

Radboud-Aalto Submission to the Voice Privacy Challenge 2026

Jiusi Zheng^{1,2}, Martha Larson¹, Tom Bäckström²

¹ Radboud University, Netherlands

² Aalto University, Finland

jiusi.zheng@ru.nl, martha.larson@ru.nl, tom.backstrom@aalto.fi

Abstract

In this work, we present a simple but effective emotion-aware voice anonymization system for the English anonymization track (Track 1) of the Voice Privacy Challenge 2026. We employ the predicted emotional state as an auxiliary cue and dynamically adjust the anonymization strength accordingly. Specifically, utterances classified as neutral are anonymized using the cascaded ASR-TTS pipeline, whereas utterances classified as non-neutral emotion are processed with reduced anonymization levels to better preserve emotion information. Our system achieves a UAR of 64.69% on the IEMOCAP-test and 63.08% on the IEMOCAP-dev. It also achieves an EER of 33.61% and 36.41% on the LibriSpeech test and development sets, outperforming the official baselines.

Index Terms: voice anonymization, voice privacy

1. Introduction

Speech signals inherently convey rich and sensitive information about speakers, including their identity, health status, emotional state, accent, and gender [1]. As a result, the collection, storage, and sharing of speech data raise significant privacy concerns and are subject to privacy regulations such as the GDPR [2]. To mitigate these risks, speech data should be anonymized before being shared with third parties. At the same time, anonymized speech should preserve utility-related information, such as linguistic content and emotion information, to remain useful for downstream speech processing tasks. In the Voice Privacy Challenge 2026 (VPC 2026), the goal is not only to conceal speaker identity but also to preserve both linguistic content and emotional information in the anonymized speech.

Compared with previous editions, VPC 2026 introduces two major novelties. First, the challenge adopts a stronger Automatic Speaker Verification (ASV) attacker, which is pre-trained on a large-scale voice-converted dataset from Source Speaker Tracing Challenge [3] and subsequently fine-tuned using the in-domain anonymized speech data. This evaluation setup poses a more realistic and challenging threat model for assessing the robustness of anonymization systems. Second, VPC 2026 expands the evaluation framework to include two tracks: Track 1 focuses on English voice anonymization and largely follows the setup of VPC 2024 [4], while Track 2 addresses multilingual anonymization across French, Spanish, English, and German. In this work, we focus on the Track 1, English anonymization.

In the English Anonymization Track, privacy and utility are evaluated using different metrics. Privacy is assessed using the Equal Error Rate (EER) of the automatic speaker verification (ASV) attacker. Compared with previous editions of the challenge, the current privacy evaluation method introduces two important changes. First, cross-gender trials are included. Sec-

ond, a stronger previously mentioned ASV attacker, WavLM-ECAPA, is adopted. This attacker is based on an ECAPA-TDNN [5] speaker verification system that combines WavLM-Large weighted features with filter-bank (Fbank) features. The model is first pre-trained on large-scale VC datasets and fine-tuned on the anonymized speech data. Following the challenge protocol, anonymization systems are evaluated under multiple target privacy protection levels, corresponding to minimum target EERs of 10%, 20%, 30%, and 40%. Utility is evaluated from two complementary perspectives: speech intelligibility and emotional information preservation.

In this paper we investigate an adaptive anonymization strategy for the Voice Privacy Challenge 2026. Unlike previous work [6], which randomly applies either an ASR-TTS system or a VC system according to a predefined mixture ratio, our approach dynamically selects the anonymization strategy based on the predicted emotional state of the input speech. Specifically, utterances classified as neutral are anonymized using an ASR-TTS pipeline, which has been shown to provide strong privacy protection in the previous study [6]. In contrast, utterances classified as non-neutral are processed using either a lightweight voice conversion-based anonymization method or no anonymization, with the aim of better preserving emotion-related information while maintaining an acceptable level of privacy protection under the VPC 2026 evaluation protocol [7].

