

Natural Language Instructions for Voice Anonymization

Orange Research submission for the VoicePrivacy 2026 Challenge

Rachid Benhsina¹, Olivier Le Blouch¹, Rayane Bakari¹, Nicolas Gengembre¹

¹ Orange Research, France

rachid.benhsina@orange.com, olivier.leblouch@orange.com,
mohamedrayane.bakari@orange.com, nicolas.gengembre@orange.com

Abstract

This paper describes our submissions to the VoicePrivacy Challenge (VPC) 2026 for Track 1 and Track 2. We propose two voice anonymization systems based on a cascaded ASR-TTS pipeline using Whisper and Qwen3-TTS VoiceDesign, the latter conditioned on natural language instructions. Both systems share the same pipeline, differing only in the instruction conditioning TTS generation. VD-Pool, submitted for Track 2, conditions generation on a voice description of the target speaker. VD-E2V, submitted for both tracks, conditions generation on the emotion detected via emotion2vec from the source utterance. Results show that VD-Pool achieves strong privacy protection and low linguistic degradation across all evaluated languages, while VD-E2V achieves competitive privacy protection with an EER exceeding 33% on Track 1 and 39% on Track 2, and significantly improves emotion preservation with a UAR of over 53% on Track 2, while maintaining a WER below 7% across both tracks.

Index Terms: VoicePrivacy 2026 Challenge, Voice Anonymization, Text-to-Speech

1. Introduction

Speech signals encapsulate a wealth of identifiable personal information, including speaker identity, age, gender, and health status. The increasing deployment of voice-based systems raises significant privacy concerns, particularly under regulations such as the European General Data Protection Regulation (GDPR) [1]. Voice anonymization addresses this challenge by transforming speech to conceal speaker identity while keeping its utility intact. It finds applications across a wide range of real-world scenarios, from healthcare consultations and legal proceedings to social media and broadcast media [2], with distinct requirements on the balance between privacy protection and speech utility.

The task is formally defined as concealing speaker identity while preserving the utility of the speech signal. Utility is not an absolute concept — it depends on the intended downstream tasks, which may target linguistic content, emotional expressiveness, naturalness, or other speech attributes [2]. In the VPC 2026 evaluation protocol [3], privacy is assessed via the equal error rate (EER) obtained by a semi-informed automatic speaker verification attacker (higher is better, optimal at 50%). Utility is evaluated for two downstream tasks: automatic speech recognition (ASR), using word error rate (WER, lower is better), and speech emotion recognition (SER), using unweighted average recall (UAR, higher is better).

Formed in 2020, the VPC initiative has established a series of competitive benchmarking challenges to advance anonymization research [4, 5, 6]. VPC 2026 [3] introduces two key advances over VPC 2024. First, the challenge

extends beyond English to a multilingual setting, introducing Track 2 covering French, Spanish, English, and German, alongside the English Track 1. Second, the ECAPA-TDNN attacker used in VPC 2024 is replaced by a stronger WavLM-ECAPA system pre-trained on 94k hours of speech including large-scale voice-converted data and fine-tuned on anonymized speech under the semi-informed scenario, applied to both tracks, making it significantly harder for anonymization systems to protect speaker privacy.

Cascaded ASR-TTS systems have demonstrated strong privacy protection in prior VPC editions, as the transcription process discards speaker-specific acoustic cues and retains only linguistic content. However, whether such systems can effectively preserve the emotional expressiveness of the source speaker remains an open question. We hypothesize that conditioning the text-to-speech (TTS) generation on the emotion automatically detected from the source enables better emotion preservation, without compromising privacy or linguistic content.

To investigate this hypothesis, we propose two systems using Whisper large-v3 [7] as the ASR component and Qwen3-TTS VoiceDesign [8] as the TTS component, the latter conditioned on natural language instructions. **VD-Pool**¹, submitted for Track 2, conditions generation on a voice description prompt randomly drawn from a pool of prompts describing target voice characteristics. **VD-E2V**², submitted for both tracks, replaces the voice description with an emotion-conditioned instruction derived from automatic emotion detection via emotion2vec_base [9]. The rest of this paper is organized as follows: Section 2 reviews related works, Section 3 describes the proposed pipeline, Section 4 describes the submitted systems, provides implementation details, presents experimental results and an ablation study, Section 5 discusses limitations, and Section 6 concludes.

