

AISRC System Description for Voice Privacy 2026 Challenge

Ngoc Hung Le¹, Kyujin Kim¹, Sangjun Park¹, Souhwan Jung^{1,**}

¹ Soongsil University, South Korea

hungln@soongsil.ac.kr, rndlwjs@ssu.ac.kr, alexjuny@soongsil.ac.kr, souhwanj@ssu.ac.kr

Abstract

Speaker anonymization aims to conceal a speaker’s identity while preserving the linguistic content and paralinguistic information of the original speech; to meet this objective in the Voice Privacy Challenge 2026, we propose a cascaded speaker anonymization system. Unlike voice conversion methods, our system employs a robust cascaded architecture to explicitly decouple and reconstruct speech attributes. In parallel, automatic speech recognition extracts the text transcript and an emotion classification module captures the underlying paralinguistic state. The recognized text is then passed to an emotional text-to-speech synthesis engine, while the predicted emotion is used to select an appropriate emotional prompt for conditioning the generation. By handling the conversion through discrete stages, our system aims to reduce speaker-identity cues while retaining the linguistic content and expressive characteristics of the source speech.

Index Terms: Speaker Anonymization, Voice Privacy 2026 challenge, Emotional Text To Speech

1. Introduction

Speech data on the Internet have proliferated exponentially due to the advent of social media, encapsulating a wealth of sensitive personal information such as the speaker’s identity, age, gender, health status, and geographical background [1]. This sensitive information can be exploited by unauthorized speaker identification or malicious speech attribute recognition systems, motivating privacy-preserving speech processing before data publication or sharing. With the enforcement of data protection frameworks like the General Data Protection Regulation (GDPR) [2] in the European Union, preserving voice privacy has become a paramount concern for the speech processing community.

Speaker anonymization serves as a proactive privacy protection solution implemented before users share their speech data [3]. Its fundamental objective is to remove a speaker’s identity traits while preserving the linguistic content and essential paralinguistic attributes of the original speech. To establish fair benchmarks and drive technical innovation, the Voice Privacy Challenge (VPC) was established, with successful editions held in 2020, 2022, and 2024 [4, 5, 6]. Building upon the paradigms established in prior challenges, particularly the VPC 2024 emphasis on preserving emotional expressiveness while reducing speaker verifiability, the current VPC 2026 edition leverage a more powerful attacker which reduces the privacy of top performing system in VPC 2024 [7] and also expand the benchmark to multilingual [8].

While prior mainstream approaches have relied on signal processing based methods [9] or neural voice conversion (VC) using feature-disentanglement architectures [10, 11], these systems can suffer from information leakage when speaker-related cues might remain entangled with content or prosody or some signature artifact from the VC that the attacker can learn in semi-informed scenario [7].

To address these limitations, we propose a cascaded speaker anonymization system for the VPC 2026 challenge. Our framework reduces speaker-identity leakage by decomposing and reconstructing the speech signal through explicit symbolic and categorical bottlenecks. The system uses two parallel analysis branches followed by a synthesis stage:

1. **Automatic Speech Recognition (ASR):** Transcribes the input waveform into pure textual linguistic content, stripping away acoustic identities.
2. **Speech Emotion Recognition (SER):** Extracts the paralinguistic state from the source speech, mapping it to descriptive emotional attributes.
3. **Emotional Text-to-Speech (TTS):** Uses the recognized text for synthesis and the predicted emotion label to select an appropriate emotional prompt for generating anonymized speech.

Experimental results on the VPC 2026 datasets demonstrate that our cascaded system provides a favorable balance between privacy protection and emotion preservation, effectively reducing speaker identifiability while maintaining linguistic and paralinguistic utility.

2. Related Works

According to the established VPC benchmarks, speaker anonymization frameworks can be broadly classified into signal-processing-based techniques [9] and learning-based generative models [12, 13]. This section provides an overview of these conventional methods alongside emerging cascaded paradigms.

2.1. Signal-Processing Based Speaker Anonymization

Signal-processing approaches anonymize speech by directly modifying acoustic characteristics, often without requiring large-scale training data. Early methods in this category primarily manipulate short-term spectral and prosodic cues, such as formant frequencies and fundamental frequency (F_0). For example, McAdams coefficient-based transformations randomly shift formant locations to distort the spectral envelope and obscure speaker identity [9]. Although these methods are computationally efficient and simple to deploy, strong transformations can substantially degrade speech naturalness and intelli-

** indicates the corresponding author.

gibility [5]. Moreover, because these transformations are typically defined by fixed mathematical rules, an attacker can potentially adapt to or partially invert them, limiting their robustness in stronger attack settings [10].