The rest of this paper is organized as follows. Section 2 describes the proposed emotion-aware anonymization framework. Section 3 presents the experimental setup and evaluation results. Section 4 discusses the potential limitations of the proposed approach and future research directions.

2. System Description

2.1. Emotion Classifier

Emotion2vec [8] is a self-supervised emotion classifier designed to learn the emotion representation from the speech signal. It adopts an online knowledge distillation strategy during pre-training, where a student model is optimized through gradient-based learning, while a teacher model is updated using an exponential moving average (EMA) of the student parameters. The teacher and student networks share the same architecture, consisting of a convolutional feature encoder and a Transformer-based backbone for contextual representation modeling. To capture emotional information at different temporal scopes, Emotion2vec employs two complementary training objectives. The frame-level loss encourages the model to learn fine-grained local emotional patterns, whereas the utterance-level loss focuses on modeling global emotional characteristics. The overall training objective allows the model to capture both local and global emotional information. In our work, we use the

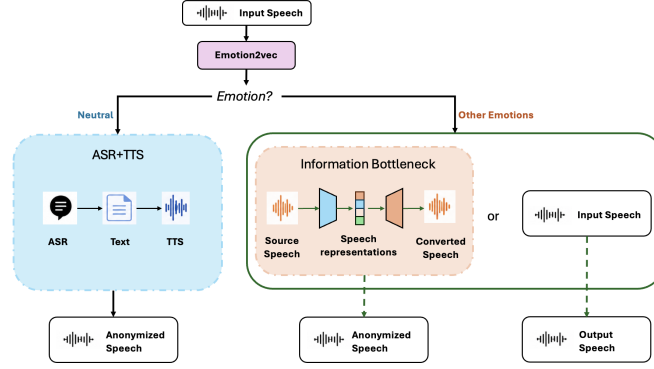


Figure 1: Overview of the proposed emotion-aware anonymization systems. The emotional state of the input speech is first predicted using Emotion2Vec. Utterances classified as neutral are anonymized using the Whisper–VITS ASR–TTS pipeline, whereas utterances classified as non-neutral are processed either by UUVc-based self-reconstruction (ours-1 method) or left unchanged (ours-2 method).

pretrained Emotion2vec+base model as the emotion classifier.

2.2. Cascaded ASR-TTS Anonymization Pipeline

This category of anonymization methods first converts the input speech into text using an Automatic Speech Recognition (ASR) system and subsequently generates a new speech signal from the transcript using a Text-to-Speech (TTS) system. In principle, such approaches can provide strong anonymization because the acoustic characteristics of the original speech are not preserved in the textual representation and, consequently, cannot be recovered during speech synthesis. As a result, the acoustic information contained in the original waveform is removed. Following the framework proposed in [6], we employ Whisper [9] as the ASR component. Specifically, the ‘medium-en’ model of Whisper is used to generate transcripts from the input speech. For speech synthesis, we adopt VITS [10], an open-source neural TTS model trained on the LibriTTS corpus. However, our implementation differs from that of [6] in the speaker selection strategy during synthesis; we use a single predefined target speaker to synthesize all anonymized speech samples.

2.3. Voice Conversion System

UUVc [11] is a voice conversion system that first decomposes speech into three complementary factors: linguistic content, prosody, and speaker characteristics. During the voice conversion process, the speaker embedding of the target speech replaces that of the source speech. The final converted speech is synthesized using the content and prosody representations from the original speech, together with the speaker embedding of the target speaker. In UUVc, a frozen HuBERT model [12] is employed to extract self-supervised speech representations, which are then quantized via k-means clustering to obtain discrete linguistic units U that represent the linguistic content. Prosodic information is modeled through pitch-energy and rhythm embeddings. Together with the content units U , these embeddings are input into a prosody predictor H_ϕ , which estimates pitch, energy, voicing, and duration-related information. Speaker characteristics are encoded as an utterance-level embedding a_s extracted from fine-tuned wav2vec 2.0 [13], capturing residual information not explained by the content and prosodic representations. In this work, we use the UUVc model to reconstruct the waveform from the disentangled representations extracted from

the original speech, including content, prosody, and speaker representations. Although the original speaker embedding is retained, the speech synthesis process inevitably introduces information loss and reconstruction artifacts, which can partially suppress fine-grained speaker-discriminative cues while largely preserving linguistic and emotional characteristics. Notably, this approach alone is insufficient to provide privacy protection under the challenge’s attack setting.