2. Related works

Voice anonymization approaches span a wide range of techniques. Signal processing methods such as the McAdams coefficient-based system [10] (B2) are naturally language-independent but provide limited privacy protection. Feature-based resynthesis systems include B4 [11], based on neural audio codec language modeling, and B5 [12], which combines vector quantized bottleneck features extracted from an ASR model with HiFi-GAN resynthesis. Among cascaded ASR-TTS approaches, B3 [13] combines phonetic transcription with GAN-based pseudo-speaker embedding generation and FastSpeech2 [14] synthesis; the HLTCOE JHU system [15] proposes a Whisper-

¹Submitted under suffix `_VD-Pool-T2` for Track 2.

²Submitted as `_VD-E2V-T1` for Track 1 and `_VD-E2V-T2` for Track 2.

VITS pipeline achieving strong privacy protection and low linguistic degradation in VPC 2024; and BM2 and BM3 [3] extend B3 to a multilingual setting using Whisper and IMS Toucan [16]. None of these systems employs natural language instructions to control voice identity or emotional expressiveness — the generation is instead conditioned on acoustic features, speaker embeddings, or fixed speaker selections.

While most voice anonymization research has focused on monolingual English speech, only a handful of works have addressed the multilingual setting. Miao et al. [17] proposed a language-independent anonymization system by replacing the language-dependent ASR acoustic model used in the VPC 2020 baseline B1 [18] with HuBERT [19], a multilingual self-supervised model, for content feature extraction, forming BM1 in VPC 2026. Meyer et al. [20] investigated the feasibility of multilingual anonymization across nine languages using an ASR-TTS cascade with IMS Toucan; the multilingual baselines BM2 and BM3 in VPC 2026 extend this work with a more recent version of IMS Toucan supporting over 7000 languages. Both systems use Whisper for transcription and ECAPA-TDNN for speaker identity encoding; BM2 additionally applies random F0 and energy modification, while BM3 estimates prosodic parameters directly from the transcription. Yao et al. [21] proposed MUSA, a disentanglement-based approach separating speaker identity from linguistic content through serial distillation, demonstrating robust cross-lingual generalization without explicit language supervision. More recently, Meyer et al. [22] presented the first investigation of voice anonymization for code-switching speech, exposing fundamental challenges in utility evaluation for spontaneous multilingual data.

3. Proposed Method

Cascaded ASR-TTS systems offer strong privacy protection by design: the transcription step discards all speaker-specific acoustic cues, preventing any leakage of the source speaker’s voice characteristics to the attacker. Linguistic utility is primarily determined by the quality of the TTS component. Among available TTS systems, Qwen3-TTS VoiceDesign stands out for its naturalness, expressiveness, and support for conditioning generation on natural language instructions — a capability that enables explicit control over both voice identity and emotional expressiveness, directly motivating our hypothesis that conditioning the TTS component on emotion can improve emotion preservation after anonymization. We propose two systems: VD-Pool, submitted for Track 2, and VD-E2V, submitted for both tracks, built upon the same pipeline and differing solely in the natural language instruction used to condition the TTS generation, as detailed hereafter.

3.1. System Overview

The proposed anonymization pipeline, illustrated in Figure 1, follows a cascaded ASR-TTS architecture. Given an original utterance, Whisper large-v3 transcribes the speech content. Although Whisper supports automatic language detection, the source language is provided as input at inference time — as done in BM2 and BM3 — to avoid potential language identification errors: Whisper achieves *only* 80.3% accuracy on the subset of FLEURS languages covered by its training data, which could introduce bias in the evaluation results. The transcript is then passed to Qwen3-TTS VoiceDesign, the core anonymization engine of this

pipeline, conditioned on both the input text and a natural language instruction. Trained on over 5 million hours of speech across 10 languages, it natively supports the four languages evaluated in this work (French, Spanish, English, and German). The final instruction is constructed by the Instruction Builder module from one of two sources, depending on which optional component is activated (see Figure 1 and Table 1). Finally, the anonymized speech is resampled to 16 kHz to comply with VPC 2026 format requirements.