2.2. Feature Disentanglement and Voice Conversion Based Anonymization

With the rise of deep learning, neural VC has become a dominant approach to speaker anonymization, often outperforming signal-processing baselines in both privacy and utility metrics [12, 14]. A typical VC-based system decomposes speech into multiple representations, including a speaker identity embedding, often extracted using an automatic speaker verification (ASV) model, and a linguistic content representation, obtained from an ASR model [14] or a self-supervised speech model [15]. During inference, the original speaker embedding is replaced with a pseudo-speaker embedding selected from an external pool, for example by choosing a distant speaker or a speaker of a different gender. Other systems generate pseudo-speaker embeddings with generative models, such as generative adversarial networks (GANs) [16] or flow-based models [17].

Although VC-based systems are effective against standard attackers, they can remain vulnerable when speaker-related cues are not fully disentangled from content, prosody, or synthesis artifacts. This limitation becomes more critical under stronger semi-informed attackers trained on diverse VC data, for which conventional VC-based anonymization can become less robust [7].

2.3. Cascaded ASR-TTS Anonymization Systems

Cascaded ASR-TTS anonymization offers an alternative strategy for reducing information leakage in feature-disentanglement-based systems [18]. By converting the input speech into text before resynthesis, these systems introduce an explicit linguistic bottleneck that breaks much of the continuous acoustic link between the source and anonymized speech. Our proposed framework extends this paradigm for VPC 2026 by adding a parallel emotion-recognition bottleneck. The predicted emotion is then used to condition an emotional TTS model, allowing the system to preserve paralinguistic expressiveness while removing speaker-specific acoustic traits.

3. Proposed Method

Figure 1 illustrates the overall pipeline of the proposed system. In parallel, our system extracts the linguistic content of the input utterance using an ASR model and predicts its emotional state using an SER model. Based on the predicted emotion, we select a suitable emotional prompt and use it to condition an emotional TTS system, which generates the anonymized speech waveform.

3.1. Speech Recognition and Emotion Recognition

For speech recognition, we use Whisper Large V3 [19]¹. To encourage normalized textual transcripts, we suppress numeric tokens during decoding so that numbers are emitted as words rather than digits. For emotion recognition, we use emotion2vec [20]² to classify the emotional state of each source utterance. Because the emotion labels produced by emotion2vec

¹<https://github.com/openai/whisper>

²https://huggingface.co/emotion2vec/emotion2vec_plus_large

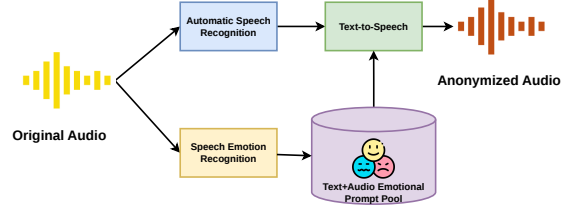


Figure 1: Overview of the proposed cascaded speaker anonymization system.

do not exactly match the VPC 2026 emotion categories, we map them to the challenge labels using the grouping shown in Table 1. Specifically, we sum the raw emotion2vec scores within each target group and select the group with the maximum summed score:

$$\hat{c} = \arg \max_{c \in \mathcal{C}} \sum_{\ell \in \mathcal{G}_c} s_\ell,$$

where s_ℓ denotes the raw score of emotion2vec label ℓ , $\mathcal{G}_c$ is the set of raw labels mapped to VPC category c , and $\mathcal{C} = \{\text{ang, hap, sad, neu}\}$.

3.2. Emotional TTS

We employ VoxCPM2 [23]³ as the waveform generation backend for the proposed cascaded anonymization system. VoxCPM2 is a multilingual controllable speech generation model that represents speech in a continuous latent space and synthesizes audio autoregressively with a hierarchical language-model and flow-matching architecture.