Table 1: Pre-trained modules used in our system and their implementations.

Module	Open-source link
Whisper [9]	https://github.com/openai/whisper
VITS [10]	https://rhasspy.github.io/piper-samples/#en_US-libritts_r-medium
UUVc [11]	https://github.com/b04901014/UUVc
Emotion2vec+base [8]	https://huggingface.co/emotion2vec/emotion2vec_plus_base

2.4. Hybrid Systems

In our submission to the VPC 2026, we demonstrate two hybrid systems, as illustrated in Figure 1. In both anonymization systems, the input speech is first analyzed using the pretrained Emotion2vec+base model to predict its emotional state. If the predicted emotion is neutral, anonymization is performed using the Whisper–VITS anonymization pipeline. For utterances predicted to be non-neutral, our first anonymization system (Ours-1 system) employs UUVc as a lightweight anonymization module. Specifically, we use the UUVc model to perform self-to-self voice conversion, reconstructing the waveform from the disentangled representations of the original speech. The second anonymization system (Ours-2 system) differs only in its treatment of predicted non-neutral speech: no anonymization is applied when the predicted emotion is not neutral. All components of the proposed hybrid systems are based on publicly available pretrained models, and no additional training data is needed. The corresponding open-source implementations are listed in Table 1.

Table 2: Privacy and utility evaluation results of different anonymization systems. Bold numbers indicate results that outperform all baseline anonymization systems on the corresponding official evaluation metric. The emotion-wise classification results for B4 baseline system are denoted by “-” because these statistics are not reported in the official Github repository.

	Libri-dev		Libri-test		IEMOCAP-dev					IEMOCAP-test				
	WER	EER	WER	EER	sad_acc	ang_acc	hap_acc	neu_acc	UAR	sad_acc	ang_acc	hap_acc	neu_acc	UAR
Orig.	1.80	6.08	1.84	3.41	63.63	79.78	66.95	65.97	69.08	72.57	72.82	67.19	71.66	71.06
B2	10.44	4.23	9.96	3.72	31.14	64.44	67.09	57.97	55.66	32.78	56.97	57.98	66.23	53.49
B3	4.31	16.32	4.31	14.24	0.42	70.45	47.45	34.00	38.08	0.30	61.07	39.78	39.79	35.24
B4	6.16	15.28	5.90	11.13	-	-	-	-	41.97	-	-	-	-	42.78
B5	4.91	20.05	4.44	17.05	8.26	64.51	37.27	50.29	40.08	4.57	58.05	33.81	56.58	38.25
Ours-1	3.87	34.94	3.79	31.49	66.17	66.07	45.43	51.46	57.28	68.72	64.06	42.15	49.81	56.18
Ours-2	3.89	35.16	3.72	32.82	59.35	81.04	61.68	50.23	63.08	68.76	75.54	62.89	51.57	64.69

3. Experimental Results

3.1. Dataset

Anonymization systems’ evaluation was conducted on the subsets of LibriSpeech and IEMOCAP provided by the challenge organizers. Specifically, emotion preservation performance was evaluated on the IEMOCAP-dev and IEMOCAP-test. Speech intelligibility was assessed on libri-dev-asr and libri-test-asr through automatic speech recognition (ASR) performance. Anonymization performance was evaluated using the official speaker verification attacker. Specifically, the libri-dev-enrolls, libri-dev-trials-mix, libri-test-enrolls, and libri-test-trials-mix subsets were used to measure speaker verification performance on anonymized speech and original speech. In addition, the anonymized LibriSpeech *train-clean-360* dataset was used to fine-tune the attacker model provided by the challenge organizers to perform the *semi-informed* attack.

