

# Emotion-Aware Speaker Anonymization with Zero-Shot TTS Foundation Models for VoicePrivacy 2026

*Alef Iury Siqueira Ferreira<sup>1</sup>, Lucas Hideki Ueda<sup>2</sup>, Lucas Rafael Stefanel Gris<sup>1</sup>, Frederico Santos de Oliveira<sup>3</sup>, Augusto Seben da Rosa<sup>4</sup>, Alexandre Costa Ferro Filho<sup>1</sup>, Victor Moreno<sup>2</sup>*

<sup>1</sup> Federal University of Goiás (UFG), Brazil

<sup>2</sup> Universidade Estadual de Campinas (Unicamp), Brazil

<sup>3</sup> Federal University of Mato Grosso (UFMT), Brazil

<sup>4</sup> São Paulo State University (Unesp), Brazil

alef.iury.c.c@discente.ufg.br, lucashueda@dac.unicamp.br, lucas.gris@discente.ufg.br

## Abstract

We present an emotion-aware text-to-speech (TTS) anonymization pipeline that shows how the recent evolution of zero-shot voice-cloning models makes them a strong fit for speaker anonymization without any task-specific training. Our pipeline transcribes the input speech and regenerates it through foundation TTS models such as Qwen3-TTS and VoxCPM2, concealing the original speaker identity while faithfully preserving the spoken content. To keep emotional consistency, we use emotion-based embeddings to retrieve pseudo-speaker references from an emotionally diverse voice bank, while also exploring prompt-based voice design. Since different TTS models excel at different emotions, we further propose an emotion-specialist ensemble that aggregates their complementary strengths. Across both Track 1 and Track 2 of VoicePrivacy 2026, our training-free pipeline and ensemble reach a strong emotion-privacy trade-off under a semi-informed attacker while preserving intelligibility.

**Index Terms:** voice anonymization, text-to-speech, ensemble, foundation models

## 1. Introduction

The widespread adoption of voice-driven interfaces has raised significant privacy concerns because speech signals inherently encode linguistic content alongside sensitive personally identifiable information like speaker identity, age, gender, and affective state [1, 2]. Voice anonymization systems aim to conceal this biometric identity while preserving the underlying utility of the speech [3, 4]. Achieving this balance presents a fundamental utility-privacy trade-off since paralinguistic features such as prosody and emotional state are essential for conveying speaker intent but are deeply entangled with individual vocal characteristics, creating a potential vector for identity leakage [5, 6]. Consequently, aggressive anonymization approaches like cascaded ASR-TTS often degrade emotional and prosodic fidelity, whereas systems designed to maintain high emotional utility like Voice Conversion invariably suffer from residual identity exposure [7].

Addressing this complex multi-objective problem requires robust privacy measurement frameworks alongside flexible generation paradigms that offer continuous control over the retention of source paralinguistics [1, 3]. The VoicePrivacy Challenge (VPC) series provides shared tasks and objective metrics to evaluate these competing dimensions. Recent editions have increasingly prioritized emotion preservation, demanding that systems remove speaker cues without stripping emotional

expressiveness [5, 8, 7]. The 2026 iteration introduces even stricter conditions by incorporating stronger automatic speaker verification attackers, cross-gender trials in the English-focused Track 1, and language-wise attacker fine-tuning in the multilingual Track 2 [2].

Our work revisits the ASR-TTS anonymization paradigm by utilizing the current generation of zero-shot TTS foundation models. Recent advancements have yielded models capable of strong voice cloning, multilingual synthesis, and instruction-based controllability [9]. Specifically, Qwen3-TTS [10] enables short-reference voice cloning and description-based control, and VoxCPM2 [11] provides zero-shot cloning and natural-language voice design within a unified architecture. These capabilities allow TTS systems to produce a new utterance from recognized text conditioned on a pseudo-speaker reference or a synthetic voice prompt.

The primary contribution of this work is an emotion-aware TTS anonymization pipeline. We build a voice bank from emotionally diverse speech corpora and employ emotion2vec [12] to select pseudo-speaker references closely aligned with the source utterance in an emotion representation space. Our methodology evaluates random voice-bank sampling, top-K emotion-bank selection, prompt-based voice design, and an emotion-specialist ensemble. The results indicate that TTS-based anonymization achieves high privacy and linguistic accuracy, though emotion preservation remains a primary limitation. To partially mitigate this, our proposed emotion-specialist ensemble leverages the distinct strengths of different TTS systems in preserving specific emotional states.

## 2. Related Work

The first VoicePrivacy Challenge formalized voice anonymization as the task of altering the speaker voice to hide identity while preserving the spoken content and other useful attributes [5]. VPC 2022 strengthened the evaluation of semi-informed attacks and introduced additional utility constraints such as intonation preservation [8]. VPC 2024 added the explicit requirement of preserving emotional state through Unweighted Average Recall (UAR) using a Speech Emotion Recognition (SER) model and intelligibility through Word Error Rate (WER) from an Automatic Speech Recognition (ASR) model, while privacy was measured using Equal Error Rate (EER) achieved by an Automatic Speaker Verification (ASV) model [7]. VPC 2026 preserves this privacy-utility framing but introduces stronger attacker models and a multilingual anonymization track [2].

