Toucan-based Framework) [19]: This architecture extracts speaker embeddings, transcriptions using the pre-trained Whisper model, as well as F0, energy, and phoneme durations. The original speaker embedding is replaced by an artificial embedding generated via a Wasserstein GAN. F0 and energy are randomly perturbed at the phoneme level, and the anonymized speech is synthesized using the IMS Toucan multilingual TTS system[4].

C. Threat Models

We evaluate voice anonymization systems under two attack scenarios. In both cases, the attacker has access to the anonymization system, multiple enrollment utterances for each speaker, and language-matched public speech data, which can be anonymized using the same system to train or adapt the attacker. The enrollment utterances are also anonymized using the same system to match the anonymized trial utterances generated by the user.

Lazy-informed attacker: The ASV model is not adapted to the corresponding anonymized evaluation data. Instead, it is

[3]English (En), German (De), French (Fr), Spanish (Es), Chinese (Cn), Japanese (Ja)

[4]VPC2026 also includes BM3 in Track 2. Since BM3 follows the same architecture as BM2 and does not incorporate prosody extraction or F0/energy modification, we do not investigate it separately.

trained on either original speech or synthetic speech generated by different generative models, as long as these are not the same as those used in the evaluation set.

Semi-informed attacker: The ASV model is adapted using anonymized speech generated by the same anonymization system as the evaluation data, resulting in a stronger attacker.

Re-identification capability of the attacker is quantified by the EER, where a lower EER indicates a stronger attacker. Note that VPC2026 focuses primarily on the semi-informed attacker, but in our experiments we consider both attacker settings. The reason is that we aim to investigate a general attacker capable of good generalization across different anonymization approaches, which can be assessed via the lazy-informed EER, while the semi-informed EER further reflects the upper-bound attack strength when the attacker is adapted to the same anonymization system.

D. Attacker Architectures

To explore the trade-offs between content-invariant matching and acoustic feature mining under cross-lingual conditions, we evaluate two fundamentally distinct self-supervised models:

Acoustic-oriented (WavLM-ECAPA): This model features a dual-stream architecture that processes raw waveforms using a frozen WavLM [23] model and spectral features via Mel-filterbanks, with parallel ECAPA-TDNN encoders for mid-level feature fusion. The model is optimized using a three-stage training strategy consisting of speaker pre-training on original speech, diversity-driven domain training on English voice-converted data from the Source Speaker Tracing Challenge (SSTC) [24] (lazy-informed), and lightweight task-specific fine-tuning on target anonymized data (semi-informed). This is the standard ASV attacker system adopted in VPC 2026.

Content-oriented (w2v-BERT-MFA): Built upon a 24-layer Conformer [25] architecture, this framework extracts intermediate block-level representations using localized layer adapters and a Multi-scale Feature Aggregation (MFA) module. Pre-trained on a large-scale 4.5-million-hour unlabeled audio corpus via joint contrastive learning and masked language modeling, it is designed to capture text-invariant semantic representations. Parameter-efficient fine-tuning is performed using Low-Rank Adaptation (LoRA) [26] inserted into attention sublayers. To ensure architectural parity during benchmarking, the pretrained w2v-BERT-MFA model[5] is first fine-tuned with LoRA adapters [27] on the SSTC [24] dataset, serving as the lazy-informed attacker. It is then further fine-tuned on the target anonymized data, serving as the semi-informed attacker.

E. Diagnostic Vulnerability Analysis

Figure 2 presents our first diagnostic view. For each of the six languages, it reports the average lazy-informed EER of the two attackers: content-oriented (w2v-BERT-MFA) and acoustic-oriented (WavLM-ECAPA), against the two anonymization baselines (BM1 and BM2) on multilingual anonymized evaluation sets described in Section II-A, with the per-language ASR WER/CER overlaid (with star/diamond). Presenting privacy leakage (EER) and utility loss (WER/CER) on a shared language axis makes it easy to see whether the two move together across languages. The trends in Fig 2 reveal a cross-lingual generalization bottleneck: while both models reliably recover identities within Indo-European cohorts, their re-identification degrades sharply on the East Asian languages (Chinese and Japanese).

