Innovation and Technology Studies
Print ISSN: 3007-6919
Online ISSN: 3007-6927
DOI: https://doi.org/10.61784/its4032

PRIVTOK-XR: PRIVACY-PRESERVING REAL-TIME SPEECH PROCESSING FOR ACCESSIBLE EXTENDED REALITY SYSTEMS

Wei Yao
University of Houston, Houston 77204, TX, USA.

Abstract: Voice is the most natural hands-free input channel of extended reality (XR) and a key enabler of accessibility for users with motor or visual impairments, yet streaming speech from a headset to an edge recognizer exposes a biometric signal that reveals identity, gender and other personal attributes. We present PrivTok-XR, a split speech-processing framework in which the headset runs a causal, frame-synchronous encoder and transmits only discrete content tokens at 249 bit/s. The encoder combines gated minimal recurrent units for constant-cost streaming, a Bayesian causal instance normalization that removes slowly varying speaker statistics online, and finite scalar quantization that bounds the transmitted information by construction; a multi-attribute maximum-entropy objective with a periodically re-initialized adversary ensemble further suppresses speaker and gender cues. On a connected spoken-command benchmark built from 60 real speakers with multi-talker babble, and against informed attackers that re-train x-vector networks on the intercepted representation, PrivTok-XR reduces closed-set speaker identification accuracy from 100.0% to 25.5% (chance 5%), raises the verification equal error rate from 4.2% to 37.4% and lowers gender inference from 99.6% to 76.1% balanced accuracy, while keeping the token error rate at 3.2% on clean and 6.8% on 10-dB babble speech and reducing the uplink rate by a factor of 257. The on-device encoder has 96,324 parameters and needs 0.066 ms of computation per 40-ms hop on a single CPU core (real-time factor 0.0017), and emits commands a median of 225 ms before the spoken word ends.
Keywords: Extended reality; Accessibility; Speech privacy; Speaker anonymization; Streaming speech recognition; Finite scalar quantization; Minimal recurrent networks; Adversarial representation learning

1 INTRODUCTION

Extended reality (XR) headsets are moving from gaming devices to general-purpose computing platforms for work, education, social interaction and health care. For a large group of users, however, the default interaction channels of XR (hand-held controllers, mid-air gestures and gaze) are difficult or impossible to use. Accessibility studies identify motor, visual and auditory impairments as major barriers to immersive technologies and recommend multimodal, hands-free alternatives [1-3]. Voice is the most natural of these alternatives: spoken commands support users with limited mobility or dexterity, speech interfaces complement low-vision rendering aids, and real-time recognition is the basis of live captioning for deaf and hard-of-hearing users. A dependable, low-latency speech channel is therefore a core building block of accessible XR.

Because untethered headsets operate under strict power and thermal budgets, speech recognition is commonly offloaded to an edge or cloud server, either by streaming audio or by splitting the network between device and server [4]. This architecture creates a serious privacy problem. Speech is a biometric signal: the same stream that carries a command also carries the speaker's identity, gender, age, accent and health-related cues [5]. XR platforms already expose a remarkable amount of personal information through motion and telemetry [6-7], and adding an always-available voice channel increases this exposure further. Speech recordings are treated as personal data by data-protection regulations, which motivates privacy preservation at the source [8].

Existing protection mechanisms do not fit the constraints of accessible XR well. Voice anonymization systems in the VoicePrivacy series transform the waveform before release [8-9]; signal-processing methods such as the McAdams-coefficient transformation are light-weight but leave substantial speaker information to an attacker who adapts to them [10], while neural-vocoder and codec-language-model approaches require heavy synthesis models that are not compatible with frame-level latency on a headset [11-12]. Learned privacy-preserving representations offer an attractive alternative for recognition-only services, but adversarially trained continuous features often provide only an illusion of privacy: an adversary that re-trains a fresh classifier on the released features still recovers the protected attributes [13-14]. Systems that sanitize speech before cloud transcription or filter paralinguistic content at the edge demonstrate the value of on-device processing [15-16], yet they target offline or utterance-level operation rather than streaming XR interaction under multi-talker interference.