Prior systems include DSP-based anonymization, x-vector and speaker-embedding transformations, neural-codec methods, voice conversion, and cascaded ASR-TTS. Meyer et al. [13] proposed a cascade of phonetic recognition, GAN-generated speaker embeddings, and multispeaker speech synthesis. Other works improved pitch modeling or prosody anonymization to reduce identity leakage while preserving utility [14, 15]. Emotion-aware approaches became more prominent after VPC 2024. For instance, Hua et al. [16] integrated emotion encoding into pseudo-speaker generation, Lee et al. [6] used emotion-enriched features in an ASR-TTS anonymization pipeline, and Le Blouch et al. [17] showed the importance of expressive datasets and emotion embeddings for improving anonymized speech expressiveness.

Cascaded ASR-TTS systems can offer strong privacy protection because the original acoustic signal is discarded after transcription, and the output speech is resynthesized from text. However, this process can also suppress paralinguistic attributes that are not fully represented in the transcript, particularly emotion and prosody [4]. This privacy-utility trade-off motivates our approach: we exploit modern voice-cloning TTS models as privacy-preserving resynthesis backbones while using emotion representations to select pseudo-speaker references that are more likely to preserve affective cues.

Recent zero-shot TTS models have improved substantially in naturalness, speaker similarity, multilingual coverage, and controllability, achieving emerging capabilities in preserving emotional information from the prompt speech. Qwen3-TTS [10] is a multilingual and controllable TTS family with short-reference voice cloning and description-based control. VoxCPM2 [11] is a 2B-parameter multilingual and controllable speech generation model supporting natural-language voice design, controllable cloning, and continuation-based cloning. This recent progress motivates the use of TTS systems as anonymization backbones in VPC 2026.

Emotion preservation is a key utility requirement in recent VoicePrivacy evaluations, including VPC 2024 and VPC 2026. To address this requirement, we use emotion2vec, a self-supervised speech emotion representation model, to extract utterance-level emotion embeddings [12]. Specifically, for each input utterance, we retrieve emotionally similar reference utterances from a voice bank and use them as prompt speech in TTS synthesis. The underlying assumption is that conditioning the TTS model on an emotionally compatible reference can improve affective fidelity in the anonymized speech.

We further investigate an emotion-specialist ensemble motivated by recent evidence that zero-shot TTS systems exhibit emotion-dependent behavior. In particular, Shen et al. [18] show that the relative performance of TTS models can vary across emotion categories, indicating that a model that is strong for one emotion is not necessarily the best for another. This observation motivates our use of class-wise development performance to select a specialist anonymizer for each predicted emotion.

### 3. System Description

#### 3.1. Overview

Figure 1 shows the overall pipeline. Given an input utterance  $x$ , Whisper Large-v3 [19] produces a transcript  $t$ . In parallel, emotion2vec+ large produces an utterance-level emotion embedding  $\mathbf{z}_x$  and a predicted emotion label  $\hat{e}_x$ . The embedding is compared with embeddings precomputed for the voice-

bank utterances and is used to retrieve emotionally similar references. The predicted label is used to construct the VoxCPM2 voice-design instruction and to route utterances in the Track 1 emotion-specialist ensemble. The anonymized waveform is then generated either from the selected voice-bank reference and its transcript or from a natural-language voice-design prompt.

#### 3.2. Emotion-Aware Voice Bank

For Track 1, we build an English-focused voice bank using MSP-Podcast [20], the English subset of CAMEO [21], and MEAD [22]. These corpora were selected because they provide emotionally diverse speech references while remaining compatible with the resource constraints of the challenge. MSP-Podcast contributes spontaneous and naturalistic speech, whereas MEAD and CAMEO are closer to the target distribution of LibriSpeech [23], representing a more controlled and prepared speech domain. Together, these datasets form a diverse voice bank that covers both spontaneous and read-style speech, as well as a broad range of emotional conditions.

For Track 2, we build language-specific voice banks. For English, we use the same voice bank as in Track 1, composed of MSP-Podcast, the English subset of CAMEO, and MEAD. For French, German, and Spanish, we use the corresponding language subsets of CAMEO. This language-specific design is intended to improve content consistency during TTS generation. In reference-based voice cloning, the prompt speech provides not only speaker and emotional cues, but also linguistic context. When the language of the reference prompt does not match the target utterance, this mismatch can encourage unstable generation, including hallucinated words, code-switching, or deviations from the target transcript. Therefore, for the multilingual track, we select references from the same language as the target utterance, reducing language-conditioning mismatch while preserving emotion-aware pseudo-speaker selection.

For each candidate reference utterance  $r_i$  in the bank, we extract an emotion embedding  $\mathbf{z}_i$  using the emotion2vec in its emotion2vec+ large version<sup>1</sup>. The result is a bank  $\mathcal{B} = \{(r_i, t_i, e_i, \mathbf{z}_i)\}_{i=1}^N$  where  $t_i$  is the reference transcription,  $e_i$  is the emotion label, and  $\mathbf{z}_i$  is the emotion embedding.

