



# Universal Semantic Disentangled Privacy-preserving Speech Representation Learning

*Biel Tura-Vecino, Subhadeep Maji, Aravind Varier, Antonio Bonafonte, Ivan Valles, Michael Owen, Constantinos Papayiannis, Leif Radel, Grant Strimel, Oluwaseyi Feyisetan, Roberto Barra-Chicote, Ariya Rastrow, Volker Leutnant, Trevor Wood*

Amazon, AGI

{bieltura, msubhade, avarier, bonafont}@amazon.com

## Abstract

The use of human speech to train LLMs poses privacy concerns due to these models' ability to generate samples that closely resemble artifacts in the training data. We propose a speaker privacy-preserving representation learning method through the Universal Speech Codec (USC), a computationally efficient codec that disentangles speech into: (i) privacy-preserving semantically rich representations, capturing content and speech paralinguistics, and (ii) residual acoustic and speaker representations that enable high-fidelity reconstruction. Evaluations show that USC's semantic representation preserves content, prosody, and sentiment, while removing identifiable traits. Additionally, we present an evaluation methodology for measuring privacy-preserving properties. We compare USC against other speech codecs and demonstrate its effectiveness on privacy-preserving representation learning, showcasing the trade-offs between speaker anonymization and paralinguistics retention.<sup>1</sup>  
**Index Terms:** representation learning, speech disentanglement

## 1. Introduction

Foundational multimodal Large-Language Models (LLMs) require massive amounts of training data. Speech and audio are essential modalities for many applications, and multimodal LLMs require exposure to them during their training process [1]. Speech is a form of individual information [2], and the development of new foundational speech-aware LLMs demands access to massive amounts of speech data to fully unlock their potential. The research community has collected, curated and released huge amounts of data over the past decades. However, in the realm of Responsible AI (RAI), every individual and organization must make proper use of individuals' data when training such speech foundational models, regardless of their public availability. Hence, it is imperative to develop privacy-preserving methods that enable advancing the state-of-the-art of speech LLMs in a manner that safeguards individual privacy.

LLMs trained on language modeling tasks model the likelihood of generating coherent text sequences from a distribution of discrete tokens. This allows them to produce expressive and varied responses during generation. Incorporating continuous signals, such as speech, into multimodal training objectives presents a representation challenge that is circumvented by discretizing the distribution of the continuous space [3]. Consequently, the model prediction quality is constrained by how well the target representations encode information from the data [4]. For natural-sounding speech modeling, these representations are required to capture rich semantic information, including content and paralinguistic information (such as prosody

<sup>1</sup>Audio samples can be found in the science blogpost at <https://www.amazon.science/usc-samples>

and sentiment) [5]. However, from a privacy perspective, they should not encapsulate any characteristic that enable individual identification. We refer to these as semantic privacy-preserving representations, which aim to capture the maximum semantic information while disentangling it from the speaker's identity.

In this study, we present the Universal Speech Codec (USC), a codec architecture that tokenizes speech into privacy-preserving discrete representations tailored for speech-aware LLMs. USC learns semantically meaningful discrete representations that capture speech content and paralinguistics such as pacing, emphasis, and sentimental aspects, while also learning the additionally required speaker residual representations for reconstructing the original waveform. Motivated by [6], we introduce a speaker privacy-preserving representation learning method with enhanced paralinguistic and anonymization biases. In addition to semantic distillation, we include a specific speaker classifier gradient reversal [7], the quantizer dropout technique [8], and the usage of Local Differential Privacy (LDP) [9] to further bias the semantic representations.

We benchmark our approach against different open-source alternatives and show that USC's semantic representations have good content preservation and low speaker-specific characteristics while encoding a large amount of paralinguistic information. Moreover, the residual representations augments the semantic ones with the remaining speaker attributes, reporting state-of-the-art metrics on up-sampling speech waveform reconstruction. Additionally, we present a set of metrics and requirements, backed up by human perceptual evaluations, to assess privacy-preservation of learned semantic representations through the speech  $k$ -anonymity factor.