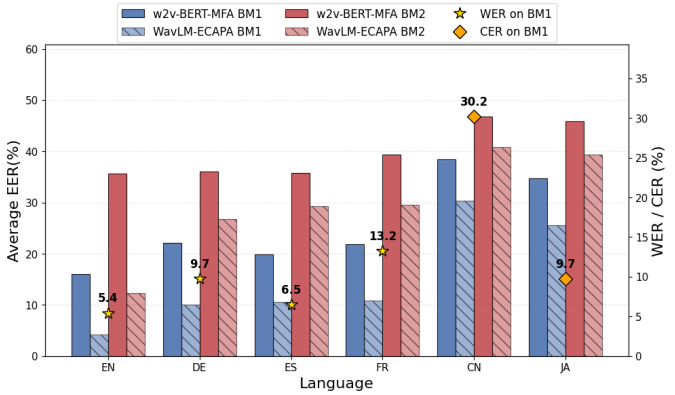


Fig. 2. Average lazy-informed EER of the content-oriented (w2v-BERT-MFA) and acoustic-oriented (WavLM-ECAPA) attackers across six languages, under the lazy-informed threat model on the BM1 and BM2 baselines. Per-language ASR WER/CER is overlaid (right axis).

To examine whether this non-uniform performance is data-driven or tied to linguistic degradation, we conduct a utility analysis, computing WER for the European languages and CER for Chinese and Japanese (where word segmentation is ill-defined) with a Whisper-Large-v3 [28] model transcribing the synthesized audio. The results are reported in Table 1.

Table I shows two trends. First, a *baseline synthesis discrepancy*: BM2 yields higher error metrics than BM1 across all languages (e.g., French WER rises from 13.2% to 44.9%), indicating that explicit prosodic and F0 extraction imposes more distortion than unified self-supervised embeddings. Second, a *typological floor*: this degradation is far larger for logographic and syllabic languages: under BM2, Japanese reaches a CER of 45.1% and Chinese a CER of 70.3%.

The parallel between utility degradation (Table I) and re-identification failure (Fig. 2) points to an acoustic–linguistic coupling. We read this as a coupling rather than a demonstrated causal link: the non-linear transformations of these anonymization systems appear to corrupt speech intelligibility, so when a language's phonetic structure is degraded, the localized identity-bearing cues are diluted, raising the effective difficulty for attackers.

III. Multilingual Voice-Converted Dataset Generation

To evaluate whether the cross-lingual performance can be improved by removing data scarcity, we construct a large-scale, multilingual voice-converted corpus. This section details

[5]https://github.com/ZXHY-82/LI-MSV-TidyVoice2026

TABLE I
ASR UTILITY METRICS ACROSS BM1 AND BM2 BASELINES. CER(%) IS REPORTED FOR JAPANESE AND CHINESE, WER(%) FOR THE REST.

Language	BM1	BM2
English (WER)	5.4	16.0
German (WER)	9.7	24.2
Spanish (WER)	6.5	25.6
French (WER)	13.2	44.9
Japanese (CER)	9.7	45.1
Chinese (CER)	30.2	70.3