For each input utterance  $x$ , we compute its emotion embedding  $\mathbf{z}_x$  and compare it against all bank embeddings using cosine similarity.

We sort the bank by similarity with the trial utterance and keep the top- $K$  candidates:

$$\mathcal{N}_K(x) = \text{TopK}_{r_i \in \mathcal{B}} \cos(\mathbf{z}_x, \mathbf{z}_i). \quad (1)$$

The final reference is sampled uniformly from this pool. This random top- $K$  strategy prevents the selection process from collapsing to the single nearest neighbor, while still biasing the reference choice toward emotionally compatible speech. As a result, it introduces emotional and speaker-reference diversity, which can further support anonymity by reducing the deterministic reuse of the same reference for similar utterances.

#### 3.3. Transcription System

We use Whisper Large-v3 [19] as the ASR front-end of our cascaded ASR-TTS anonymization pipeline. Each input utterance is first transcribed, and the resulting transcript is used as the content condition for the TTS model. We also transcribe

<sup>1</sup>[https://huggingface.co/emotion2vec/emotion2vec\\_plus\\_large](https://huggingface.co/emotion2vec/emotion2vec_plus_large)

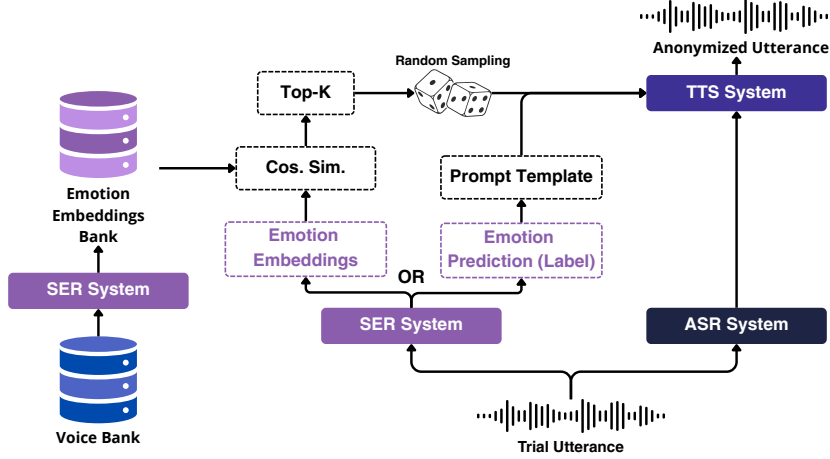


Figure 1: Overview of the proposed TTS-based anonymization pipeline. The *emotion2vec+* large model is used both to precompute voice-bank embeddings offline and to extract an emotion embedding and predicted emotion label from each input utterance at inference time. The embedding is used for emotion-similarity retrieval, while the predicted label is used for prompt-based voice design and emotion-specialist routing. Whisper Large-v3 provides the target transcript, and the TTS system receives either an emotion-matched reference utterance or a natural-language voice-design instruction.

the selected reference prompt speech and provide this reference transcript together with the reference audio to the TTS model, enabling a stronger voice-cloning condition based on both the pseudo-speaker acoustics and its corresponding linguistic content. Thus, the anonymized utterance is synthesized from the input transcript, while its voice characteristics are derived from the selected anonymizing reference. Whisper Large-v3 is selected due to its robustness and multilingual transcription capability, which are useful for both Track 1 and Track 2.

### 3.4. Text-to-Speech Anonymization Backbones

We evaluate two TTS backbones. The first is Qwen3-TTS 1.7B Base [10]<sup>2</sup>, used in reference-based voice cloning mode. The recognized text  $t$  is synthesized using a voice-bank reference selected by one of the retrieval strategies described above. We compare three reference selection strategies: random sampling from the full bank, top-1 emotion similarity, and random sampling from the top-50 emotion-similar candidates.

The second backbone is VoxCPM2 [11]<sup>3</sup>. We use it in two modes. In the reference-based mode, VoxCPM2 receives a selected voice-bank reference. In the voice-design mode, no reference audio is used. Instead, our pipeline builds a dynamic natural-language instruction from the predicted emotion label.

### 3.5. Prompt-based Voice Design

VoxCPM2 supports natural-language voice design, allowing the system to synthesize a pseudo-speaker directly from a textual description instead of relying on reference audio. We use this capability to generate emotion-conditioned anonymized speech. First, the predicted *emotion2vec* label is mapped to the IEMO-CAP [24] dataset emotion set used in the evaluation: *angry*, *happy*, *sad*, and *neutral*. Labels outside this set are mapped to the closest evaluation class; for instance, *disgusted* is mapped to *angry*, *fearful* to *sad*, and *surprised*, *other*, and unknown labels

to *neutral*.