In this work we show:

1. A speaker privacy-preserving representation learning method based on the USC model, that disentangles speech semantics from speaker-identifiable traits, surpassing all baselines in jointly encoding content and paralinguistic information.
2. An ensemble of speaker disentanglement techniques focused on privacy through the addition of Local Differential Privacy for scalable privacy-preserving representation learning.
3. A new privacy-preserving evaluation that defines a set of metrics to assess the level of anonymization in speech representations, backed by human perceptual subjective tests.

## 2. Method

The proposed model (Figure 1) is based on a modified version of DAC [8]. It encodes speech into discrete residual representations and decodes them back to the original waveform. The disentanglement module biases the representations to capture semantics without speaker-specific traits. Only the first quantizer is biased to obtain a single non-residual semantic codebook.

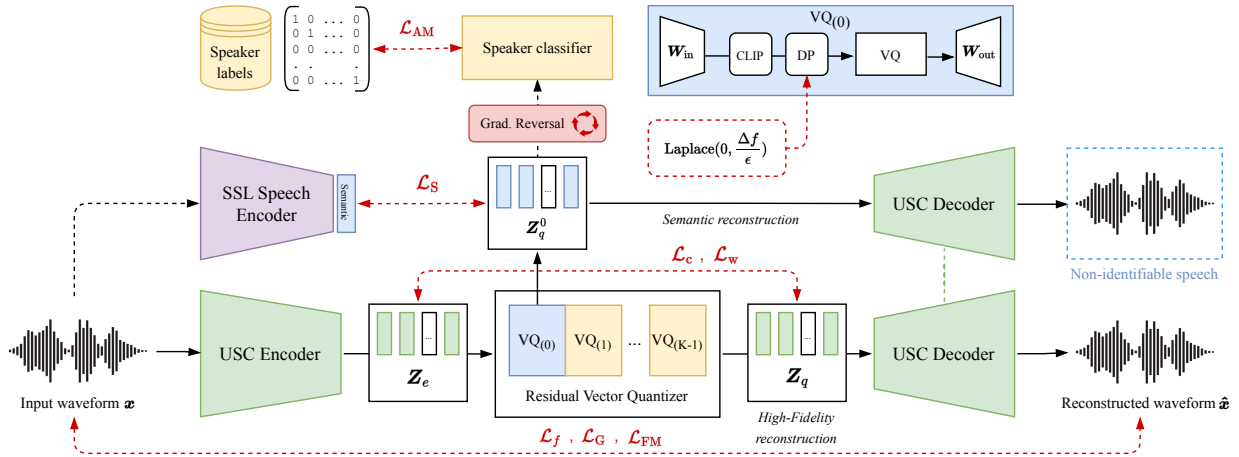


Figure 1: USC architecture. Dashed lines denote training and continuous the inference for high-fidelity and semantic reconstruction.

## 2.1. Universal Speech Codec

**Encoder & Decoder:** The encoder performs temporal down-sampling of the waveform  $x$  through strided and residual convolutional blocks to obtain the encoded representations  $z_e$ . The decoder mirrors the encoder and reconstructs the waveform  $\hat{x}$  from the quantized representation  $z_q$  of  $z_e$ . Following the work in [10], we use log-scale Snake-beta [11] activations and we remove the final  $\tanh$  activation to reduce harmonic distortions.

**Residual Vector Quantizer:** In RVQ, multiple  $K$  Vector Quantizers (VQ) are employed in a hierarchical manner to achieve a more fine-grained quantized representation  $z_q$  of the input latent  $z_e$  [12]. The input vector is first quantized using the initial quantizer  $VQ_{(0)}$ , and the difference between the input and the quantized representation is then recursively discretized using the subsequent codebooks. We employ factorized and  $L_2$ -normalized codes [13], which include an input  $W_{in}$  and output  $W_{out}$  projection in the quantizer to improve codebook usage.