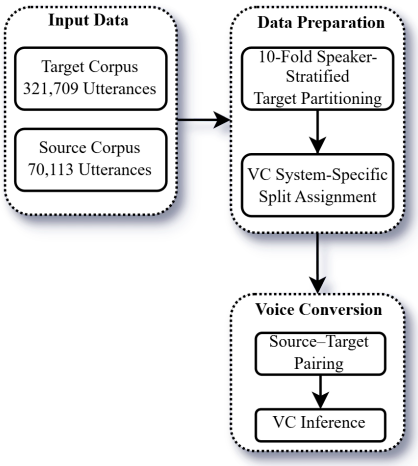


Fig. 3. Voice Converted Dataset Pipeline

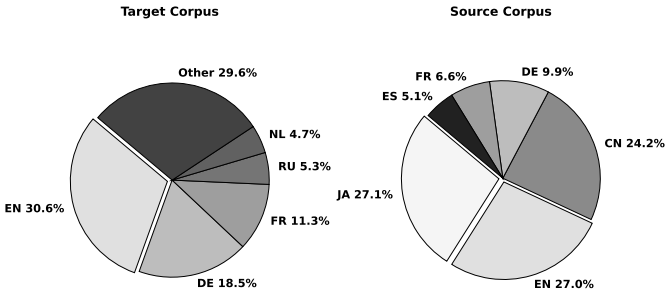


Fig. 4. Language distribution of the evaluation corpora. (Left) Target corpus distribution (TidyVoiceX), where the top 5 languages comprise roughly 70% of the 321,709 utterances. (Right) Source corpus composition, where 70,113 utterances serve as the content reference pool for all seven voice conversion systems.

TABLE II
VOICE CONVERSION SYSTEMS AND THEIR ASSIGNED TARGET SPLITS. EACH SYSTEM CONVERTS 34,022 TARGET UTTERANCES USING 70,113 SOURCE UTTERANCES, PRODUCING 102,066 CONVERTED SAMPLES PER SYSTEM.

ID	Voice conversion system	Target fold
VC-1	CosyVoice2 [30]	Split 1
VC-2	Seed-VC [31]	Split 2
VC-3	YingMusic-SVC [32]	Split 3
VC-4	FreeVC [33]	Split 4
VC-5	KDVC [34]	Split 5
VC-6	LVC-VC [35]	Split 6
VC-7	TriAAN-VC [36]	Split 7
Converted utterances per system		**102,066**
Total converted utterances		**714,462**

the source and target speech distributions, our speaker-oriented partition mechanics, and the deployment of seven advanced voice conversion architectures used to synthesize our evaluation. Fig. 3 shows the overall pipeline for the voice converted dataset generation.

A. Source Corpus

We construct an expanded multilingual source pool comprising same utterances across all targeted languages (see Fig 4 Right) as from training set mention in section II-A. These source utterances serve as the references for our downstream voice conversion pipeline.

B. Target Corpus

For the target speaker pool, we employ TidyVoiceX from [29], a diverse multilingual speech corpus containing 321,709 utterances from 4,474 unique speakers across 40 distinct languages. The structural distribution of the top 5 languages (English, German, French, Russian, Dutch) in this target corpus is outlined in (see Fig 4 Left). This corpus provides the necessary acoustic and demographic variance required to create target voice profiles.

C. Split Strategy

We partition the TidyVoiceX corpus into 10 balanced folds using a per-speaker round-robin distribution strategy. For each speaker, utterances are assigned sequentially to splits 1 through 10 in cyclic order. When a speaker's total utterance count is not a perfect multiple of 10, the surplus utterances are uniformly distributed via duplication into the remaining splits. This ensures that all 10 folds contain an identical volume of exactly 34,022 utterances while strictly guaranteeing speaker-level stratification; every individual split contains representations from all 4,474 speakers, preventing any speaker profile from appearing exclusively within a single fold.

D. Voice Conversion Systems

We utilize seven VC architectures, assigning each to a distinct target split to enforce target-disjoint evaluation and prevent identity cross-contamination. As detailed in Table II, each architecture processes a fixed allocation of target utterances against the global source pool, yielding a total generated dataset of 714,462 converted samples. Splits 8 through 10 are held in reserve as baseline control sets for future ablation tracking.

E. Pair Generation

To synthesize pairs for each VC system, we map each of the 34,022 target style references to three distinct source content utterances, yielding $34{,}022 \times 3 = 102{,}066$ conversion utterances per system. The pairing framework follows a

TABLE III
EER (%) ON BM1 AND BM2 ACROSS SIX LANGUAGES. LOWER EER INDICATES A STRONGER ATTACK. BEST RESULTS ARE BOLDED INDEPENDENTLY FOR EACH BLOCK.