3.2. Attack Model

In VPC 2026, the attacker is based on an ECAPA-TDNN speaker verification system that combines WavLM-Large weighted features with filter-bank (Fbank) features. The model is first pretrained on approximately 9,000 hours of voice-converted speech from the Source Speaker Tracing Challenge [3]. Subsequently, it is fine-tuned on the anonymized LibriSpeech, *train-clean-360* dataset resulting in a *semi-informed* attacker adapted to the anonymized speech domain.

3.3. Evaluation Metrics

Voice anonymization aims to conceal speaker identity while preserving the utility of the speech signal. Accordingly, evaluation considers three complementary aspects: anonymization performance, speech intelligibility, and emotion preservation. Privacy protection is evaluated using Equal Error Rate (EER) under the *semi-informed* attacker; higher EER values correspond to stronger anonymization performance. Speech intelligibility is measured using Word Error Rate (WER). Emotion preservation is assessed using Unweighted Average Recall (UAR) across four emotion categories (happy, neutral, angry, and sad). Lower WER and higher UAR indicate better utility preservation.

3.4. Experimental Results

As shown in Table 2, we present the performance of the proposed anonymization systems, namely Ours-1 and Ours-2, in comparison with the official baselines (B2–B5). Overall, both systems consistently outperform all baselines across privacy

(EER) and utility (WER and UAR) metrics, demonstrating the effectiveness of the proposed framework.

A key difference between the two systems lies in the handling of speech segments predicted to carry non-neutral emotional content. In the Ours-1 system, such segments are processed using UUVC, resulting in a weak anonymization strategy that suppresses part of the speaker-related information. In contrast, Ours-2 directly preserves the original speech classified as non-neutral emotional speech in order to better maintain emotional expressiveness, thereby prioritizing utility over privacy protection. In terms of voice anonymization performance, Ours-1 achieves EERs of 36.57% and 33.49% on libri-dev and libri-test, respectively. In comparison, Ours-2 obtains lower EERs of 36.41% and 33.61% on the same evaluation sets. Both systems substantially outperform all baseline systems (B2–B5). For speech intelligibility, Ours-1 system achieves WERs of 3.87% (dev) and 3.79% (test), while Ours-2 yields 3.89% and 3.72%, respectively. Both systems outperform all baseline methods, demonstrating that linguistic content is largely preserved after anonymization. However, there remains a noticeable degradation compared to the original speech (1.80% on dev and 1.84% on test). For emotion preservation, Ours-1 obtains UAR scores of 57.28% (dev) and 56.18% (test). In contrast, Ours-2 achieves higher UARs of 63.08% and 64.69% on the corresponding sets, indicating better retention of emotional characteristics. This improvement is consistent with the design choice of Ours-2 system, which doesn’t anonymize the predicted non-neutral emotional speech.

4. Limitation and future work

4.1. Analysis of Privacy Risks

In this section, we investigate the limitation of the proposed approach that is not captured by the official VPC 2026 evaluation protocol. Specifically, we would like to examine whether the proposed anonymization systems introduce substantial privacy leakage for emotionally expressive speech, thereby causing the current evaluation protocol to underestimate the actual privacy risk of the proposed systems. In VPC 2026, privacy is evaluated using speaker verification attacks on LibriSpeech, while emotion preservation is assessed on IEMOCAP. Consequently, privacy and utility are measured under different evaluation conditions. The proposed emotion-aware anonymization strategy implicitly exploits this mismatch. We observe that 86.45% of utterances in the LibriSpeech *test-clean* set are classified as neutral by the Emotion2vec+base emotion classifier, whereas only 32.53% of utterances in IEMOCAP-test receive the same label. As a result, the proposed system applies strong anonymiza-