Table 1: Configuration of optional components for each system.

System	(1) Emotion2vec	(2) Prompt Pool
VD-Pool	✗	✓
VD-E2V	✓	✗

3.2. Conditioning the TTS Generation

The Instruction Builder module constructs the natural language instruction passed to Qwen3-TTS VoiceDesign at each utterance, depending on which optional component is activated (see Table 1).

Voice-conditioned instruction (VD-Pool). When only component (2) is active, the instruction consists of a voice description prompt of the target speaker randomly drawn at the utterance level from a pool of 100 English prompts, some of which are illustrated in Table 2. Each prompt combines five vocal attributes: *age*, *gender*, *pitch*, *speaking pace*, and *speaker role*, according to the following pattern:

"A [age] [gender] voice, [pitch] and [characteristic], [pace], [role]."

The pool covers a balanced set of 50 female and 50 male voice descriptions, ensuring diversity across all attribute combinations. The utterance-level random selection minimizes the probability that consecutive utterances share the same voice identity.

Table 2: Examples of voice description prompts used by VD-Pool.

"A middle-aged female voice, deep and rich, steady pace, professor."
"A young male voice, medium and smooth, brisk pace, scientist."
"A middle-aged male voice, medium and smooth, relaxed pace, journalist."

Emotion-conditioned instruction (VD-E2V). When only component (1) is active, the source emotion is automatically detected by emotion2vec_base across seven categories: *angry*, *disgusted*, *fearful*, *happy*, *neutral*, *sad*, and *surprised*. The detected label is mapped to a fixed English instruction phrase following the pattern: *"speaking with [emotion], [adjective] and [adjective] tone,"* as illustrated in Table 3. No voice description is included, allowing Qwen3-TTS VoiceDesign to freely generate the most expressive voice for the injected emotion.

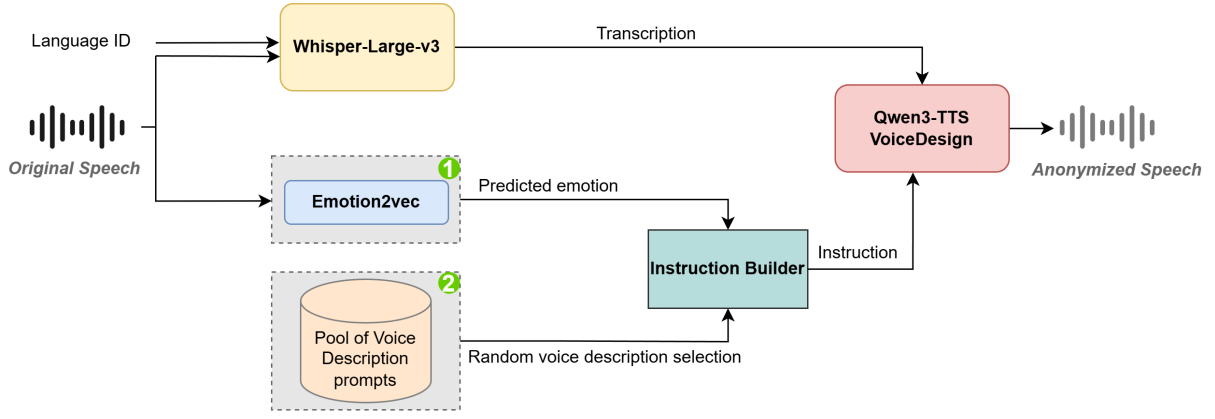


Figure 1: Overview of the proposed anonymization pipeline. Dashed blocks indicate optional components: (1) emotion detection via emotion2vec_base, active in VD-E2V; (2) voice description prompt pool, active in VD-Pool.

Table 3: Emotion-conditioned prompts used by VD-E2V.