For each utterance, we build a randomized voice-design instruction by sampling demographic and speaking-style attributes. The prompt always includes an age group, gender, delivery verb, and an emotion-preserving phrase containing the mapped emotion word, ensuring that the intended affective condition is explicitly provided to the TTS model. For non-neutral emotions, an intensity modifier is sampled uniformly from *none*, *clearly*, *strongly*, and *very*, corresponding to a 75% chance of adding an explicit intensity cue. For neutral speech, intensity modifiers are omitted to avoid unnatural prompts such as “very neutral”. In addition, an emotion-specific prosodic detail is added with 50% probability, using cues such as firm emphasis or sharper intonation for *angry*, bright intonation or lively energy for *happy*, slower pacing or subdued delivery for *sad*, and balanced pacing or even intonation for *neutral*. An example instruction is:

*A middle-aged male speaker is saying the line in a clearly angry voice with firm emphasis.*

The final VoxCPM2 voice design input is formed by concatenating the instruction with the target transcript using the format (instruction) text.

## 4. Experimental Setup

### 4.1. Challenge tracks and metrics

We follow the VoicePrivacy 2026 evaluation protocol. Track 1 evaluates English anonymization, while Track 2 evaluates multilingual anonymization over French, Spanish, English, and German. Privacy is measured by ASV equal error rate (EER), where higher values indicate stronger anonymization and 50% corresponds to chance-level verification. We also report the semi-informed ASV EER, which evaluates robustness against a stronger attacker with access to anonymized training data. Linguistic utility is measured by ASR word error rate (WER), where lower is better. Emotion preservation is measured by SER unweighted average recall (UAR), where higher values indicate better preservation of the affective state.

<sup>2</sup><https://huggingface.co/Qwen/Qwen3-TTS-12Hz-1.7B-Base>

<sup>3</sup><https://huggingface.co/openbmb/VoxCPM2>

Table 1: *Track 1 development and test comparison of the final anonymization systems on IEMOCAP in terms of UAR and class-wise accuracy (%).* Blue cells and values in **bold** indicate the selected specialist for each emotion in the final ensemble.

| System              | Development  |             |              |              |              | Test         |             |              |              |              | Selected for   |
|---------------------|--------------|-------------|--------------|--------------|--------------|--------------|-------------|--------------|--------------|--------------|----------------|
|                     | UAR ↑        | SAD ↑       | NEU ↑        | ANG ↑        | HAP ↑        | UAR ↑        | SAD ↑       | NEU ↑        | ANG ↑        | HAP ↑        |                |
| Qwen3-TTS + Top-50  | <b>35.90</b> | 4.29        | <b>49.33</b> | <b>57.38</b> | 32.58        | <b>34.98</b> | 2.89        | <b>53.33</b> | <b>50.17</b> | 33.54        | Neutral, Angry |
| VoxCPM2 + Top-50    | 26.10        | <b>4.47</b> | 24.55        | 4.07         | 71.33        | 26.72        | <b>6.65</b> | 26.45        | 2.24         | 71.55        | Sad            |
| VoxCPM2 VoiceDesign | 25.25        | 4.45        | 14.84        | 4.31         | <b>77.41</b> | 25.57        | 4.30        | 14.31        | 6.22         | <b>77.46</b> | Happy          |

Table 2: *Track 1 test set results. Higher is better for SER UAR, ASV EER, ASV Semi EER, and emotion–privacy H-mean (Trade-off); lower is better for WER. The emotion–privacy H-mean (Trade-off) is computed as the harmonic mean between SER UAR and ASV Semi EER. Bold indicates the best result within each group, underline indicates the second-best result within each group, and blue cells indicate the best overall result for each metric.*

| System                        | SER UAR ↑    | ASV EER ↑    | WER ↓       | ASV Semi EER ↑ | Trade-off ↑  |
|-------------------------------|--------------|--------------|-------------|----------------|--------------|
| <i>Challenge Baselines</i>    |              |              |             |                |              |
| B2 – McAdams DSP              | <b>53.49</b> | 7.22         | 9.96        | 4.71           | 8.66         |
| B3 – TTS                      | 35.24        | <u>23.18</u> | <b>4.31</b> | <u>16.85</u>   | <u>22.80</u> |
| B4 – Neural audio codec       | <u>42.78</u> | 17.76        | 5.90        | 14.33          | 21.47        |
| B5 – ASR-based                | 38.25        | <b>28.87</b> | <u>4.44</u> | <b>21.28</b>   | <b>27.35</b> |
| <i>Proposed TTS Pipelines</i> |              |              |             |                |              |
| Qwen3-TTS + Random            | 31.50        | 47.94        | 5.23        | 34.60          | 32.98        |
| Qwen3-TTS + Top-1             | 33.33        | 47.73        | 8.76        | 34.30          | 33.81        |
| Qwen3-TTS + Top-50            | <u>34.98</u> | 47.34        | 4.24        | 35.18          | 35.08        |
| VoxCPM2 + Top-50              | 26.72        | <b>49.55</b> | 5.23        | <b>51.64</b>   | <u>35.22</u> |
| VoxCPM2 + VoiceDesign         | 25.57        | 47.45        | <b>2.67</b> | <u>45.69</u>   | 32.79        |
| Emotion Specialist Ensemble   | <b>42.71</b> | 44.53        | <u>4.01</u> | 36.91          | <b>39.60</b> |