## 2.2. Speaker reversal

It is the first component that removes speaker-specific information from the semantic representations. It consists of a cross-entropy based speaker classifier that uses a stack of transformer encoder layers as a speaker extractor and a gradient reversal layer [7]. The speaker classifier is trained to identify the speaker from the semantic representations. Then, we reverse the computed gradients during backpropagation [14] to suppress speaker identifiable information. The speaker classifier’s output is projected to a finite number  $N$  of known speakers, and it is trained with the AM Softmax loss,  $\mathcal{L}_{AM}$  [15], with an additive margin of  $m = 0.4$  and a constant scaling factor of  $s = 30$ .

## 2.3. Semantic distillation

Biasing the representations with speaker reversal alone leads to heavy degradation of meaningful semantic information. The simplest solution is to destroy enough information from the semantic codebook to remove identifiable information. Therefore, we apply a bias to the representations via distillation from a SSL Speech Encoder [6]. Specifically, we select the 9th layer of the multilingual HuBERT [16] to extract semantic targets, as it contains rich semantic information without speaker-identifiable traits [17]. We apply the continuous DistillHuBERT loss,  $\mathcal{L}_S$  [18], as distillation objective on the representations  $z_q^0$ .

## 2.4. Quantizer dropout

Waveform reconstruction and perceptual loss contains rich semantic information, but they are highly entangled with speaker-specific characteristics. To leverage this for semantic learning, we use quantizer dropout [12], which enables variable bit-rate capabilities during training. We apply quantizer dropout with a probability of  $p = 0.5$ , ensuring that the decoder reconstructs faithful waveform at different levels of the RVQ, and lead it to learn from the most to the least significant information with each additional residual quantizer. However, to avoid speaker leakage into the semantic codebook, we limit its influence by stopping the gradients from propagating to the encoder when the semantic quantizer is being used alone during training. By doing so, we train the decoder to reconstruct faithful speech from semantic representations while propagating the perceptual loss to just the decoder and the quantizer, which encourages the semantic codebook to capture relevant paralinguistics.

## 2.5. Local differential speaker privacy

Speaker gradient reversal technique does not guarantee that information is being reliably removed from the semantic representations for an unseen identity. To ensure stronger guarantees of speaker information removal, we employ tools from Local Differential Privacy (LDP) [9]. LDP protects the privacy of individual records and provides strong theoretical guarantees on anonymization. We employ a widely used variant of LDP known as the Laplace mechanism, which anonymizes a function  $f$  by adding Laplace noise. The noise is controlled by a hyper-parameter  $\epsilon$  and the  $L_1$  sensitivity,  $\Delta f$ . We apply the Laplace noise block to the semantic quantizer as shown in Figure 1. We clip the norm of the projection output to  $C$ , which results in an  $L_1$  sensitivity upper bounded by  $\Delta f = 2C$ . The clipping value  $C$  was estimated by computing the average of the  $L_1$ -norm over several training batches from a USC variant without the Laplace noise block. During training we add the samples noise  $n \sim \text{Laplace}(0, 2C/\epsilon)$  to the output of the quantizer projection layer. The smaller the value of  $\epsilon$ , the more spread out the Laplace distribution is. The choice of hyper-parameter  $\epsilon$  dictates the degree of privacy-utility tradeoff. Privacy is quantified as speaker re-identification accuracy and utility as speaker fidelity of the generated speech. During inference we omit the noise block [19]. We show the impact of LDP in Section 3.4.

## 2.6. Training objectives

**Reconstruction Loss:**  $\mathcal{L}_f$  follows [8] through a combination of multi-scale spectral losses, computed as the L1 distance between the multiple scales of mel-spectrograms.

**Perceptual loss:** GAN-based loss as a combination of a multi-period MPD and a multi-band multi-resolution MB-MRSD [8] discriminators. The latter helps suppress high-frequency aliasing artifacts from the upsampling decoder. We use least squares adversarial loss  $\mathcal{L}_G$ , and L1 feature matching loss,  $\mathcal{L}_{FM}$ .