Among the generation interfaces supported by VoxCPM2, we use the *Controllable Cloning* mode. According to the VoxCPM2 formulation, this mode combines a reference-audio pathway, which provides the voice identity, with a natural-language style description, which controls attributes such as emotion, speaking rate, pitch, and energy. To provide rich conditioning beyond simple emotion labels, we designed descriptive style prompts for each target emotion as shown in Table 1 by pairing primary affective terms with explicit acoustic and vocal delivery descriptors (such as vocal tremor, shouting, or flatness). Therefore, in our system the emotion is not assumed to come only from the audio prompt. Instead, the emotion-specific prompt

³<https://github.com/OpenBMB/VoxCPM2>

Table 1: Mapping from emotion2vec labels to VPC 2026 emotion categories.

VPC label	emotion2vec labels	TTS style prompt
ang	angry, disgusted	furious, shouting, intense voice
hap	happy, surprised	very happy, excited, cheerful voice
sad	sad, fearful	deeply sad, sorrowful, trembling voice
neu	neutral, other, unknown	neutral, calm, flat voice

Table 2: *Track 1 results. EER and UAR are better when higher, while WER is better when lower. EER is reported for the semi-informed ASV attacker. Bold indicates the best anonymization result for each metric and split.*

System	WER, % ($\downarrow$)		UAR, % ($\uparrow$)		EER, % ($\uparrow$)	
	LS-dev	LS-test	IEMOCAP-dev	IEMOCAP-test	LS-dev	LS-test
Original	1.80	1.84	69.08	71.06	6.08	3.41
B2 [9]	9.61	9.11	55.66	53.52	4.23	3.72
B3 [21]	4.16	4.26	38.47	37.70	16.32	14.24
B4 [22]	6.16	5.90	41.97	42.78	15.28	11.13
B5 [14]	4.82	4.53	38.83	38.09	20.05	17.05
Ours	4.57	3.25	55.40	55.78	34.87	31.08

Table 3: *Track 2 multilingual results for French (fr), Spanish (es), English (en), and German (de). Bold indicates the best anonymization result for each metric and split.*

Dataset	Metric	Development					Evaluation				
		Orig.	BM1	BM2	BM3	Ours	Orig.	BM1	BM2	BM3	Ours
MLS-fr	EER	4.35	1.57	15.45	46.11	33.90	6.48	2.08	15.30	46.25	40.51
	WER	5.80	12.34	52.17	13.77	12.86	5.13	9.90	42.24	11.99	10.91
MLS-es	EER	5.83	0.93	10.74	46.23	40.84	3.03	1.58	8.74	45.92	39.96
	WER	3.78	6.17	20.37	10.56	7.05	3.75	6.20	28.92	10.12	7.41
MLS-en	EER	1.93	1.19	5.25	22.02	26.01	4.21	3.40	7.53	20.40	26.39
	WER	4.58	7.29	14.29	6.71	6.59	5.46	8.55	15.41	7.73	7.66
MLS-de	EER	1.84	1.12	10.05	46.54	40.92	3.28	1.36	10.90	46.05	37.61
	WER	5.25	7.76	19.28	9.74	10.43	5.39	9.50	24.40	10.10	11.56
MLS-Avg.	EER	3.49	1.20	10.37	40.22	35.42	4.25	2.10	10.62	39.65	36.12
	WER	4.85	8.39	26.52	10.20	9.23	4.93	8.53	27.74	9.98	9.39
EmoTrack2	UAR	72.09	41.30	27.13	26.24	57.34	78.90	42.63	29.04	24.84	62.22

and the corresponding style description are used jointly: the prompt provides a consistent external reference voice and affective example, while the text instruction explicitly specifies the predicted emotional state.

3.3. Emotion Prompt Construction

To condition emotional speech synthesis, we construct a compact prompt bank with one reference prompt for each target emotion. Candidate utterances are selected from CREMA-D [24] and CAMEO [25] according to their annotated emotion labels. CREMA-D is used for English prompts, whereas CAMEO is used for the other languages, and the prompt language is matched to the language of the source speech. Since the expressiveness of the reference audio directly affects the generated speech, we further assess candidate quality with the SER model described in Section 3.1. The resulting emotion scores are mapped to the target label using Table 1, and candidates with stronger evidence for the desired emotion are preferred. For each target emotion, the selected prompt is formed by concatenating several high-confidence utterances from the same speaker until an 8-second reference duration is obtained.

4. Experimental Setup

4.1. Dataset

We follow the official VPC 2026 protocol, which defines English and multilingual anonymization tracks. Track 1 uses LibriSpeech (LS) [26] subsets for privacy and linguistic-utility evaluation and IEMOCAP [27] for emotion-utility evaluation, following the VPC 2024 setup but adding cross-gender trials to the same-gender trials. The privacy score is computed on the mixed trial condition.