To summarize the balance between emotion-category preservation and anonymization robustness, we report an emotion-privacy harmonic mean between SER UAR and semi-informed ASV EER:

$$H_{\text{emo-priv}} = \frac{2 \cdot \text{SER UAR} \cdot \text{ASV Semi EER}}{\text{SER UAR} + \text{ASV Semi EER}} \quad (2)$$

The harmonic mean is used because both components are higher-is-better and represent competing objectives: SER UAR measures preservation of the target emotion category, whereas ASV Semi EER measures anonymization robustness against the stronger semi-informed attacker. It therefore penalizes systems that perform well on only one dimension. SER UAR is used as the utility term because emotion preservation is the primary bottleneck addressed in this work. Cascaded ASR–TTS systems often retain linguistic content through the intermediate transcript, while prosodic and emotional cues, which are not fully encoded in text, may be lost during resynthesis. WER is thus reported separately, and the resulting score should be interpreted as an emotion–privacy diagnostic rather than a complete privacy–utility measure.

## 4.2. Implementation Details

For the reference-based Qwen3-TTS and VoxCPM2 systems, a single anonymizing reference is selected for each input utterance according to the corresponding sampling strategy with our proposed pipeline. The selected reference audio, its transcript, and the target transcript are then provided to the TTS backbone to synthesize the anonymized waveform. In the VoxCPM2 voice-design setting, no reference audio is used; instead, the target transcript is conditioned on the generated natural-language voice instruction described above.

For VoxCPM2, both reference-based voice cloning and prompt-based voice design are decoded using a classifier-free guidance value of 2.0, 10 inference timesteps, and temperature 1.0. For Qwen3-TTS, we use temperature 0.9, top- $k$  sampling with  $k = 50$ , top- $p = 1.0$ , and a repetition penalty of 1.05. These decoding settings are kept fixed across all experiments for each backbone to ensure that differences between systems mainly reflect the reference-selection or voice-design strategy rather than changes in generation hyperparameters.

## 4.3. Computational Cost

The submitted systems require no task-specific parameter training or fine-tuning. They instead compose pretrained ASR, emotion-representation, and TTS models at inference time. The anonymized speech for both Qwen3-TTS and VoxCPM2 was generated using a single NVIDIA L40S GPU. The computational cost is therefore dominated by ASR transcription, emotion embedding extraction, voice-bank retrieval, and TTS waveform generation.

## 4.4. Submitted Systems and Challenge Compliance

The proposed systems follow the VPC 2026 utterance-level anonymization requirement. Each input utterance is anonymized independently, without using speaker labels. For the reference-based systems, the pseudo-speaker reference is selected per utterance from the corresponding voice bank using either random sampling or emotion-similarity-based retrieval. The same anonymization configuration is applied to the development, evaluation, and attacker-training data required by each track.

Table 3 maps each submitted anonymized-data suffix to the corresponding system variant described in this paper. The Track 1 submission corresponds to the emotion-specialist en-

semble, while the Track 2 submission corresponds to VoxCPM2 with random sampling from the top-50 emotion-similar references in the language-specific voice bank.

## 5. Results and Discussion

### 5.1. Track 1

Table 1 analyzes the emotion-wise behavior of the final anonymization systems for Track 1 on IEMOCAP. The results show that no single TTS configuration is consistently optimal across all emotion classes. Qwen3-TTS + Top-50 achieves the highest UAR among the individual systems and performs best for neutral and angry utterances on both development and test splits. VoxCPM2 + Top-50 is the strongest option for sad utterances, although this class remains difficult for all systems. Sad speech remains difficult for all Track 1 systems, with very low class-wise accuracy. This may reflect errors in emotion2vec prediction, limited sad references, domain mismatch, or weak transfer of low-arousal prosodic cues by the TTS backbones. In contrast, VoxCPM2 VoiceDesign is clearly superior for happy speech, suggesting that prompt-based control can encourage more expressive synthesis when the target emotion is explicit and easy to represent or verbalize.

This complementary behavior motivates the proposed emotion-specialist ensemble. Instead of relying on a single TTS backbone to generate all emotions, the ensemble selects the anonymization system according to the predicted emotion label. The specialist assigned to each emotion is chosen using development-set performance: Qwen3-TTS + Top-50 is selected for neutral and angry utterances, VoxCPM2 + Top-50 for sad utterances, and VoxCPM2 VoiceDesign for happy utterances. The same assignment is then applied at test time, allowing each emotion to be handled by the system that best preserved it during development.

Table 2 reports the main Track 1 test results and illustrates the expected tension between emotion preservation and anonymization among the challenge baselines. B2 retains the most emotion information but offers weak protection against the semi-informed attacker, whereas the remaining baselines improve privacy at the cost of lower SER UAR. B5 provides the strongest baseline emotion-privacy balance, but all proposed TTS systems substantially improve semi-informed privacy.