This paper addresses the following question: *can an XR headset stream exactly the information needed for real-time spoken interaction, and as little speaker information as possible, under a strict compute and bandwidth budget?* We answer it with **PrivTok-XR**, a privacy-by-design split speech-processing framework whose uplink consists of discrete content tokens instead of audio. The main contributions are as follows. (i) We formulate privacy-preserving streaming speech processing for XR as a joint utility, privacy, latency and rate problem and design a causal on-device encoder that emits one 10-bit token every 40 ms, so that the transmitted information is bounded by construction at about 250 bit/s. (ii)

© By the Author(s) 2024, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

We combine three complementary mechanisms that operate in the streaming regime: gated minimal recurrent units (minGRU) for constant-cost frame-synchronous inference [17], a Bayesian causal instance normalization (BCIN) that removes slowly varying speaker and channel statistics online, and finite scalar quantization (FSQ) as an information bottleneck that [18], unlike vector quantization, is free of codebook collapse. (iii) We introduce a multi-attribute maximum-entropy adversarial objective with a periodically re-initialized adversary ensemble, which counteracts the tendency of a single co-trained adversary to be fooled without removing the underlying information. (iv) We build a reproducible connected spoken-command benchmark from 60 real speakers with social-XR babble interference, and evaluate utility, accessibility-relevant group robustness, real-time cost and privacy against informed attackers that re-train x-vector-style networks on the intercepted representation.

2 RELATED WORK

2.1 Privacy and Accessibility in XR

Accessible XR research has documented interaction barriers across physical, visual, auditory and cognitive impairments and argues for accessibility as a design principle rather than a retrofit [1-3]. In parallel, security studies show that XR telemetry enables covert inference of dozens of personal attributes and that VR data privacy calls for dedicated protection tools [6-7]. Voice interaction lies at the intersection of both lines of work: it is essential for inclusive interaction, and it is one of the richest biometric channels a headset captures. Our work treats speech privacy as an integral part of accessible XR design.

2.2 Speaker Anonymization and Private Speech Representations

The VoicePrivacy initiative formalized voice anonymization with attack models of increasing strength, from ignorant to semi-informed attackers that re-train speaker verification on anonymized data [8-9]. Baseline anonymizers include x-vector replacement with neural waveform synthesis [11], McAdams-coefficient formant shifting [10], vector-quantized bottleneck features and neural audio codec language models [12, 19]; the 2024 edition additionally requires the preservation of linguistic content and emotional state [8]. Preserving privacy in speech has also been studied from legal, cryptographic and evaluation perspectives [5]. For recognition-only services, an alternative is to release a representation instead of a waveform. Srivastava *et al.* added a speaker-adversarial branch with gradient reversal to an end-to-end recognizer and observed that closed-set identification drops while open-set verification remains largely unaffected [13, 20], consistent with the general finding that adversarial removal leaves recoverable information [14]. Maximum-entropy adversarial representation learning provides more stable optimization than likelihood-based adversaries [21]. Instance normalization is known to strip speaker information in voice conversion [22], and quantized codes bottleneck speaker information in codec-based and VQ-based anonymization [12, 19]. PrivTok-XR combines these insights in a streaming, bandwidth-bounded design and evaluates it with re-trained attackers.

2.3 Efficient Streaming Sequence Models and Discrete Tokens

Streaming recognition requires causal models whose per-frame cost does not grow with context. Recurrent networks such as LSTMs satisfy this requirement but train sequentially [23]; Transformers train in parallel but need a key-value cache that grows with the stream [24]. Selective state-space models (Mamba) and their structured-duality successor provide linear-time training and constant-time inference [25-26]. Feng *et al.* recently showed that minimal GRUs and LSTMs whose gates do not depend on the previous state are trainable in parallel and competitive with these models [17]. For discretization, VQ-VAE learns a codebook with auxiliary losses and is prone to codebook collapse [27], whereas FSQ rounds a few bounded scalar channels to fixed levels [18], yielding an implicit product codebook with full utilization. CTC enables frame-synchronous decoding of such token streams [28]. We build on minGRU and FSQ as the 2024 building blocks of an on-device speech encoder.