**Discretization loss:** The RVQ is trained with both codebook  $\mathcal{L}_c$  and commitment  $\mathcal{L}_w$  loss functions with straight-through gradient estimation [3]. The commitment loss encourages the encoder’s output to be close to the quantized value while the codebook loss encourages the codewords to better represent the data distribution by minimizing the output-codeword distance. The final objective is the  $\lambda$ -weighted balanced loss  $\mathcal{L}$  over:

$$\underbrace{\lambda_f \mathcal{L}_f + \lambda_G \mathcal{L}_G + \lambda_{FM} \mathcal{L}_{FM}}_{\text{Reconstruction + Perceptual}} + \underbrace{\lambda_c \mathcal{L}_c + \lambda_w \mathcal{L}_w}_{\text{Codebook + Commit.}} + \underbrace{\lambda_{AM} \mathcal{L}_{AM} + \lambda_S \mathcal{L}_S}_{\text{Speaker disentanglement}} \quad (1)$$

## 3. Experiments and results

### 3.1. Experimental setup

**Datasets:** We used the same custom speech dataset as in [20], which consisted of more than 100K hours of public domain multilingual speech data, with English being the dominant one. We further added more than 1K hours of internal labeled studio-quality speakers. We ensured that 20% of the samples in a batch were from this labeled set to train the speaker classifier. For objective evaluation, we used 10 different internal speakers with various speaking styles: excited, cheerful, mindful, conversational, and long-form. For the privacy-preservation evaluation, we took a labeled pool of 7974 speakers from different sources.

**Training:** first, a 16 kHz USC variant is trained from scratch for 1M steps to leverage the maximum available speech data. Following Equation 1, the loss is weighted with  $\lambda_f = 15$ ,  $\lambda_{GAN} = 1$ ,  $\lambda_{FM} = 2.0$ ,  $\lambda_c = 1$  and  $\lambda_w = 0.25$ . For speaker disentanglement we set  $\lambda_{AM} = 25$  and  $\lambda_S = 45$ . For LDP,  $\epsilon = 15$  provided the best privacy-utility trade-off. For high-quality speech, a new up-sampling decoder is trained on 24 kHz filtered data by freezing the encoder and RVQ for 2.5M steps.

**Model:** USC encodes waveforms at 16 kHz with a temporal downsampling of  $640\times$ . Each encoded latent corresponds to 40ms of speech. The frame-rate of USC is half of the temporal dimension of the teacher semantic model, thus we apply 2-dim average pooling to get the semantic targets. We use a 6-layer RVQ,  $C_{0.5}$  with 16,384 tokens in  $C_0$  to encode a larger number of semantic variations. We use 1024 tokens for the residual layers. USC achieves a bit-rate of 1.6 kbps for all the discretized RVQ tokens and of 0.35 kbps for the semantic representations.

### 3.2. Evaluation metrics

We evaluate against EnCodec [21], DAC [8], SpeechTokenizer [6] and FaCodec [22]. We extend the Voice Privacy Challenge [23] privacy and utility evaluation with SOTA models. For privacy metrics, we measure the retention of speaker-identifiable traits (SIM) by computing the cosine similarity between speaker embeddings extracted from a pre-trained TitaNet model [24]. For utility metrics, we measure content preservation via the Word Error Rate (WER) by transcribing the resynthesized speech using the Whisper v2-large [25], and quantify sentimental information via the Concordance Correlation Coefficient (CCC) [26] between the outputs of sentiment logits

through a proprietary Wav2Vec2-XLSR-based [27] sentiment extractor. For intonation faithfulness, we use the F0 Spearman’s Correlation Coefficient (SCC) [28] to measure the monotonic non-absolute correlation. To report quality, we use ViSQOL [29] and PESQ [30] metric.