The proposed TTS pipeline provides substantially stronger anonymization than the official baselines. Qwen3-TTS + Random reaches 47.94% ASV EER and 34.60% ASV Semi EER, highlighting the privacy advantage of the ASR-TTS resynthesis paradigm. However, its emotion preservation remains limited. Emotion-aware reference selection improves this balance, as Qwen3-TTS + Top-50 obtains the best UAR among the individual proposed systems, reaching 34.98%, while preserving low WER at 4.24% and strong anonymization. This result supports the hypothesis that emotionally similar references can improve affective preservation relative to random reference sampling in this configuration.

A relevant observation is that Qwen3-TTS + Random does not reach chance-level semi-informed EER, even though the original waveform is discarded and the anonymizing reference is sampled without emotion-based matching and independently of the source-speaker identity. This suggests that residual linkability is not determined solely by the pseudo-speaker timbre. Prior work has shown that speaker identity may leak through lexical and semantic content patterns due to speaker-specific vocabularies [25, 26]. Linguistic-content leakage is therefore

one plausible contributor to the observed result. Other possible contributors include speaker-dependent ASR error patterns and stable artifacts introduced by the synthesis model.

The VoxCPM2 systems exhibit a different trade-off profile in Track 1. VoxCPM2 + Top-50 achieves the strongest anonymization, reaching 49.55% ASV EER and 51.64% ASV Semi EER, but its lower SER UAR limits emotion preservation on IEMOCAP. This result indicates that strong zero-shot TTS systems can differ substantially in their voice-cloning behavior depending on the dataset, emotion distribution, and reference-conditioning setup. In this Track 1 setting, VoxCPM2 appears less consistent in transferring prosodic and emotional cues for angry and neutral utterances, as shown in Table 1. However, this should not be interpreted as a general limitation of VoxCPM2, since the model remains highly effective for privacy and performs strongly for other emotional conditions. Rather, these results motivate the use of an emotion-specialist ensemble when different systems show complementary emotion-wise strengths. VoxCPM2 VoiceDesign obtains the best intelligibility, with 2.67% WER, but also shows low overall UAR, suggesting that prompt-only voice design improves lexical consistency without reliably preserving all emotion cues. A possible explanation is that VoiceDesign does not require a reference audio prompt or an associated reference transcript. This simplifies the generation process, since the model only needs to follow the target transcript and the natural-language voice instruction, avoiding potential mismatches introduced by conditioning on both a voice prompt and its transcript. However, this advantage comes at the cost of weaker emotion preservation across classes, indicating that textual voice descriptions alone may be insufficient to represent the full range of emotional and prosodic variation required by the task.

The emotion-specialist ensemble provides the best overall balance among the Track 1 systems. It increases SER UAR to 42.71%, substantially improving emotion preservation over all individual proposed systems, while maintaining strong privacy with 44.53% ASV EER and 36.91% ASV Semi EER. This leads to the highest emotion-privacy trade-off, 39.60, indicating that class-wise specialist selection can recover part of the emotional expressiveness lost by individual TTS configurations without discarding the privacy benefits of ASR-TTS resynthesis. Based on this balance between anonymization robustness and affective consistency, the emotion-specialist ensemble was selected as our Track 1 submission.

### 5.2. Track 2

Table 4 reports the Track 2 multilingual results. Compared with the challenge baselines, the proposed TTS systems achieve a stronger balance between anonymization and emotion preservation. BM1 obtains the best baseline SER UAR, but provides weak privacy, especially under the semi-informed attacker, which leads to a low emotion-privacy trade-off. BM3 follows the opposite pattern, achieving much stronger ASV and ASV Semi EER values across languages, but with substantially lower SER UAR. This confirms that, for the official baselines, stronger anonymization often comes at the cost of degraded emotion preservation.

Among the proposed systems, VoxCPM2 + Top-50 achieves the best overall Track 2 performance. It obtains the highest SER UAR, 56.34%, the strongest average ASV EER, 46.76%, the lowest average WER, 7.85%, and the highest average ASV Semi EER, 43.90%. This leads to the best emotion-privacy trade-off, 49.35. Qwen3-TTS + Top-50 remains com-

Table 3: Mapping between submitted anonymized-data suffixes and system variants.

| Track   | \$anon.data.suffix             | Corresponding system variant                                                                |
|---------|--------------------------------|---------------------------------------------------------------------------------------------|
| Track 1 | _ensemble2                     | Emotion Specialist Ensemble                                                                 |
| Track 2 | _voxcpm2_nanovllm_emosim_top50 | VoxCPM2 + Top-50 with language-specific voice banks and emotion-similar reference selection |

Table 4: Track 2 results by language and averaged over English (EN), Spanish (ES), German (DE), and French (FR). Higher is better for SER UAR, ASV EER, ASV Semi EER, and emotion-privacy H-mean (Trade-off), and lower is better for WER. The emotion-privacy H-mean (Trade-off) is computed as the harmonic mean between SER UAR and ASV Semi EER and does not include WER. **Bold** indicates the best result within each group, underline indicates the second-best result within each group, and **blue** cells indicate the best overall result for each metric/language.