Threat Model	ID	ASV Model	Training Data	En	De	Fr	Es	Cn	Ja	Avg.
Lazy-BM1	1	w2v-BERT-MFA	+SSTC	16.76	22.20	22.73	18.84	39.76	34.86	25.86
	2	w2v-BERT-MFA	+SSTC + MultiVC	11.108	15.336	13.141	10.565	28.160	27.670	17.66
	3	WavLM-ECAPA	+SSTC	5.52	11.57	11.93	8.23	31.25	26.13	15.77
	4	WavLM-ECAPA	+SSTC + MultiVC	**2.887**	**4.526**	**2.520**	**1.636**	**17.727**	**19.622**	**8.99**
Lazy-BM2	5	w2v-BERT-MFA	+SSTC	35.80	34.90	39.82	34.46	46.86	45.85	39.61
	6	w2v-BERT-MFA	+SSTC + MultiVC	35.708	34.507	38.948	33.972	46.069	44.602	38.97
	7	WavLM-ECAPA	+SSTC	13.27	26.79	30.57	27.23	41.10	39.61	29.76
	8	WavLM-ECAPA	+SSTC + MultiVC	**11.017**	**21.053**	**24.781**	**18.443**	**37.884**	**35.814**	**24.83**
Semi-BM1	9	w2v-BERT-MFA	+SSTC + BM1	4.142	5.684	6.086	4.313	14.364	15.206	8.30
	10	w2v-BERT-MFA	+SSTC + MultiVC + BM1	3.991	5.409	5.463	4.210	14.592	14.867	8.09
	11	WavLM-ECAPA	+SSTC + BM1	2.887	1.081	1.091	1.217	7.875	11.496	4.27
	12	WavLM-ECAPA	+SSTC + MultiVC + BM1	**2.786**	**0.914**	**1.071**	**0.992**	**6.285**	**9.910**	**3.66**
Semi-BM2	13	w2v-BERT-MFA	+SSTC + BM2	48.131	32.206	38.121	31.482	46.279	44.034	40.04
	14	w2v-BERT-MFA	+SSTC + MultiVC + BM2	26.270	30.604	35.433	29.207	45.141	42.633	34.88
	15	WavLM-ECAPA	+SSTC + BM2	7.185	10.561	13.389	7.139	28.992	28.994	16.04
	16	WavLM-ECAPA	+SSTC + MultiVC + BM2	**7.013**	**9.899**	**12.610**	**6.650**	**26.786**	**28.409**	**15.23**

two-phase allocation strategy engineered to maximize global source coverage:

1) **Guaranteed Coverage (Phase 1):** The entire pool of 70,113 source utterances (Table II) is randomized and assigned to the initial 70,113 pair slots, ensuring every source language token acts as a content reference at least once.
2) **Stochastic Backfill (Phase 2):** The remaining $102{,}066 - 70{,}113 = 31{,}953$ slots are populated via uniform random sampling from the source pool with replacement.

This strategy ensures that the entire source corpus is represented in every conversion set. Across systems, the average source usage ranges from $1.35\times$ to $1.46\times$, with a maximum cap of 6 uses per source utterance. The distribution of source usage follows a long-tail pattern: approximately 70% of sources are utilized once, 20% twice, and the remainder three or more times.

IV. EXPERIMENT

We first evaluate lazy-informed attackers on multilingual test sets. Since the training data combine multiple languages, we expect improved cross-lingual generalization ability against different anonymization systems. This setting allows us to study the robustness of both acoustic-oriented (WavLM-ECAPA) and content-oriented (w2v-BERT-MFA) ASV models under different synthetic training conditions, including monolingual (SSTC) and multilingual (MultiVC) data.

We then evaluate semi-informed attackers, where the lazy-informed models are further fine-tuned on the target anonymization baselines, i.e., BM1 or BM2. We further analyze the bottlenecks of the multilingual attacker system.