Emotion	Instruction - "Speaking with ..."
Happy	"happiness, cheerful and lively tone."
Surprised	"surprise, lively and startled tone."
Neutral	"a calm and gentle tone."
Angry	"anger, intense and serious tone."
Disgusted	"disgust, cold and dismissive tone."
Fearful	"fear, tense and uneasy tone."
Sad	"sadness, soothing and subdued tone."

Combining voice description and emotion conditioning was not proposed as a system for the challenge, as the ablation study (Section 4.4) shows that such combination consistently degrades emotion preservation compared to emotion conditioning alone.

4. Experiments

4.1. Submitted Systems

4.1.1. VD-Pool System

The VD-Pool system is submitted for Track 2 only. Following the pipeline described in Section 3.1, only component (2) is active: at each utterance, a voice description prompt is randomly drawn from the predefined pool (see Section 3.2), with no emotion conditioning. This design targets strong privacy protection and low linguistic degradation across the four evaluated languages, without explicit control over emotion.

4.1.2. VD-E2V System

VD-E2V is submitted for Track 1 and Track 2, sharing the same configuration across the two tracks. It shares the same ASR-TTS backbone as VD-Pool (Section 3.1), but activates only component (1) instead: the instruction consists of an emotion-conditioned prompt derived from the source utterance (see Section 3.2), with no voice description. This design aims at better emotion preservation.

4.2. Implementation Details

All models are used off-the-shelf without fine-tuning. VD-Pool totals 3.25B parameters, combining Whisper large-v3 (1.55B) for transcription and Qwen3-TTS VoiceDesign (1.7B) for speech synthesis. VD-E2V additionally incor-

porates emotion2vec_base (164M) for emotion detection, bringing the total to 3.41B parameters. Inference runs on two NVIDIA A100 80GB GPUs with 4 parallel workers per GPU (8 workers total), an ASR batch size of 8 and a TTS batch size of 4 per worker, automatically configured based on available VRAM. The pipeline achieves an average inference time of approximately 2.3 seconds per utterance.

4.3. Results

4.3.1. Track 1 Results

Table 4 reports Track 1 results according to the VPC 2026 evaluation protocol. In terms of privacy, measured under the semi-informed attacker scenario, VD-E2V achieves an EER of 33.39% on the evaluation set, approximately 12 points above B5, the strongest baseline in this regard. The linguistic content is well preserved, with a WER of 2.45%, close to the original speech (1.84%). Regarding emotion preservation, VD-E2V reaches a UAR of 42.43%, comparable to B4 (42.78%), while B2, a DSP-based system that bypasses speech resynthesis entirely, retains a higher UAR (53.49%) at the cost of its poor privacy protection (EER of 4.71%).

4.3.2. Track 2 Results

Table 5 reports Track 2 results across four languages (French, Spanish, English, and German). In terms of privacy, VD-Pool achieves the best average EER of 41.67% on the evaluation set, slightly above BM3 (40.57%), and outperforming BM3 notably on English (29.86% vs. 22.25%), where BM3 shows weaker privacy protection than on other languages. VD-E2V reaches a competitive average EER of 39.08%, confirming that both proposed systems achieve strong privacy across all four languages. In terms of linguistic utility, VD-Pool obtains the lowest average WER of 6.27%, substantially below all multilingual baselines, and close to the original speech (4.93%). Although emotion conditioning slightly increases WER in VD-E2V (6.77%), linguistic content remains well preserved compared to all baselines. For emotion preservation, VD-E2V largely outperforms all systems with a UAR of 53.59%, compared to 42.63% for BM1, the baseline achieving the highest UAR, and only 26.96% for VD-Pool, confirming the effectiveness of emotion conditioning.