3 PROBLEM FORMULATION AND THREAT MODEL

3.1 System Model

The headset captures a waveform and computes log-Mel features $X = (x_1, ..., x_T)$. A causal encoder f_θ maps the stream to transmitted symbols $Q = (q_1, ..., q_{T'})$ at a frame period of 40 ms, and an edge decoder g_ϕ produces the command or caption token sequence y. Utility is measured by the token error rate (TER) of $g_\phi(f_\theta(X))$; real-time operation requires that q_t depends only on past input and is computed within a fraction of the frame period on the headset.

3.2 Adversary

We consider an honest-but-curious edge operator, or an eavesdropper on the wireless link, who observes Q and wants to infer the speaker identity s and gender g. Following the semi-informed attacker of VoicePrivacy [8-9], the adversary knows f_θ, can run it on its own labelled speech, and re-trains an attack network on the resulting tokens. We evaluate three attacks: (a) closed-set speaker identification, where the attacker additionally holds labelled enrolment data of the

target users; (b) open-set speaker verification with an embedding extractor trained on other speakers; and (c) gender inference. Each attack is trained from scratch after the encoder is frozen, so no evaluation relies on a co-trained adversary.

3.3 Objective

We seek parameters (θ, ϕ) that minimize the recognition loss while keeping the attack success close to chance, subject to a rate constraint. With a finite token alphabet $\mathcal{C}$, the data-processing inequality gives, for any attribute S with Markov chain $S \to X \to Q$,

$$I(S;Q) \;\le\; I(X;Q) \;\le\; H(Q) \;\le\; T'\log_2|\mathcal{C}|. \tag{1}$$

The transmitted information is therefore capped by the token rate before any learning takes place; the remaining design problem is to spend this small budget on linguistic content rather than on speaker cues.

4 PROPOSED METHOD: PRIVTOK-XR

4.1 Overview

Figure 1 shows the architecture. On the headset, 40-dimensional log-Mel features (25-ms window, 10-ms hop) are normalized with fixed global statistics and passed through a causal convolutional front-end, two gated minGRU blocks, the BCIN layer and an FSQ tokenizer. The edge server de-quantizes the tokens with a linear layer, applies two further gated minGRU blocks and a CTC output layer over 10 command tokens plus blank. During training only, an ensemble of adversaries observes the tokens and drives the encoder towards speaker- and gender-agnostic representations. Because the privacy modules are inserted into an existing recognizer and fine-tuned, the framework can be retrofitted to a deployed streaming model. All components are causal, so the same network runs frame-synchronously at inference time.

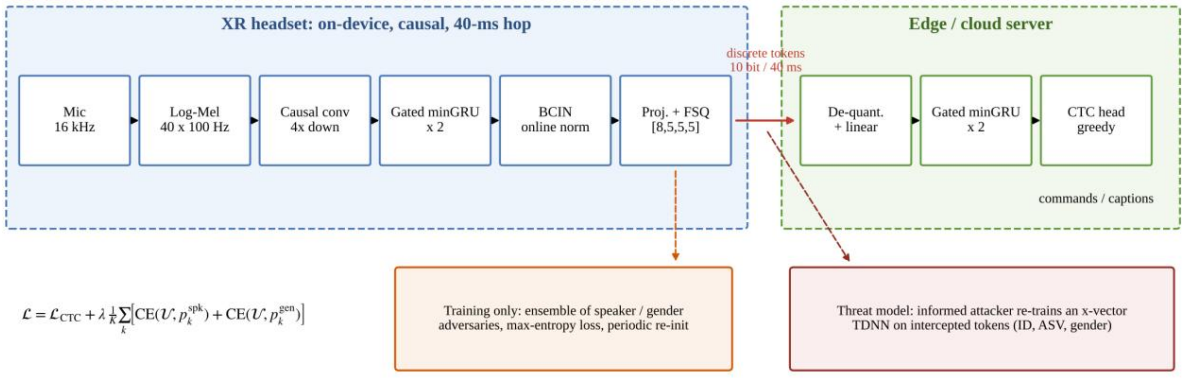

Figure 1 PrivTok-XR Architecture

Note: The headset transmits only FSQ tokens (10 bit per 40-ms frame). Adversaries are used during training only; privacy is evaluated with attackers re-trained on intercepted tokens.