Track 2 evaluates multilingual anonymization in French, Spanish, English, and German. Privacy and linguistic utility are evaluated on MLS [28] subsets, while emotion utility is evaluated on EmoTrack2, sampled from multiple emotional speech datasets [8].

4.2. Metrics

Privacy is measured by the equal error rate (EER) of the official ASV attacker, where higher EER indicates stronger anonymization. VPC 2026 uses a semi-informed attacker: an ASV model pretrained on large-scale voice-converted speech is fine-tuned on anonymized training data produced by each submitted system [7]. Track 1 uses anonymized English data for fine-tuning, while Track 2 uses language-specific anonymized MLS subsets and averages scores across languages.

Utility is measured by word error rate (WER) for linguistic content preservation and unweighted average recall (UAR) for emotion preservation. Lower WER and higher UAR indicate better utility. Track 1 uses English ASR and IEMOCAP-trained SER models; Track 2 uses Whisper Large V3 for multilingual ASR and emotion2vec-large for multilingual SER.

4.3. Configuration

The base configuration uses one prompt for each target emotion and an emotion-specific text style prompt during controllable TTS generation. This base configuration was submitted to the challenge with the suffix “**_emotts_variant**”. To analyze the effect of prompt conditioning, we additionally evaluate four Track 1 variants: a large emotion prompt pool with 100 prompts per emotion, an all-neutral prompt pool constructed from LS, the base emotion prompt pool without the emotion text prompt, and an all-neutral LS prompt pool without any text prompt.

Table 4: *Intermediate ASR-stage WER before speech synthesis.*

Dataset	Subset	Dev WER, %	Test WER, %
IEMOCAP	Emotion eval.	11.69	11.86
LS	Enrolls	3.49	2.54
LS	Trials	3.79	2.65
MLS-fr	Enrolls	7.30	6.86
MLS-fr	Trials	7.69	6.43
MLS-es	Enrolls	5.08	4.71
MLS-es	Trials	5.00	5.46
MLS-en	Enrolls	5.56	6.77
MLS-en	Trials	5.55	6.51
MLS-de	Enrolls	7.80	9.41
MLS-de	Trials	8.47	8.98

Table 5: *Intermediate emotion-classification UAR before speech synthesis.*

Dataset	UAR, %
IEMOCAP-dev	83.02
IEMOCAP-test	84.26
EmoTrack2-dev	74.49
EmoTrack2-test	80.40

These variants isolate the contributions of audio-prompt diversity, emotion-specific audio prompts, and text-based style instructions.

5. Experimental Results

5.1. Track 1

Table 2 reports the English Track 1 results. For privacy, the proposed system achieves the highest EER among all systems. These values substantially exceed the strongest baseline B5, indicating that the cascaded ASR&SER-to-TTS design effectively suppresses speaker-specific acoustic information under the semi-informed ASV attacker.

For linguistic utility, our system obtains the best LS-test WER. For emotion utility, our system achieves 55.40% UAR on IEMOCAP-dev and the best IEMOCAP-test UAR of 55.78%. B2 is slightly higher on the development set, with 55.66%, but our system provides the strongest test-set emotion preservation. Overall, Track 1 shows a favorable privacy-utility trade-off, combining the highest EER, the best test-set WER, and the best test-set UAR.

5.2. Track 2

Table 3 reports multilingual Track 2 results. Our system obtains average EERs of 35.42% on development and 36.12% on evaluation, below BM3 at 40.22% and 39.65%, but still substantially higher than BM1 and BM2. It also achieves the best English EER on both splits, with 26.01% on MLS-en development and 26.39% on MLS-en evaluation.

For linguistic utility, our average WERs are 9.23% on development and 9.39% on evaluation, close to BM1 at 8.39% and 8.53%. The system performs best among anonymization systems for English WER, remains competitive for Spanish, while French and German are more challenging. For emotion utility, it achieves the highest EmoTrack2 UAR. Overall, Track 2 shows strong emotion preservation and competitive multilin-

Table 6: *Track 1 test-set ablation for audio-prompt and text-prompt conditioning.*

Audio prompt	Text prompt	WER, %	UAR, %	EER, %
Base emotion	Emotion	3.25	55.78	31.08
Large emotion	Emotion	3.75	55.31	32.39
Base emotion	None	3.11	51.16	33.30
All-neutral	Emotion	3.35	36.49	33.58
All-neutral	None	3.03	32.51	33.04

gual content preservation, with average privacy lower than BM3 but stronger than BM1 and BM2.