A. Training Setup for the Attacker Models

As introduced in Section II-C and Section II-D, we consider two attacker scenarios, i.e., lazy-informed and semi-informed, under two ASV architectures, i.e., WavLM-ECAPA and w2v-BERT-MFA.

During lazy-informed attacker training, WavLM-ECAPA follows the training strategy in [5]. The content-oriented w2v-BERT-MFA framework is fine-tuned using LoRA, following the architectural constraints in [27]. Both architectures are trained on SSTC (English voice conversion data) or MultiVC, where training data from multiple languages are combined to improve generalization.

During semi-informed attacker training, we follow the official VPC 2026 ASV fine-tuning script to optimize both ASV architectures under target anonymization data, i.e., BM1 and BM2. The models are further fine-tuned on the corresponding target dataset for each scenario.

B. Experimental Results

1) Results on Lazy-informed Attackers: We first analyze the lazy-informed attacker results, where the top portion of Table III reports the EER on multilingual test sets under both BM1 and BM2 anonymization settings.

From the training data perspective, incorporating multilingual data consistently improves performance. For BM1 (ID1–ID4), the average EER is reduced from 25.86% to 17.66% for w2v-BERT-MFA, and from 15.77% to 8.99% for WavLM-ECAPA. Similar trends are observed on BM2 (ID5–ID8), where the average EER decreases from 39.61% to 38.97% for w2v-BERT-MFA, and from 29.76% to 24.83% for WavLM-ECAPA. These results demonstrate that multilingual training data provides consistent but moderate gains across both architectures.

However, the improvement brought by multilingual training is substantially smaller on BM2 compared to BM1. This suggests that stronger anonymization significantly limits the benefit of cross-lingual generalization. A possible explanation is that BM2 employs ASR–TTS-based speech generation,

which introduces stronger prosodic distortion and partial semantic degradation (see WER/CER results on BM2 in table I). As a result, voice conversion-based training data (e.g., SSTC, MultiVC) becomes less effective in preserving or recovering speaker-discriminative cues. This hypothesis is further investigated in the semi-informed evaluation.

Comparing ASV architectures, WavLM-ECAPA consistently outperforms w2v-BERT-MFA across both BM1 (ID1–2 vs. ID3–4) and BM2 (ID5–6 vs. ID7–8). This indicates that acoustic-oriented representations are more robust under anonymized conditions, whereas content-oriented models such as w2v-BERT-MFA rely more heavily on semantic consistency. When anonymization distorts semantic information, the effectiveness of content-based speaker modeling is reduced, making acoustic representations more reliable for attacker construction.

2) Results on Semi-informed Attackers: We further evaluate semi-informed attackers by fine-tuning each lazy-informed model on anonymized speech generated by the corresponding target anonymization system. The results are reported in the lower part of Table III.

Unlike in the lazy-informed setting, the relative effectiveness of different attackers depends on the utility of the anonymized speech. On BM1, WavLM-ECAPA with MultiVC + SSTC + BM1 (ID12) achieves the best performance, yielding the lowest average EER of 3.66%. WavLM-ECAPA also outperforms w2v-BERT-MFA on BM2 (ID15–16 vs. ID13–14), with the best result obtained using MultiVC augmentation (ID16, 15.23% average EER). However, the performance gap between the two attacker types differs substantially across BM1 and BM2.

This difference is consistent with the utility results in Table I. BM1 achieves lower WER/CER across all languages, indicating better preservation of linguistic information. Under this condition, the content-oriented attacker performs more closely to the acoustic-oriented attacker. In contrast, BM2 introduces stronger speech distortion, which degrades the linguistic information available to the content-oriented attacker and widens the performance gap in favor of WavLM-ECAPA. MultiVC augmentation further improves attack performance in most settings.

Overall, acoustic-oriented attackers achieve better performance across both anonymization settings. However, the relative effectiveness of content-oriented attacks depends more strongly on speech utility: when linguistic information is well preserved, their performance approaches that of acoustic-oriented attacks, whereas stronger distortion increases the advantage of acoustic-oriented attacks.