4.2 Causal Front-End and Gated minGRU Encoder

Two causal convolutions with kernel 3 and stride 2 reduce the frame rate from 100 Hz to 25 Hz without look-ahead: $h_j^{(l)} = \mathrm{SiLU}\left(\sum_{k=0}^{2} W_k^{(l)}\, x_{2j-k}\right)$ and $e_t = \mathrm{SiLU}\left(\sum_{k=0}^{2} W_k^{(2)}\, h_{2t-k}^{(1)}\right)$ (biases omitted). Each gated minGRU block follows the Mamba macro-structure but replaces the selective SSM with a minGRU mixer [17, 25]:

$$[u_t; g_t] = W_{\mathrm{in}}\,\mathrm{RMSNorm}(e_t), \qquad \bar{u}_t = \mathrm{SiLU}\left(\textstyle\sum_{k=0}^{3} c_k \odot u_{t-k}\right), \tag{2}$$

$$z_t = \sigma(W_z \bar{u}_t), \quad \tilde{h}_t = W_h \bar{u}_t, \quad h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t, \tag{3}$$

$$e'_t = e_t + W_{\mathrm{out}}\left(h_t \odot \mathrm{SiLU}(g_t)\right). \tag{4}$$

Because neither z_t nor $\tilde{h}_t$ depends on h_{t-1}, the recurrence (3) is a first-order linear scan that is computed in parallel during training, whereas inference keeps only h_{t-1} and the last three convolution inputs per block. With model width $d = 64$ and inner width 96, the complete on-device encoder has 96,324 parameters and a recurrent state of 1,105 floating-point values.

4.3 Bayesian Causal Instance Normalization

Speaker and channel characteristics manifest largely as slowly varying offsets and scales of the encoder features, which instance normalization removes in voice conversion [22]. Utterance-level statistics are, however, unavailable in a stream. BCIN estimates them online with a conjugate-style prior of strength n_0 frames:

$$\mu_t = \frac{n_0\mu_0 + \sum_{\tau \le t} e_\tau}{n_0 + t}, \qquad \sigma_t^2 = \frac{n_0(\sigma_0^2 + \mu_0^2) + \sum_{\tau \le t} e_\tau^2}{n_0 + t} - \mu_t^2, \tag{5}$$

$$\hat{e}_t = \gamma \odot \frac{e_t - \mu_t}{\sqrt{\sigma_t^2 + \epsilon}} + \beta, \tag{6}$$

where $\mu_0, \sigma_0, \gamma, \beta$ are learned and $n_0 = 20$ (0.8 s). At the beginning of a stream the prior dominates and the layer behaves like a fixed normalization; as evidence accumulates it converges to per-session instance normalization. The update costs $O(d)$ per frame and two running sums of state.

4.4 Finite Scalar Quantization Bottleneck

A linear projection maps $\hat{e}_t$ to $\zeta_t \in \mathbb{R}^4$, and each channel i is bounded and rounded to L_i levels [18]:

$$q_{t,i} = \frac{1}{\lfloor L_i/2 \rfloor} \operatorname{round}\big((L_i - 1)(1-\epsilon)/2 \tanh(\zeta_{t,i} + s_i) - o_i\big), \tag{7}$$

with offsets $o_i = 0.5$ and shifts $s_i = \operatorname{artanh}(o_i/(L_i - 1)(1-\epsilon)/2)$ for even L_i and zero otherwise; gradients pass through the rounding with the straight-through estimator. We use $L = [8,5,5,5]$, i.e. $|\mathcal{C}| = 1000$ symbols or 9.97 bit per 40-ms frame (249 bit/s), and later vary the levels to trace the rate-privacy trade-off. FSQ has no learned codebook, no commitment loss and no dead codes, which matters here because the adversarial gradient otherwise tends to destabilize VQ codebooks.