5.3. Intermediate ASR and SER Analysis

Tables 4 and 5 summarize the intermediate ASR and SER stages before synthesis. The ASR module produces accurate transcripts for the non-emotional linguistic-evaluation data, where the challenge uses the trial sets for WER computation and WER remains below 9.5% for all LS and MLS enrollment/trial subsets. We do not report EmoTrack2 ASR WER because ground-truth transcripts are not available for this emotional dataset.

The intermediate emotion classifier reaches 84.26% UAR on IEMOCAP-test and 80.40% on EmoTrack2-test, while the final synthesized speech obtains lower but still best-in-baseline UARs of 55.78% for Track 1 and 62.22% for Track 2 evaluation. This gap indicates that correct emotion prediction must also be reliably expressed by the TTS model.

5.4. Impact of Prompt and Text Conditioning

For this ablation, privacy is evaluated with the same semi-informed ASV attacker fine-tuned on anonymized training data produced by the base configuration, which uses one prompt per emotion. Increasing the emotion prompt pool does not improve UAR and also increases WER, suggesting that more emotion prompts do not necessarily improve emotion preservation. Removing the emotion text prompt from the base emotion-prompt configuration reduces UAR from 55.78% to 51.16%, showing that text-based emotion conditioning provides additional paralinguistic control. The all-neutral prompt variants yield only modest EER gains, but they substantially degrade UAR, especially without the text prompt. This indicates that neutral prompts may slightly increase privacy, but removing emotion-specific conditioning is not an effective privacy-utility trade-off.

5.5. Privacy Discussion for the TTS-Based Approach

The proposed system reconstructs speech from an ASR transcript and a predicted emotion label, thereby removing most direct acoustic speaker cues from the source waveform. This text-based bottleneck is expected to provide strong privacy, but the remaining EER degradation under the semi-informed attacker suggests that information beyond speaker acoustics may still be exploitable. One possible source is the SER stage, which may assign speaker-dependent emotion patterns. However, the prompt-conditioning ablation in Table 6 does not show a clear EER improvement when emotion-specific conditioning is removed, indicating that emotion conditioning is unlikely to be the main factor.

We therefore hypothesize that part of the remaining privacy loss may be due to content leakage introduced by the construction of enrollment and trial sets. If utterances from the same

speaker contain characteristic lexical content, or if enrollment and trial material are not sufficiently separated by content, a semi-informed attacker may exploit text-related cues even after the original acoustic identity has been removed. This hypothesis is consistent with prior observations on content leakage [29], but requires further controlled evaluation. Future privacy evaluations should therefore consider more systematic enrollment-trial construction to better separate speaker-identity leakage from linguistic-content leakage.

6. Conclusion

This paper presented the AISRC system for the VPC 2026, using a cascaded architecture to anonymize speech through explicit text and emotion bottlenecks. The system achieves a strong privacy-utility trade-off in Track 1, with the highest EER, best LS-test WER, and best IEMOCAP-test UAR among the compared anonymization systems. In Track 2, it provides competitive multilingual WER, the best English EER, and the highest EmoTrack2 UAR, although its average multilingual EER remains below BM3. The intermediate-stage analysis and prompt ablation show that accurate emotion prediction and emotion-conditioned prompting are important for preserving expressiveness, while future work should use controlled enrollment-trial construction to examine residual content leakage and further improve multilingual TTS robustness.

7. Generative AI Use Disclosure

Generative AI tools were used only to assist with grammar, wording, and readability. All scientific ideas, system design, experiments, data analysis, and conclusions are the authors' original work.