V. Conclusion

This study investigated the bottleneck of cross-lingual generalization for attacker ASV systems against voice anonymization. To determine whether the performance gap across languages is driven by data scarcity or linguistic intelligibility, we compared an acoustic-oriented dual-stream model (WavLM-ECAPA) with a content-oriented unified semantic model (w2v-BERT-MFA). Our experiments establish a clear coupling between speech utility (measured by WER/CER) and re-identification vulnerability; as well in semi-informed scenario, content-oriented models (w2v-BERT-MFA) gives competitive results when semantic structure is preserved, but under heavily distorted, high-WER conditions, acoustic-oriented architectures (WavLM-ECAPA) demonstrate far superior robustness. This indicate that the optimal attackers outcome depends on how well the target anonymization system preserves semantic information. Furthermore, while fine-tuning on our constructed multi-speaker voice-converted corpus (MultiVC) mitigates data scarcity and partially narrows the cross-lingual EER gap, a residual, intelligibility-bound gap inherently persists. Consequently, these observations highlight the need for multifaceted attacker modeling and evaluation protocol that can weigh privacy preservation againts utility degradation.

VI. Acknowledgments

This work is supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant R-R13-A405-0005.

VII. Generative AI Use Disclosure

Large language models (LLMs) were used in a limited capacity as auxiliary tools to assist with the language polishing of the manuscript. All research ideas, methodological design, experiments, analyses, and final writing decisions were conducted and verified by the authors.

References

[1] A. Das, C. Franzreb, T. Herzig, P. Pirlet, and T. Polzehl, "Comparing speech anonymization efficacy by voice conversion using knn and disentangled speaker feature representations," in *Proceedings of the Symposium on Security and Privacy in Speech Communication (SPSC 2024)*. ISCA, 2024.

[2] 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*, vol. 32, pp. 3477–3491, 2024.

[3] 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*, vol. 74, p. 101362, 2022.

[4] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the VoicePrivacy Initiative," in *Proc. Interspeech*, 2020, pp. 1693–1697.

[5] R. Arefeen, X. Miao, R. Tong, A. B. Ng, S. See, and T. Liu, "Dast: A dual-stream voice anonymization attacker with staged training," 2026. [Online]. Available: https://arxiv.org/abs/2603.12840

[6] C. O. Mawalim, A. Adila, and M. Unoki, "Fine-tuning titanet-large model for speaker anonymization attacker systems," in *Proc. ICASSP*, 2025, pp. 1–2.

[7] C. Franzreb, A. Das, T. Polzehl, and S. Möller, "Content leakage in librispeech and its impact on the privacy evaluation of speaker anonymization," *arXiv preprint arXiv:2601.13107*, 2026.

[8] Ünal Ege Gaznepoglu, A. Leschanowsky, A. Aloradi, P. Singh, D. Tenbrinck, E. A. P. Habets, and N. Peters, "You are what you say: Exploiting linguistic content for voiceprivacy attacks," 2026. [Online]. Available: https://arxiv.org/abs/2506.09521

[9] N. Tomashenko, E. Vincent, and M. Tommasi, "Exploiting Context-dependent Duration Features for Voice Anonymization Attack Systems," in *Interspeech 2025*, 2025, pp. 5128–5132.

[10] R. Arefeen, X. Miao, R. Tong, A. B. Ng, and S. See, "Segreconcat: A data augmentation method for voice anonymization attack," in *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2025, pp. 2080–2085.

[11] X. Lyu, Y. Wang, T. Zhao, and H. Liu, "Fast adaptation of pretrained speaker verification system for source speaker tracking," in *Proc. ICASSP*, 2025, pp. 1–2.

[12] 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 *Proc. ICASSP*, 2025, pp. 1–2.

[13] N. Tomashenko, X. Miao, P. Champion, S. Meyer, M. Panariello, X. Wang, N. Evans, E. Vincent, J. Yamagishi, and M. Todisco, "The third voiceprivacy challenge: Preserving emotional expressiveness and linguistic content in voice anonymization," *Computer Speech & Language*, vol. 100, p. 101988, 2026.