4.5 Multi-Attribute Maximum-Entropy Adversary Ensemble

Each adversary a_k, $k = 1, ..., K$, is a light time-delay network with statistics pooling that predicts a speaker posterior p_k^{spk} over the training speakers and a gender posterior p_k^{gen} from the token sequence. Adversaries minimize their cross-entropy on the true labels, while encoder and decoder minimize

$$\mathcal{L} = \mathcal{L}_{\text{CTC}} + \frac{\lambda}{K} \sum_{k=1}^{K} \left[\text{CE}\big(\mathcal{U}, p_k^{\text{spk}}\big) + \text{CE}\big(\mathcal{U}, p_k^{\text{gen}}\big) \right], \tag{8}$$

where $\text{CE}(\mathcal{U}, p) = -\frac{1}{N} \sum_c \log p_c - \log N \geq 0$ vanishes only when the adversary posterior is uniform. In contrast with gradient reversal, which rewards confidently wrong predictions and can oscillate [20], the maximum-entropy target is bounded and stable [21]. Every R steps one adversary, chosen round-robin, is re-initialized. A freshly initialized adversary re-learns whatever speaker evidence remains in the current tokens, so the encoder cannot satisfy the objective by exploiting the blind spots of a stale adversary; this mimics the re-trained attacker used at test time. The weight λ is ramped linearly during the first 30% of training. The procedure is summarized below.

Algorithm 1. Training of PrivTok-XR

1: initialize θ, ϕ from a converged non-private streaming recognizer; initialize BCIN, projection and adversaries $a_1, ..., a_K$ randomly
2: **for** step $n = 1, ..., N$ **do**
3: sample a noisy command batch (X, y, s, g); $Q \leftarrow f_\theta(X)$
4: update (θ, ϕ) with $\nabla \mathcal{L}$ of (8), $\lambda \leftarrow \lambda_{\max} \min(1, n/0.3N)$
5: update every a_k with $\nabla \left[\text{CE}\big(s, p_k^{\text{spk}}\big) + \text{CE}\big(g, p_k^{\text{gen}}\big) \right]$ on $\text{sg}(Q)$
6: **if** $n \bmod R = 0$ **then** re-initialize $a_{(n/R) \bmod K}$
7: **end for**

4.6 Streaming Decoding

The edge decoder runs the recurrence (3) frame by frame and applies greedy CTC decoding [28]: a token is emitted at the first frame of every non-blank run that differs from the previous label. A token decoded at encoder frame j is available after input frame $4j$, i.e. at time $(4jH + W)/f_s$ with hop $H = 160$, window $W = 400$ and $f_s = 16$ kHz. We define the emission latency as this time minus the true end time of the spoken token.

5 EXPERIMENTAL SETUP

5.1 Data: Connected Spoken Commands from Real Speakers

We use AudioMNIST [29], which contains 30,000 recordings of the English digits zero to nine spoken 50 times by each of 60 speakers (12 female, 48 male; 16 accent labels), recorded at 48 kHz in several rooms, one of which is annotated as a VR room. Recordings are resampled to 16 kHz, trimmed with a 35-dB energy threshold and normalized to -25 dBFS. The speakers are split into 40 training and 20 test speakers, stratified by gender and by German versus other accents (test: 4 female, 16 male). To emulate XR voice commands such as menu selections, room numbers or PIN entry, we compose connected command strings of 1-5 tokens from one speaker with random pauses of 50-300 ms. The training pool contains 24,000 strings; 60% of them are mixed with four-talker babble from other training speakers at an SNR drawn uniformly from 0-20 dB, emulating co-located users in social XR.

The utility test set contains 1,000 strings (2-5 tokens) from the 20 unseen speakers, evaluated clean and with babble from other unseen talkers at 10, 5 and 0 dB SNR. For the privacy evaluation, each test speaker contributes 80 enrolment strings

built from repetitions 0-24 and 40 probe strings built from repetitions 25-49, so that enrolment and probe never share a recording; attackers for verification and gender inference are trained on 4,000 strings of the 40 training speakers.