Taken together, these results support our hypothesis that conditioning the TTS component on emotion improves

Table 4: *Track 1 evaluation results: Privacy (semi-informed EER,%), and utility (WER,% and UAR,%) of the VPC 2026 baselines (B2, B3, B4, B5) and proposed system (VD-E2V) on anonymized data vs. original (Orig.). Bold values indicate the best result per row. Grey-highlighted values will be used for ranking by VPC 2026.*

Metric	Development						Evaluation					
	Orig.	B2	B3	B4	B5	VD-E2V	Orig.	B2	B3	B4	B5	VD-E2V
EER	7.34	6.97	19.14	17.96	24.20	36.66	3.91	4.71	16.85	14.33	21.28	33.39
WER	1.80	10.44	4.31	6.16	4.91	2.63	1.84	9.96	4.31	5.90	4.44	2.45
UAR	69.08	55.66	38.08	41.97	40.08	42.37	71.06	53.49	35.24	42.78	38.25	42.43

emotion preservation, while keeping privacy protection strong and linguistic degradation low across both tracks.

4.4. Ablation Study

We conduct an ablation study on Track 2 to assess two aspects:

- first, the upper bound of UAR achievable when replacing emotion2vec predictions with ground-truth emotion labels, and the impact of emotion2vec classification errors on emotion preservation;
- second, the effect of combining emotion conditioning with a target voice description in the instruction, compared to emotion conditioning alone.

For this purpose, we evaluate four configurations:

- **VD-Pool**: instruction contains a voice description prompt only, with no emotion conditioning;
- **VD-E2V**: instruction contains an emotion2vec-predicted emotion phrase only, with no voice description;
- **VD-GT**: instruction contains a ground-truth emotion phrase only, with no voice description;
- **VD-GT-Pool**: instruction combines a ground-truth emotion phrase with a voice description prompt.

VD-Pool and VD-E2V correspond exactly to the systems submitted for Track 2.

Ground-truth emotion labels provided by VPC 2026 for the SER evaluation data of Track 2 are used in VD-GT and VD-GT-Pool, to eliminate any bias introduced by emotion2vec prediction errors and isolate the effect of the voice description on emotion preservation. Results are reported in Table 6.

Compared to VD-Pool, the system without emotion conditioning (UAR of 26.96%), VD-GT achieves a UAR of 62.94%, representing a gain of approximately 36 points. This demonstrates both the upper bound of UAR attainable through perfect emotion conditioning and the strong capacity of Qwen3-TTS VoiceDesign to reproduce emotional expressiveness when conditioned on accurate emotion labels.

Comparing VD-GT and VD-GT-Pool reveals the impact of adding a voice description to the instruction: VD-GT-Pool achieves a better WER (6.14% vs. 7.31%) and a slightly better EER (40.39% vs. 38.73%), but at the cost of a significant drop of 17.61 points in UAR (45.33% vs. 62.94%). This confirms that describing the target voice identity in the instruction interferes with emotion preservation, likely because the model tends to prioritize speaker timbre matching when a voice description is present, at the expense of emotional expressiveness.

The gap between VD-GT (62.94%) and VD-E2V (53.59%) reflects the impact of emotion2vec_base classification errors on emotion preservation. Moreover, this impact may differ substantially across languages, since SER

accuracy is itself strongly affected by the language spoken in the source speech ([9], Table 4).

These results justify the submission of VD-E2V alongside VD-Pool for Track 2: while VD-Pool achieves stronger privacy and linguistic utility, VD-E2V significantly improves emotion preservation (53.59% vs. 26.96% UAR).

5. Discussion

Although the proposed systems achieve strong privacy protection and low linguistic degradation, the evaluation of emotion preservation through UAR raises a fundamental concern. The pipeline does not preserve the original emotional expressiveness of the source speaker — it recreates a new expression conditioned on a detected emotion label. This constitutes a typical instance of Goodhart’s Law [23]: by optimizing for UAR, the system satisfies the emotion classifier rather than faithfully preserving the speaker’s authentic way of conveying emotion. The emotional label is preserved, but the speaker’s individual expressiveness is transformed. We believe this distinction between *emotion label preservation* and *expressiveness preservation* deserves further attention in future voice anonymization evaluation frameworks.