[14] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi, "The First VoicePrivacy Attacker Challenge evaluation plan," *arXiv preprint arXiv:2410.07428*, 2024.

[15] ——, "Privacy attacks on voice anonymization systems: Overview and key findings from the first voiceprivacy attacker challenge," 2026.

[16] X. Miao, N. Tomashenko, R. Arefeen, S. Meyer, M. Panariello, X. Wang, E. Vincent, N. Evans, J. Yamagishi, and M. Todisco, "The voiceprivacy 2026 challenge evaluation plan," 2026. [Online]. Available: https://www.voiceprivacychallenge.org/vp2026/docs/VPC_2026_may12.pdf

[17] Z. Li, X. Miao, J. Liu, and M. Li, "Language-invariant multilingual speaker verification for the tidyvoice 2026 challenge," *ICASSP*, 2026.

[18] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, "Language-independent speaker anonymization approach using self-supervised pre-trained models," *The Speaker and Language Recognition Workshop (Odyssey)*, 2022.

[19] S. Meyer, F. Lux, and N. T. Vu, "Probing the Feasibility of Multilingual Speaker Anonymization," in *Interspeech 2024*, 2024, pp. 4448–4452.

[20] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the twelfth language resources and evaluation conference*, 2020, pp. 4218–4222.

[21] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "Jvs corpus: free japanese multi-speaker voice corpus," *arXiv preprint arXiv:1908.06248*, 2019.

[22] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A Multi-Speaker Mandarin TTS Corpus," in *Interspeech 2021*, 2021, pp. 2756–2760.

[23] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[24] Z. Li, Y. Lin, T. Yao, H. Suo, P. Zhang, Y. Ren, Z. Cai, H. Nishizaki, and M. Li, "The database and benchmark for the source speaker tracing challenge 2024," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 1254–1261.

[25] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[26] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *Proc. ICLR*, vol. 1, no. 2, p. 3, 2022.

[27] Z. Li, M. Cheng, and M. Li, "Enhancing speaker verification with w2v-bert 2.0 and knowledge distillation guided structured pruning," in *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026, pp. 16 462–16 466.

[28] 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*. PMLR, 2023, pp. 28 492–28 518.

[29] A. Farhadipour, J. Marquenie, S. Madikeri, T. Vukovic, V. Dellwo, K. Reid, F. M. Tyers, I. Siegert, and E. Chodroff, "Tidyvoice 2026 challenge evaluation plan," *arXiv preprint arXiv:2601.21960*, 2026.

[30] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang *et al.*, "Cosyvoice 2: Scalable streaming speech synthesis with large language models," *arXiv preprint arXiv:2412.10117*, 2024.

[31] S. Liu, "Zero-shot voice conversion with diffusion transformers," 2024. [Online]. Available: https://arxiv.org/abs/2411.09943

[32] G. Chen, X. Zhang, Z. Weng, J. Zheng, D. Shen, C. Ding, W.-Q. Zhang, and Z. Chen, "Yingmusic-svc: Real-world robust zero-shot singing voice conversion with flow-grpo and singing-specific inductive biases," *arXiv preprint arXiv:2512.04793*, 2025.

[33] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[34] H. T. Tu, L. Thanh Long, V. Huan, N. T. Phuong Thao, N. Van Thang, N. T. Cuong, and N. Thi Thu Trang, "Voice conversion for low-resource languages via knowledge transfer and domain-adversarial training," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[35] W. Kang, M. Hasegawa-Johnson, and D. Roy, "End-to-end zero-shot voice conversion with location-variable convolutions," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2023, 2023, pp. 2303–2307.

[36] H. J. Park, S. Woo Yang, J. S. Kim, W. Shin, and S. W. Han, "Triaan-vc: Triple adaptive attention normalization for any-to-any voice conversion," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.