### 3.3. Privacy-preserving test: linkability and singling out

Eliminating speaker-specific traits while retaining paralinguistic richness is a conflicting task [31]. Certain paralinguistic aspects are characteristic traits that facilitate speaker identification, yet crucial to be preserved for natural-sounding speech modeling. Motivated by this, we assess the level of privacy-preservation in our semantic representations through a speech privacy-preserving test based on the  $k$ -anonymity metric [32].  $k$ -anonymity is a property of data that guarantees that the information for each person contained in a set cannot be distinguished from at least  $k-1$  other individuals in the same set. This allows the preservation of certain aspects of the voice, without revealing the individual’s identity. We define two metrics based on  $k$ -anonymity, adhering to EU data protection legislation [33]. **Linkability:** ability to link two anonymized speech samples pertaining to the same individual within the dataset.

**Singling out:** ability to locate an individual’s sample within the original dataset from an anonymized sample.

Consider a dataset  $\mathcal{D}$  with utterances from a set  $\mathbb{S} = \{s_1, \dots, s_N\}$  of  $N$  speakers. The dataset is split into two partitions, the reference dataset  $\mathcal{D}_r$ , and the evaluation dataset,  $\mathcal{D}_e$ , each of them including recordings from all the  $N$  speakers. For each speaker  $s \in \mathbb{S}$ , we run  $L$  speaker identification tests, comparing a random speaker utterance  $x_s^l \in \mathcal{D}_e$  with  $N$  utterances, one for each speaker,  $Y_s^l = y_{s_1}^l, \dots, y_{s_N}^l \in \mathcal{D}_r$ , randomly selected for the  $l$  test. We calculate the speaker similarity between the evaluation utterance  $x_s^l$  and the  $N$  reference utterances in  $Y_s^l$  across the  $L$  test cases. The similarity measurement relies on the objective SIM metric. We use automatic metrics, as they have proven more accurate than humans for speaker identification [34]. Then, we compute the classification rank,  $r_s^l$ , defined as the position in the descending list of similarities of the utterance from the same speaker. Finally, we compute the mean rank per speaker,  $\bar{r}_s$  as the average of  $r_s^l$  across the  $L$  tests:

$$r_s^l = \text{rank}_{|N_{\downarrow}}(\text{sim}(x_s^l, y_{s_n}^l)) \in \mathbb{N}^{L \times N}, \quad \bar{r}_s = \frac{1}{L} \sum_{l=1}^L r_s^l \quad (2)$$

Having an average rank  $\bar{r}_s \geq k$  means that, on average, there at least  $k - 1$  different speakers which are more similar than samples from the original speaker. For non-anonymized speech and a perfect similarity metric, the rank would be 1. For completely indistinguishable samples, random speaker guessing would generate ranks that follow a uniform distribution over the  $N$  rank options. Thus, the expected value of a uniformly distributed rank is  $\mathbb{E}[r_s] = (N + 1)/2$ . For linkability, the similarities are computed with anonymized  $\mathcal{D}_r$  and  $\mathcal{D}_e$ . For singling out, the similarities compare anonymized  $\mathcal{D}_r$  and original  $\mathcal{D}_e$ .

**Perceptual privacy evaluations:** We introduce an extra evaluation with human preference to check if the proposed test, based on objective measurements, is correlated with human perception. We choose to validate the singling-out scenario as it poses the greatest challenge for individual privacy. We have randomly selected 20 unique speakers and built 20  $A/B/X$  triplets as: **X:** Unidentifiable speech sample (semantic). **A:** Sample with different content of same speaker. **B:** Sample with different content from a similar speaker. Listeners are asked to identify which speaker ( $A$  or  $B$ ) is the one that generated the semantic

reconstruction  $X$ . To get the  $B$  samples from similar speakers, we identify a pool of speakers who got a higher objective singling out ranking than the original speaker (hard case).