5.2 Compared Systems

All learned systems share the same front-end, width ($d = 64$) and depth (2 device + 2 edge blocks). The non-private recognizer is trained with CTC from scratch for 1,000 steps of 32 strings (AdamW, peak learning rate 2×10^{-3}, cosine decay). Every privacy-oriented system follows the same two-stage recipe: it is initialized from this converged recognizer, its privacy module is inserted, and the model is fine-tuned for 1,000 steps with a peak learning rate of 10^{-3}. The recipe matters for low-rate bottlenecks: training PrivTok-XR from scratch with the same budget reaches only 29.8% clean TER, against 3.2% with two-stage training. **Cloud-Raw** streams log-Mel features (64 kbit/s at 16-bit precision) to the server, which runs the full model. **McAdams** applies the VoicePrivacy B2 transformation ($\alpha = 0.8$) on the device [10]; for noisy conditions the complete mixture is anonymized, and the recognizer is trained on anonymized speech. **Split-Cont** transmits the continuous 64-dimensional encoder output (25.6 kbit/s in half precision). **Adv-GRL** adds a speaker and gender adversary with gradient reversal ($\lambda = 0.5$) to Split-Cont, following [13]. **VQ-1024** inserts a VQ-VAE bottleneck with 1,024 codes of dimension 16 (10 bit per frame) [27], representative of VQ content features [19]. **FSQ-1000** uses the FSQ bottleneck without adversaries or BCIN. **PrivTok-XR** uses BCIN, FSQ with $L = [8,5,5,5]$, $K = 3$ adversaries re-initialized every $R = 150$ steps and $\lambda = 1$.

5.3 Attackers and Metrics

Attackers follow the x-vector design [30]: three dilated time-delay convolution layers with 128 channels and layer normalization, mean and standard-deviation pooling, a 128-dimensional embedding layer, and speaker and gender heads with class-balanced loss weights. Inputs are standardized per dimension; models are trained with AdamW and selected on a held-out 10% split. We report closed-set speaker identification accuracy (SID, chance 5%), the equal error rate (EER) of cosine scoring between probe embeddings and 10-string enrolment models (16,000 trials), and gender balanced accuracy (BA, chance 50%). Utility is reported as TER (edit distance divided by the number of reference tokens) and string error rate; real-time behaviour as per-hop compute time on one CPU thread, real-time factor (RTF) and emission latency. All experiments run on a single core of an Intel Xeon CPU at 2.3 GHz with JAX.

6 RESULTS AND DISCUSSION

6.1 Privacy-Utility Comparison

Table 1 summarizes utility, privacy and uplink rate for all systems; Figure 2 visualizes the trade-off between average noisy TER and the two speaker-privacy metrics. Unprotected representations reveal the speaker almost perfectly. An attacker that intercepts log-Mel features (Cloud-Raw) identifies the 20 test speakers with 100.0% accuracy, verifies them at 4.2% EER and infers gender with 99.6% balanced accuracy. The continuous encoder output of a split recognizer is hardly safer (99.4% SID, 5.6% EER, 97.3% gender BA), which confirms that recognition training alone does not remove speaker information from intermediate features [13]. Signal-level McAdams anonymization keeps recognition close to the unprotected system (1.2% clean TER) but offers little protection once the attacker re-trains on anonymized speech: SID remains at 99.9% and EER at 4.5%. A deterministic spectral-envelope transformation changes every voice in a consistent way, so speakers stay separable after the transformation. Gradient-reversal training on the continuous features (Adv-GRL) yields 97.0% SID and 9.1% EER at 2.2% clean TER, while its co-trained adversary reports only 9.4% speaker accuracy at the end of training — the gap between the training-time adversary and the re-trained attacker illustrates why co-trained adversaries are not a reliable privacy measure [13-14]. The VQ-1024 bottleneck transmits the same 10 bit per frame as our tokenizer and reaches 69.5% SID and 20.4% EER at 1.3% clean TER; it uses 606 of its 1,024 codes (perplexity 89). Discretization alone is a strong privacy mechanism: FSQ-1000 without adversaries lowers SID to 59.2% and raises EER to 25.3% with no loss in clean TER (1.5%), which is the empirical counterpart of the rate bound (1). PrivTok-XR adds BCIN and the adversary ensemble and achieves the strongest protection of all compared systems: SID drops to 25.5% (3.9 times lower than Cloud-Raw), EER rises to 37.4%, and gender inference falls to 76.1% BA. The price is an increase of the clean TER from 1.5% to 3.2% and of the 10-dB TER from 5.3% to 6.8%. The uplink carries 249 bit/s, or 111 bit/s after entropy coding of the observed token distribution (perplexity 22), which is 257 and 576 times less than streaming 16-bit log-Mel features.