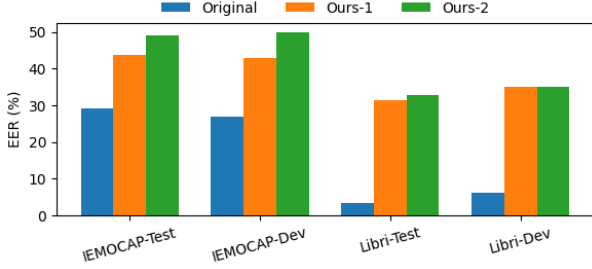


Figure 2: Attack performance across different test sets. EERs on the original unprotected test sets are obtained using the pre-trained VPC 2026 attacker. EERs on anonymized test sets are obtained using the VPC 2026 attackers fine-tuned on anonymized LibriSpeech train-clean-360.

Table 3: Statistics of the ASV evaluation protocol.

Subset		# Speakers		# Utt.	# ASV trials	
		M	F		Labels	# Trials
IEMOCAP-dev	Enrollment	2	2	40	Same-speaker	2706
	Trial	3	2	2706	Different-speaker	10824
IEMOCAP-test	Enrollment	2	2	40	Same-speaker	2745
	Trial	2	3	2745	Different-speaker	10980

tion to a substantial portion of LibriSpeech utterances, while emotionally expressive speech is processed using either a weak anonymization strategy or no anonymization at all in order to better preserve emotional information. Although this design leads to a favorable privacy–utility trade-off under the official VPC evaluation protocol, it also raises an important question: does the reduced anonymization strength for emotional speech introduce significant privacy leakage that remains undetected by the current benchmark?

To answer this question, we further evaluate speaker verification performance on IEMOCAP. While such an analysis is not required by the VPC framework, it provides additional insight into the robustness and generalizability of the proposed anonymization strategy under different conditions. We constructed speaker verification test pairs on both the IEMOCAP development and test sets and evaluated the attacker EER under the conditions illustrated in Figure 2. The statistics of the test pairs constructed by ourselves are summarized in Table 3. Interestingly, the attackers exhibit significantly different performance across different test sets as shown in Figure 2. Notably, directly interpreting the absolute EER values as privacy protection performance may lead to an overestimation of privacy protection [14]. Our intention in reporting these additional results is to motivate the need for stronger cross-domain attackers and unified privacy–utility evaluation.

A possible explanation for the performance discrepancy of the attacker across different datasets is that the speech data used for attacker training differ substantially from emotional speech corpora such as IEMOCAP in terms of speaking style, emotional expressiveness, recording conditions, and acoustic variability. Due to the limited availability of large-scale emotional speech resources, emotionally expressive speech is often under-represented during attacker training. As a result, a domain mismatch may exist between the attacker training data and the emotional speech used during evaluation, potentially leading to different attack performance across datasets. These findings sug-

gest that privacy evaluation results obtained on a single speech domain may not fully reflect the privacy risk under diverse real-world conditions. Future work should investigate privacy protection under more heterogeneous speech domains, particularly emotionally expressive and spontaneous speech. Developing stronger cross-domain attackers and establishing evaluation protocols that jointly assess privacy and utility on the same speech data may provide a more comprehensive assessment of privacy protection.

4.2. The Role of Emotion in Privacy Leakage

As discussed earlier, we noticed that LibriSpeech speech data are predominantly classified as neutral by the Emotion2vec classifier, whereas IEMOCAP contains a much larger proportion of emotionally expressive speech. This observation motivated us to investigate another question: are the large EER differences observed between LibriSpeech and IEMOCAP primarily caused by differences in their emotion distributions?