| System                 | SER UAR ↑ | ASV EER ↑ |       |       |       |       | WER ↓ |       |       |       |       | ASV Semi EER ↑ |       |       |       |       | Trade-off ↑ |
|------------------------|-----------|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------------|-------|-------|-------|-------|-------------|
|                        |           | EN        | ES    | DE    | FR    | AVG   | EN    | ES    | DE    | FR    | AVG   | EN             | ES    | DE    | FR    | AVG   |             |
| Challenge Baselines    |           |           |       |       |       |       |       |       |       |       |       |                |       |       |       |       |             |
| BM1                    | 42.60     | 5.60      | 9.10  | 11.40 | 12.00 | 9.53  | 9.30  | 7.20  | 11.60 | 10.10 | 9.55  | 3.70           | 1.70  | 1.60  | 2.50  | 2.38  | 4.50        |
| BM2                    | 32.40     | 13.70     | 27.60 | 29.00 | 30.10 | 25.10 | 17.10 | 34.10 | 32.40 | 44.20 | 31.95 | 8.50           | 10.00 | 12.10 | 16.60 | 11.80 | 17.30       |
| BM3                    | 24.80     | 37.60     | 49.00 | 48.30 | 48.50 | 45.85 | 13.60 | 16.30 | 31.30 | 22.50 | 20.93 | 22.70          | 45.70 | 47.30 | 46.50 | 40.55 | 30.78       |
| Proposed TTS Pipelines |           |           |       |       |       |       |       |       |       |       |       |                |       |       |       |       |             |
| Qwen3-TTS + Top-50     | 49.30     | 39.69     | 47.20 | 46.36 | 46.55 | 44.95 | 14.25 | 6.49  | 19.17 | 9.46  | 12.34 | 29.82          | 44.33 | 42.09 | 43.82 | 40.02 | 44.17       |
| VoxCPM2 + Top-50       | 56.34     | 47.95     | 46.90 | 45.10 | 47.08 | 46.76 | 7.26  | 5.38  | 10.74 | 8.00  | 7.85  | 44.65          | 44.92 | 41.44 | 44.60 | 43.90 | 49.35       |

Table 5: Track 2 development and test emotion preservation results on EmoData in terms of UAR and class-wise accuracy (%). Original speech is shown as an upper-bound reference. **Blue** cells indicate the best anonymized system for each split and metric.

| System             | Development  |              |              |              |              | Test         |              |              |              |              |
|--------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                    | UAR ↑        | HAP ↑        | NEU ↑        | ANG ↑        | SAD ↑        | UAR ↑        | HAP ↑        | NEU ↑        | ANG ↑        | SAD ↑        |
| Ground Truth       | 72.09        | 75.61        | 82.31        | 72.33        | 58.10        | 78.90        | 68.08        | 90.64        | 83.40        | 73.50        |
| Qwen3-TTS + Top-50 | 44.44        | <b>39.84</b> | 69.68        | 35.06        | 33.20        | 49.30        | 36.92        | 74.89        | 39.25        | 46.15        |
| VoxCPM2 + Top-50   | <b>52.17</b> | 35.77        | <b>74.73</b> | <b>53.51</b> | <b>44.66</b> | <b>56.34</b> | <b>44.62</b> | <b>75.75</b> | <b>55.85</b> | <b>49.15</b> |

petitive in terms of privacy, but its higher average WER and lower SER UAR indicate weaker multilingual robustness compared with VoxCPM2.

Unlike Track 1, Track 2 exhibits limited emotion-wise complementarity between the evaluated systems. As shown in Table 5, VoxCPM2 + Top-50 achieves the highest overall development UAR and the best class-wise accuracy for neutral, angry, and sad speech, whereas Qwen3-TTS + Top-50 is only marginally better for happy speech. A specialist ensemble would therefore differ from VoxCPM2 for only one emotion category while introducing additional sensitivity to emotion-prediction errors during routing. We consequently selected VoxCPM2 + Top-50 as the simpler and more robust Track 2 configuration. This choice is further supported by the language-wise results: VoxCPM2 + Top-50 achieves the lowest WER in all four languages and strong ASV Semi EER values, indicating effective multilingual content preservation and privacy under the stronger semi-informed attacker. The test results are also consistent with this decision, as VoxCPM2 + Top-50 obtains the highest anonymized accuracy for every emotion class.

### 5.3. Limitations

Because the original waveform is discarded, the system preserves the recognized linguistic content and coarse emotion categories rather than the source utterance’s exact prosody, timing, and other fine-grained acoustic characteristics. Performance may also be affected by error propagation across the multi-model inference pipeline, including transcription errors from Whisper and emotion-classification errors from emotion2vec.

Finally, although the proposed approach does not require task-specific parameter training, it relies on several large pretrained models and therefore may incur non-negligible computational and inference costs.