Several practical limitations also apply. The pipeline is restricted to languages jointly supported by Whisper large-v3 and Qwen3-TTS VoiceDesign, which may limit its applicability to other languages. Additionally, Qwen3-TTS VoiceDesign can produce hallucinations in the generated audio. During the anonymization of the Track 2 corpus (60,261 utterances), we observed 20 files with abnormally long duration (exceeding 42 seconds, compared to the 10–20 second range of source MLS utterances), which were confirmed by listening to contain hallucinated content. These hallucinations can arise either from transcription errors produced by Whisper or from the TTS model itself. This observation is based on duration analysis and may not capture all hallucination cases, as shorter hallucinations may also occur. Finally, emotion control via natural language instructions remains coarse-grained: Qwen3-TTS VoiceDesign does not distinguish between different levels of emotional intensity within the same category (e.g., *happy* vs. *very happy*), which may reduce the fidelity of emotion preservation for subtle or intense expressions.

Beyond these practical limitations, a question raised by this work remains open: since the ASR step converts source audio to text, theoretically discarding all speaker-specific attributes before resynthesis, one would expect both VD-Pool and VD-E2V to reach the optimal EER of 50%, as both systems follow a cascaded ASR-TTS architecture. Yet this is not observed, and the gap is particularly pronounced for English in Track 2, where both the proposed systems and the multilingual baselines consistently show lower EER

Table 5: *Track 2 evaluation results: Privacy (semi-informed EER,%) and utility (WER,% and UAR,%) of the VPC 2026 multilingual baselines (BM1, BM2, BM3) and proposed systems (VD-Pool, VD-E2V) on anonymized data vs. original (Orig.) across four languages. Bold values indicate the best result per row. Grey-highlighted values will be used for ranking by VPC 2026.*

Language	Metric	Development						Evaluation					
		Orig.	BM1	BM2	BM3	VD-Pool	VD-E2V	Orig.	BM1	BM2	BM3	VD-Pool	VD-E2V
MLS-fr	EER	6.15	2.10	16.93	47.09	46.75	39.47	7.51	2.63	16.11	46.54	46.92	44.20
	WER	5.80	12.42	52.17	13.77	8.35	10.60	5.13	9.93	42.24	11.99	7.32	7.89
MLS-es	EER	8.97	1.63	11.91	46.23	44.35	40.69	4.86	1.83	10.23	46.46	44.52	41.33
	WER	3.79	6.15	20.37	10.56	6.04	5.40	3.75	6.21	28.92	10.12	5.17	5.47
MLS-en	EER	2.24	1.40	6.34	23.49	29.97	28.07	4.75	3.73	8.82	22.25	29.86	28.83
	WER	4.57	7.40	14.29	6.71	5.58	5.62	5.46	8.54	15.41	7.73	6.29	6.30
MLS-de	EER	3.00	1.43	11.38	46.84	46.77	41.95	11.39	1.75	11.88	47.02	45.41	41.98
	WER	5.25	7.81	19.28	9.74	5.70	6.66	5.38	9.38	24.40	10.10	6.33	7.45
MLS-Avg.	EER	5.09	1.64	11.64	40.91	41.96	37.54	7.13	2.49	11.76	40.57	41.67	39.08
	WER	4.85	8.45	26.52	10.20	6.41	7.07	4.93	8.52	27.74	9.98	6.27	6.77
EmoTrack2	UAR	72.09	41.30	27.13	26.24	27.85	52.30	78.90	42.63	29.04	24.84	26.96	53.59

Table 6: *Track 2 ablation study: Privacy (semi-informed EER,%) and Utility (WER,% and UAR,%) for four system configurations. Bold values indicate the best result per row.*

Language	Metric	Development				Evaluation			
		VD-Pool	VD-GT	VD-GT-Pool	VD-E2V	VD-Pool	VD-GT	VD-GT-Pool	VD-E2V
MLS-Avg.	EER	41.96	37.60	39.94	37.54	41.67	38.73	40.39	39.08
	WER	6.41	6.63	6.26	7.07	6.27	7.31	6.14	6.77
EmoTrack2	UAR	27.85	65.12	47.28	52.30	26.96	62.94	45.33	53.59

compared to French, Spanish, and German. We leave this question for future investigation.