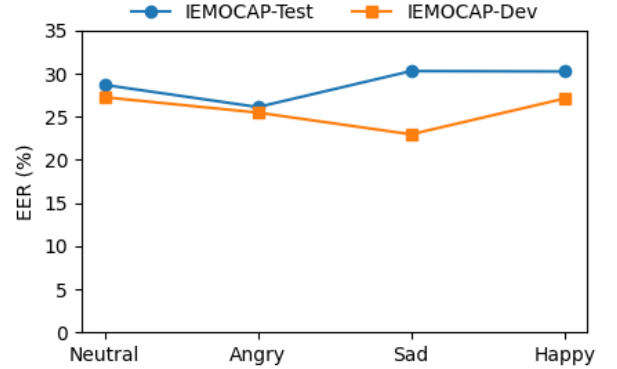


Figure 3: Emotion-wise attack performance analysis on the original unprotected IEMOCAP dataset. Results are obtained using the pre-trained VPC 2026 speaker verification attacker.

To this end, we constructed speaker verification test pairs grouped by emotion categories and evaluated attacker performance separately for each emotion condition. The results are presented in Figure 3. Interestingly, we observed only minor EER differences across different emotion categories. Although certain emotions exhibit slightly higher or lower privacy leakage, the overall variation remains relatively small compared with the substantial verification performance gap observed between LibriSpeech and IEMOCAP datasets.

These observations suggest that emotion alone cannot explain the domain-dependent speaker verification attack performance difference. While emotion serves as an effective auxiliary cue in our hybrid anonymization system and leads to a favorable privacy–utility trade-off under the challenge evaluation protocol, other factors are likely to contribute more substantially to the observed differences in attacker performance. Potential factors include speaking style, recording conditions, acoustic variability, speaker diversity, and other dataset-specific characteristics. Understanding which factors are most responsible for privacy leakage across speech domains remains an open research question. Future work will investigate these factors in a more controlled manner and explore whether domain-aware anonymization strategies can provide robust anonymization performance across diverse speech conditions.

5. Acknowledgments

This work was funded by the European Union’s Horizon Europe research and innovation programme grant No 101168193.

6. Generative AI Use Disclosure

Generative AI tools were used for the purpose of improving language and readability. The authors solely developed and verified all scientific aspects of this work, including the study design, analysis, experimental results, and conclusions. The authors have full responsibility for the manuscript and no generative AI tool is listed as an author.

7. References

- [1] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mubaa *et al.*, “Preserving privacy in speaker and speech characterisation,” *Computer Speech & Language*, vol. 58, pp. 441–480, 2019.
- [2] European Parliament and Council, “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation),” Apr. 2016, Official Journal of the European Union, L 119, pp. 1–88. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.
- [3] M. Li, P. Zhang, Y. Ren, Z. Cai, and H. Nishizaki, “The SSTC 2024 challenge evaluation plan,” 2024, [Online]. Available: https://sstc-challenge.github.io/file/Evaluation_Plan.pdf.
- [4] 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,” 2024, Version 2.0. [Online]. Available: https://www.voiceprivacychallenge.org/docs/VoicePrivacy_2024_Eval_Plan.v2.0.pdf.
- [5] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification,” in *Interspeech 2020*, 2020, pp. 3830–3834.
- [6] H. L. Xinyuan, Z. Cai, A. Garg, K. Duh, L. P. García-Perera, S. Khudanpur, N. Andrews, and M. Wiesner, “HLTCOE JHU submission to the Voice Privacy Challenge 2024,” in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 61–66.
- [7] 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, Version 1.2. [Online]. Available: https://www.voiceprivacychallenge.org/vp2026/docs/VPC_2026_june14.pdf.
- [8] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, “emotion2vec: Self-supervised pre-training for speech emotion representation,” in *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 15 747–15 760.
- [9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 28 492–28 518.
- [10] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 5530–5540.
- [11] L.-W. Chen, S. Watanabe, and A. Rudnicky, “A unified one-shot prosody and speaker conversion system with self-supervised discrete speech units,” in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [12] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.
- [13] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.
- [14] M. Panariello, S. Meyer, P. Champion, X. Miao, M. Todisco, N. T. Vu, and N. Evans, “The risks and detection of overestimated privacy protection in voice anonymisation,” in *5th Symposium on Security and Privacy in Speech Communication*, 2025, pp. 8–12.