## 6. Conclusion

We revisited the ASR-TTS anonymization paradigm in light of the recent emergence of zero-shot voice-cloning foundation models. The proposed system resynthesizes speech from ASR transcriptions using modern TTS backbones, while emotion-based embeddings guide the selection of emotionally compatible pseudo-speaker references from a diverse voice bank. Motivated by evidence that zero-shot TTS systems can behave differently across emotions, we also explored an emotion-specialist ensemble that selects the anonymizer according to the predicted emotion class. Our results show that current TTS foundation models can perform effective speaker anonymization without task-specific training or highly complex pipelines, reaching strong ASV EER and low WER across both VoicePrivacy 2026 tracks. However, preserving emotion and fine-grained prosody remains challenging. In Track 1, the emotion-specialist ensemble achieved the best emotion-privacy trade-off, while in Track 2, VoxCPM2 + Top-50, which samples from the 50 voice-bank references most emotionally similar to the trial utterance, provided the strongest multilingual balance between privacy, intelligibility, and emotion preservation. Overall, these findings indicate that zero-shot TTS-based resynthesis is a practical and promising direction for speaker anonymization, especially when combined with emotion-aware reference selection.

## 7. References

- [1] S. Zhang, Z. Li, and A. Das, “Voicepm: A robust privacy measurement on voice anonymity,” in *Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, ser. WiSec ’23. New York, NY, USA: Association for Computing Machinery, 2023, p. 215–226. [Online]. Available: <https://doi.org/10.1145/3558482.3590175>
- [2] 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,” VoicePrivacy Challenge, Tech. Rep., 2026, version 1.1.
- [3] I. R. Ulgen, Z. Cai, N. Andrews, P. Koehn, and B. Sisman, “Diff-fanon: Diffusion-based prosody control for voice anonymization,” 2026. [Online]. Available: <https://arxiv.org/abs/2604.26281>
- [4] Z. Cai, H. L. Xinyuan, A. Garg, L. P. García-Perera, K. Duh, S. Khudanpur, N. Andrews, and M. Wiesner, “Privacy versus emotion preservation trade-offs in emotion-preserving speaker anonymization,” in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 409–414.
- [5] 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.
- [6] J. Lee, T. Park, and Y. You, “Voice anonymization using emotion-enriched feature integration with stt and tts models,” *Proc. SPSC*, pp. 50–54, 2024.
- [7] 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,” *arXiv preprint arXiv:2601.11846*, 2026.
- [8] 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.
- [9] H. Azzuni and A. E. Saddik, “Voice cloning: Comprehensive survey,” *arXiv preprint arXiv:2505.00579*, 2025.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] S. Meyer, P. Tilli, F. Lux, P. Denisov, J. Koch, and N. T. Vu, “Cascade of phonetic speech recognition, speaker embeddings gan and multispeaker speech synthesis for the voiceprivacy 2022 challenge,” in *Proc. 2nd Symposium on Security and Privacy in Speech Communication*, 2022.
- [14] Ü. E. Gaznepoglu, A. Leschanowsky, and N. Peters, “Voiceprivacy 2022 system description: speaker anonymization with feature-matched f0 trajectories,” *arXiv preprint arXiv:2210.17338*, 2022.
- [15] J. He, J. Zhou, H. Sun, H. Wang, and Y. Qin, “Facodec-based anonymization solution enhanced with prosody anonymization,” *The VoicePrivacy*, 2024.
- [16] H. Hua, Z. Shang, X. Li, P. Shi, C. Yang, L. Wang, and P. Zhang, “Emotional speech anonymization: Preserving emotion characteristics in pseudo-speaker speech generation,” in *Proc. SPSC 2024*, 2024, pp. 55–60.
- [17] O. Le Blouch, R. Bakari, and N. Gengembre, “Tuning dissc for voice privacy challenge 2024,” *Proc. SPSC*, 2024.
- [18] S. Shen, D. Wu, X. Song, D. Zhou, L. Xue, M. Meng, J. Luan, and S. Wang, “Iterate to differentiate: Enhancing discriminability and reliability in zero-shot tts evaluation,” *arXiv preprint arXiv:2603.24430*, 2026.
- [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] C. Busso, R. Lotfian, K. Sridhar, A. N. Salman, W.-C. Lin, L. Goncalves, S. Parthasarathy, A. R. Naini, S.-G. Leem, L. Martinez-Lucas *et al.*, “The msp-podcast corpus,” *arXiv preprint arXiv:2509.09791*, 2025.
- [21] 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.
- [22] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, “Mead: A large-scale audio-visual dataset for emotional talking-face generation,” in *European conference on computer vision*. Springer, 2020, pp. 700–717.
- [23] 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.
- [24] 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.
- [25] Ü. E. Gaznepoglu, A. Leschanowsky, A. Aloradi, P. Singh, D. Tenbrinck, E. A. Habets, and N. Peters, “You are what you say: Exploiting linguistic content for voiceprivacy attacks,” *arXiv preprint arXiv:2506.09521*, 2025.
- [26] 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.