### 3.4. Results

**Objective evaluation:** As shown in Table 1, USC achieves competitive performance in waveform reconstruction across all RVQ levels (*High-fidelity reconstruction*) in both PESQ and ViSQOL, outperforming SpeechTokenizer and FaCodec while reducing its bit-rate by 60% and 80% respectively for 24 kHz reconstruction. DAC slightly outperforms all baselines for high-fidelity reconstruction due to its 44.1 kHz data selection. When using disentangled levels of the RVQ (*Semantic reconstruction*), SpeechTokenizer achieved the lowest speaker similarity metric, demonstrating better anonymization characteristics at the cost of removing substantial paralinguistic information, resulting in the lowest sentiment CCC metric. FaCodec, while conditioned on a mean average speaker embedding [35], does not drastically modify the semantic waveform compared to the original recording, revealing speaker leakage in its independent content and prosody VQ. Effectively preserving paralinguistics leads to increased speaker similarity, as certain prosodic characteristics facilitate speaker identification. In terms of CCC sentiment and F0 SCC, USC closes the gap from SpeechTokenizer’s semantic representations by 47.97% and 46.25% respectively, but falls behind FaCodec, whose semantic waveform is closest to the recording at the cost of a  $6.85\times$  larger bit-rate. EnCodec and DAC do not apply disentanglement; thus, their  $C_0$  reconstructions are merely low-quality acoustic versions of speech.

Table 1: Evaluation metrics on 1200 samples with varied speaking styles. If no statistically significant difference between best scores ( $p_{value} > 0.05$ ), multiple systems are highlighted in bold.

Model	RVQ	kbps	WER	PESQ	ViSQOL	SIM	CCC	SCC
Recordings	-	-	0.053	4.500	5.000	1.000	1.000	1.000
<i>High-Fidelity Reconstruction</i>								
EnCodec	$C_{0:7}$	6.00	<b>0.056</b>	2.327	3.686	0.802	0.914	0.891
DAC	$C_{0:8}$	7.75	0.059	<b>3.311</b>	<b>3.975</b>	<b>0.910</b>	<b>0.969</b>	<b>0.962</b>
SpeechTok.	$C_{0:7}$	4.00	<b>0.057</b>	2.332	3.539	0.811	0.915	<b>0.957</b>
FaCodec	$C_{0:5}$	4.80	<b>0.056</b>	2.724	3.566	0.864	0.951	0.961
USC	$C_{0:5}$	<b>1.60</b>	<b>0.056</b>	2.991	3.706	0.884	<b>0.957</b>	<b>0.959</b>
<i>Semantic Reconstruction</i>								
EnCodec	$C_0$	0.75	0.226	1.147	1.786	0.145	0.433	0.641
DAC	$C_0$	0.86	0.171	<b>1.195</b>	<b>2.077</b>	0.248	0.440	0.728
SpeechTok.	$C_0$	0.50	0.077	1.101	1.095	<b>0.056</b>	0.273	0.118
FaCodec	$C_{0:2}$	2.40	<b>0.067</b>	1.086	1.632	0.313	<b>0.629</b>	<b>0.815</b>
USC	$C_0$	<b>0.35</b>	0.091	1.067	1.687	0.218	0.526	0.526

**Privacy evaluations:** We created reference ( $\mathcal{D}_r$ ) and evaluation ( $\mathcal{D}_e$ ) sets of 45 utterances per speaker. To compute the mean rank per speaker  $\bar{r}_s$ , we used  $L = 100$  tests. We validated two variants of USC: with and without LDP to assess its impact in the speaker privacy-preserving task. Table 2 shows the ranking distributions. We report the median (p50) and the first percentile (p1), defined as the speech  $k$ -anonymity factor, which shows the worst privacy-preserving case, as it represents the minimum number of speakers who are indistinguishable from the ones with the most unique representation in the set.

For *Linkability*, when USC is not trained with LDP, 50% of the anonymized speakers (p50) are not distinguishable from at least 495 other anonymized speakers. For the final LDP variant, this number scales to 1029. Focusing on the first percentile, the  $k$ -anonymity factor, we show that for 99% of the speakers,

there are at least 35 indistinguishable anonymized speakers for the variant without LDP and 159 for the final USC variant. This result shows that LDP improves the linkability metric by 368% relatively to not using LDP. Regarding *Singling Out*, when using the final USC with LDP, 50% of the anonymized speakers are not distinguishable from at least other 816, while for 99% of the anonymized speakers, the  $k$ -anonymity factor, is 68 speakers that are closer than the original speaker identity. This is a relative improvement of 508% compared to not using LDP.