6. Conclusions

This work addresses the VPC 2026, which extends prior editions by introducing Track 2, a multilingual track covering French, Spanish, English, and German, in addition to the English Track 1. We propose two systems built on a cascaded ASR-TTS pipeline with Whisper large-v3 and Qwen3-TTS VoiceDesign, differing only in the natural language instruction conditioning the TTS generation: VD-Pool, submitted for Track 2, conditions generation on a voice description of the target speaker; VD-E2V, submitted for both tracks, conditions generation on the emotion automatically detected from the source utterance. The ablation study confirms that conditioning Qwen3-TTS VoiceDesign on emotion alone outperforms the combination of emotion and voice description of the target speaker in terms of UAR. Nevertheless, optimizing for UAR does not guarantee faithful preservation of the speaker’s authentic expressiveness, a distinction that deserves further attention in future evaluation frameworks.

7. References

- [1] European Parliament and Council of the European Union, “Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data,” Official Journal of the European Union, Tech. Rep., 2016, general Data Protection Regulation (GDPR).
- [2] S. Meyer and N. T. Vu, “Use cases for voice anonymization,” in *Proc. SPSC Symposium*, 2025.
- [3] 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,” <https://www.voiceprivacychallenge.org/vp2026/>, 2026.
- [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, “The VoicePrivacy 2020 challenge evaluation plan,” *arXiv preprint arXiv:2005.01387*, 2020.
- [5] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J.-F. Bonastre, “The VoicePrivacy 2022 challenge evaluation plan,” *arXiv preprint arXiv:2203.12468*, 2022.
- [6] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and M. Todisco, “The VoicePrivacy 2024 challenge evaluation plan,” *arXiv preprint arXiv:2404.02677*, 2024.
- [7] 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, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 28 492–28 518. [Online]. Available: <https://proceedings.mlr.press/v202/radford23a.html>
- [8] H. Hu, X. Zhu, T. He, D. Guo, B. Zhang, X. Wang, Z. Guo, Z. Jiang, H. Hao, Z. Guo *et al.*, “Qwen3-tts technical report,” *arXiv preprint arXiv:2601.15621*, 2026.
- [9] 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*, 2024, pp. 15 747–15 760.
- [10] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and

- N. Evans, "Speaker anonymisation using the McAdams coefficient," in *Proc. Interspeech*, 2021, pp. 1099–1103.
- [11] M. Panariello, F. Nespola, M. Todisco, and N. Evans, "Speaker anonymization using neural audio codec language models," in *Proc. ICASSP*, 2024, pp. 4725–4729.
- [12] P. Champion, "Speaker anonymization," Ph.D. dissertation, Université de Lorraine, 2023.
- [13] 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 *Proc. SPSC Symposium*, 2022.
- [14] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, 2021.
- [15] 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," *arXiv preprint arXiv:2409.08913*, 2024.
- [16] F. Lux *et al.*, "Meta learning text-to-speech synthesis in over 7000 languages," in *Proc. Interspeech*, 2024.
- [17] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, "Language-independent speaker anonymization approach using self-supervised pre-trained models," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 279–286.
- [18] B. M. L. Srivastava, M. Maouche, M. Sahidullah, E. Vincent, A. Bellet, M. Tommasi, N. Tomashenko, X. Wang, and J. Yamagishi, "Privacy and utility of x-vector based speaker anonymization," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 30, pp. 2383–2395, 2022.
- [19] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, 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.
- [20] S. Meyer, F. Lux, and N. T. Vu, "Probing the feasibility of multilingual speaker anonymization," in *Proc. Interspeech*, 2024, pp. 4448–4452.
- [21] J. Yao, Q. Wang, P. Guo, Z. Ning, Y. Yang, Y. Pan, and L. Xie, "MUSA: Multi-lingual speaker anonymization via serial disentanglement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 33, pp. 1664–1674, 2025.
- [22] S. Meyer, E. Kolos, and N. T. Vu, "First steps towards voice anonymization for code-switching speech," in *Proc. Interspeech*, 2025.
- [23] C. A. E. Goodhart, "Problems of monetary management: The UK experience," *Papers in Monetary Economics*, vol. 1, 1975.