8. Acknowledgement

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT), (RS-2025-02215142, Development of Pseudonymization Technology for Suspected Criminal Information, 50%) and Korea Internet & Security Agency (KISA) grant funded by the Korea government(PIPC) (No.RS-2026-25530576, Development of Robust Real-time Deepfake Disruption Technology via Adversarial Noise Injection for Privacy Protection, 50%)

9. References

- [1] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, M. A. Abdelraheem, A. Abad, F. Teixeira, D. Matrouf, M. Gomez-Barrero, D. Petrovska-Delacrétaz, G. Chollet, N. Evans, T. Schneider, J.-F. Bonastre, B. Raj, I. Trancoso, and C. Busch, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441–480, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0885230818303875>
- [2] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps Towards a Common Understanding," in *Interspeech 2019*, 2019, pp. 3695–3699.
- [3] 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 *Interspeech 2020*, 2020, pp. 1693–1697.
- [4] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien, A. Chanclu, J.-F. Bonastre, M. Todisco, and M. Maouche, "The voiceprivacy 2020 challenge: Results and findings," *Computer Speech & Language*, vol. 74, p. 101362, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0885230822000080>
- [5] 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 Trans. Audio, Speech and Lang. Proc.*, vol. 32, p. 3477–3491, Jul. 2024. [Online]. Available: <https://doi.org/10.1109/TASLP.2024.3430530>
- [6] 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. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0885230826000513>
- [7] R. Arefeen, X. Miao, R. Tong, A. B. Ng, S. See, and T. Liu, "Dast: A dual-stream voice anonymization attacker with staged training," *arXiv preprint arXiv:2603.12840*, 2026.
- [8] 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," Master's thesis, 2026, eURECOM. Personal use of this material is permitted. The definitive version of this paper was published in Report, March 2026 and is available at :.
- [9] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker Anonymisation Using the McAdams Coefficient," in *Interspeech 2021*, 2021, pp. 1099–1103.
- [10] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, "Speaker anonymization using orthogonal household neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3681–3695, 2023.
- [11] X. Miao, R. Tao, C. Zeng, and X. Wang, "A benchmark for multi-speaker anonymization," *IEEE Transactions on Information Forensics and Security*, 2025.
- [12] J. He, S. Zhao, J. Zhou, H. Sun, H. Wang, and Y. Qin, "Emotion-preserving prosody anonymization network for voice privacy protection," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [13] I. R. Ulgen, Z. Cai, N. Andrews, P. Koehn, and B. Sisman, "Diffanon: Diffusion-based prosody control for voice anonymization," *arXiv preprint arXiv:2604.26281*, 2026.
- [14] P. Champion, "Anonymizing speech: Evaluating and designing speaker anonymization techniques," *arXiv preprint arXiv:2308.04455*, 2023.
- [15] Z. Zhang, W.-C. Huang, X. Wang, X. Miao, and J. Yamagishi, "Mitigating Language Mismatch in SSL-Based Speaker Anonymization," in *Interspeech 2025*, 2025, pp. 5133–5137.
- [16] S. Meyer, P. Tilli, P. Denisov, F. Lux, J. Koch, and N. T. Vu, "Anonymizing speech with generative adversarial networks to preserve speaker privacy," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 912–919.
- [17] Y. Xiao, Y. Zhao, Z. Ma, S. Wang, K. Yu, J. Liao, and X. Chen, "Anonymization, not elimination: Utility-preserved speech anonymization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 2763–2773, 2026.
- [18] S. Meyer, F. Lux, P. Denisov, J. Koch, P. Tilli, and N. T. Vu, "Speaker Anonymization with Phonetic Intermediate Representations," in *Interspeech 2022*, 2022, pp. 4925–4929.
- [19] 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.

- [20] 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*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 15 747–15 760. [Online]. Available: <https://aclanthology.org/2024.findings-acl.931/>
- [21] S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and N. T. Vu, "Prosody is not identity: A speaker anonymization approach using prosody cloning," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [22] M. Panariello, F. Nespoli, M. Todisco, and N. Evans, "Speaker anonymization using neural audio codec language models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4725–4729.
- [23] Y. Zhou, G. Zeng, X. Liu, X. Li, R. Yu, J. Gui, J. Wu, Z. Wang, X. Shen, R. Ye *et al.*, "Voxcpm2 technical report," *arXiv preprint arXiv:2606.06928*, 2026.
- [24] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.
- [25] I. Christop and M. Czajka, "Cameo: Collection of multilingual emotional speech corpora," in *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026, pp. 19 457–19 461.
- [26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [27] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
- [28] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Interspeech 2020*, 2020, pp. 2757–2761.
- [29] C. Franzreb, A. Das, T. Polzehl, and S. Möller, "Content leakage in librispeech and its impact on the privacy evaluation of speaker anonymization," in *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026, pp. 13 577–13 581.