EnCodec and DAC do not apply any disentanglement, thus they report lower privacy-preserving metrics. SpeechTokenizer reports the highest privacy-preserving metric of all, at the cost of destroying most of the sentimental information. FaCodec reports low singling out metrics, suggesting that, in large-scale, FaCodec’s representations suffer from speaker leakage.

Table 2: Percentiles p50 and p1 for Linkability and Singling out.

Model	Linkability		Singling out	
	Rank p50	Rank p1	Rank p50	Rank p1
Recordings	1.01	1.00	1.01	1.00
EnCodec	435.61	37.20	673.63	32.10
DAC	266.66	12.04	181.74	4.52
SpeechTokenizer	<b>1929.61</b>	<b>774.57</b>	<b>2459.81</b>	<b>601.68</b>
FaCodec	465.12	41.72	414.54	14.15
USC (w/o LDP)	495.21	34.98	320.75	12.22
USC	1029.03	159.96	816.49	68.91
Random	3987.50	3452.06	3987.50	3452.06

We corroborate the results through the subjective *A/B/X* test evaluated by human raters using a crowd-sourcing platform. We evaluated the final version of USC with LDP. The analysis of the test shows that the probability of finding out that  $X$  is the same speaker as  $A$  is  $0.51 \pm 0.02$ . The test does not allow concluding that the speaker can be singled out from the anonymized speech. We repeated the test, using the same samples but with original waveforms. In this case, the probability of detecting the speaker is  $0.61 \pm 0.02$ , showing that according to human testers, it is possible to identify the source speaker with non-anonymized speech. Note that a probability of 0.61 may not seem high, but  $B$  samples are chosen from the most similar speakers, making the task non-trivial for a human evaluator.

## 4. Conclusions

We presented a method for speech disentanglement and privacy-preserving representation learning based on the Universal Speech Codec (USC), a low bit-rate speech codec that disentangles speech into two representation in its RVQ. Extensive evaluation showed that the main codebook,  $C_0$  learns rich semantic speech representations that encode speech content and paralinguistic information while preserving non-prosodical speaker privacy, generating sentimental speech with inconsistent identities. Moreover, USC learns the complementary speaker-specific information to enable high-fidelity 24 kHz speech reconstruction while being more efficient than any other baselines. We proposed a new speech privacy assessment test based on  $k$ -anonymity, evaluated our solution on it and corroborated that our semantic representations preserve privacy, making it infeasible for state-of-the-art identification models to link speakers between anonymized sets (linkability) or recognize the original identity of an anonymized sample (singling out). We showcase the trade-off between obfuscating speaker-identifiable traits and preserving rich semantic information. As shown in the privacy test, the way someone speaks is closely related to their identity.

## 5. References

- [1] S. C. Team, "Joint speech and text machine translation for up to 100 languages," *Nature*, vol. 637, 2025.
- [2] 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 and Language*, 2019.
- [3] A. v. d. Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," *Advances in neural information processing systems (NeurIPS)*, 2017.
- [4] L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, A. Gupta, X. Gu *et al.*, "Language model beats diffusion-tokenizer is key to visual generation," *International Conference on Learning Representations (ICLR)*, 2024.
- [5] Y. Yang, F. Shen, C. Du, Z. Ma, K. Yu, D. Povey, and X. Chen, "Towards universal speech discrete tokens: A case study for ASR and TTS," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.
- [6] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "Speech-tokenizer: Unified speech tokenizer for speech large language models," *International Conference on Learning Representations (ICLR)*, 2024.
- [7] Á. Martín-Cortinas, D. Sáez-Trigueros, J. L. Trueba, G. Beringer, I. Valles, R. Barra-Chicote, B. T. Vecino, A. Gabrys, P. Bilinski, and T. Merritt, "Investigating self-supervised features for expressive, multilingual voice conversion," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.
- [8] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," *Advances in neural information processing systems (NeurIPS)*, 2023.
- [9] A. S. Shamsabadi, B. M. L. Srivastava, A. Bellet, N. Vauquier, E. Vincent, M. Maouche, M. Tommasi, and N. Papernot, "Differentially private speaker anonymization," *Privacy Enhancing Technologies Symposium (PoPETS/PETS)*, 2023.
- [10] J. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Long-form music generation with latent diffusion," *arXiv preprint arXiv:2404.10301*, 2024.
- [11] L. Ziyin, T. Hartwig, and M. Ueda, "Neural networks fail to learn periodic functions and how to fix it," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [12] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 2021.
- [13] J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu, "Vector-quantized image modeling with improved VQGAN," *International Conference on Learning Representations (ICLR)*, 2022.
- [14] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *International Conference on Machine Learning (ICML)*, 2015.
- [15] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, 2018.
- [16] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia *et al.*, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 2021.
- [17] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing (J-STSP)*, 2022.
- [18] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "DistilHuBERT: Speech representation learning by layer-wise distillation of hidden-unit bert," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.
- [19] O. Chouchane, M. Panariello, O. Zari, I. Kerenciler, I. Chihaoui, M. Todisco, and M. Önen, "Differentially private adversarial auto-encoder to protect gender in voice biometrics," *ACM Workshop on Information Hiding and Multimedia Security*, 2023.
- [20] M. Łajszczak, G. Cámbara, Y. Li, F. Beyhan, A. van Korlaar, F. Yang, A. Joly, Á. Martín-Cortinas, A. Abbas, A. Michalski *et al.*, "BASE TTS: Lessons from building a billion-parameter text-to-speech model on 100k hours of data," *arXiv preprint arXiv:2402.08093*, 2024.
- [21] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research (TMLR)*, 2023.
- [22] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang *et al.*, "Natural-speech 3: Zero-shot speech synthesis with factorized codec and diffusion models," *arXiv preprint arXiv:2403.03100*, 2024.
- [23] 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," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.
- [24] N. R. Koluguri, T. Park, and B. Ginsburg, "Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.
- [25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *International Conference on Machine Learning (ICML)*, 2023.
- [26] B. T. Atmaja and M. Akagi, "Evaluation of error-and correlation-based loss functions for multitask learning dimensional speech emotion recognition," *International Conference on Acoustics and Vibration (ANV)*, 2020.
- [27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems (NeurIPS)*, 2020.
- [28] C. Spearman, "The proof and measurement of association between two things," *Appleton-Century-Crofts*, 1961.
- [29] M. Chinen, F. S. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "Visqol v3: An open source production ready objective speech and audio metric," *International Conference on quality of multimedia experience (QoMEX)*, 2020.
- [30] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 2001.
- [31] Z. Cai, H. L. Xinyuan, A. Garg, L. P. García-Perera, K. Duh, S. Khudanpur *et al.*, "Privacy versus emotion preservation trade-offs in emotion-preserving speaker anonymization," *IEEE Spoken Language Technology Workshop (SLT)*, 2024.
- [32] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," *SRI International*, 1998.
- [33] EU, "Data protection working party. article 29 on anonymisation techniques (wp216)," 2014. [Online]. Available: [https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216\\_en.pdf](https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf)
- [34] J. Kahn, N. Audibert, S. Rossato, and J.-F. Bonastre, "Speaker verification by inexperienced and experienced listeners vs. speaker verification system," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.
- [35] J. Yao, N. Kuzmin, Q. Wang, P. Guo, Z. Ning, D. Guo, K. A. Lee, E.-S. Chng, and L. Xie, "Npu-ntu system for voice privacy 2024 challenge," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.