Table 1 Utility (TER, %), Uplink Rate and Privacy against Re-Trained Attackers (SID: Closed-Set Identification Accuracy, %; EER: Verification Equal Error Rate, %; Gender: Balanced Accuracy, %)

System	bit/s	TER clean	TER 5 dB	SID ↓	EER ↑	Gender ↓
Cloud-Raw	64.0k	1.5	12.4	100.0	4.2	99.6
McAdams [10]	64.0k	1.2	16.4	99.9	4.5	99.1
Split-Cont	25.6k	1.5	12.4	99.4	5.6	97.3
Adv-GRL [13]	25.6k	2.2	10.8	97.0	9.1	92.2
VQ-1024 [27]	250 (162)	1.3	12.3	69.5	20.4	84.1
FSQ-1000 [18]	249 (142)	1.5	12.8	59.2	25.3	87.9

System	bit/s	TER clean	TER 5 dB	SID ↓	EER ↑	Gender ↓
PrivTok-XR	**249 (111)**	**3.2**	**15.1**	**25.5**	**37.4**	**76.1**

Note: Rates in parentheses are entropy-coded; results at 10 and 0 dB are shown in Figure 3.

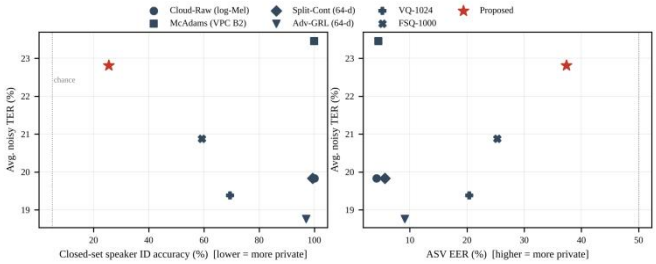

Figure 2 Privacy-Utility Trade-Off

Note: Left: closed-set speaker identification accuracy (lower is more private, chance 5%). Right: verification EER (higher is more private). The vertical axis is the TER averaged over 10, 5 and 0 dB babble.

6.2 Robustness to Multi-Talker Interference

Social XR applications place several talkers in the same physical or virtual room, so interference from other voices is the dominant acoustic challenge. Figure 3 reports TER as a function of babble SNR. All systems degrade gracefully down to 5 dB and strongly at 0 dB, where the competing talkers are as loud as the user. PrivTok-XR follows the same curve as the unprotected recognizer with a moderate offset (15.1% versus 12.4% at 5 dB and 46.6% versus 41.9% at 0 dB), showing that the 4-dimensional token stream retains the information needed to separate the target speaker's commands from interfering speech that the recognizer has learned to ignore.

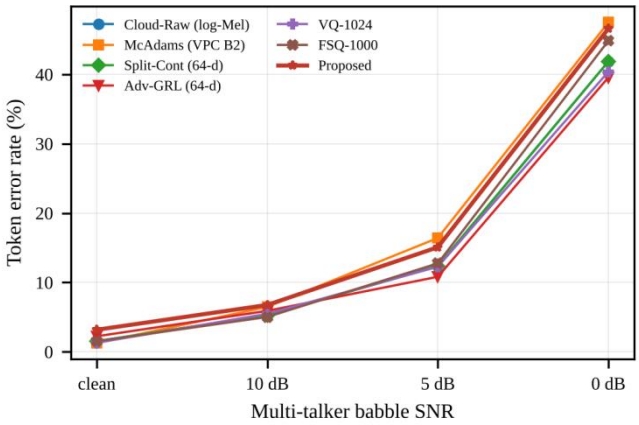

Figure 3 Token Error Rate under Four-Talker Babble from Unseen Speakers at Decreasing SNR

6.3 Ablation and the Illusion of Privacy

Table 2 Ablation of PrivTok-XR (TER, SID, EER and Gender BA in %)

Variant	TER clean	TER 5 dB	SID ↓	EER ↑	Gender ↓
Proposed (full)	3.2	15.1	25.5	37.4	76.1
w/o adversary, w/o BCIN	1.5	12.8	59.2	25.3	87.9

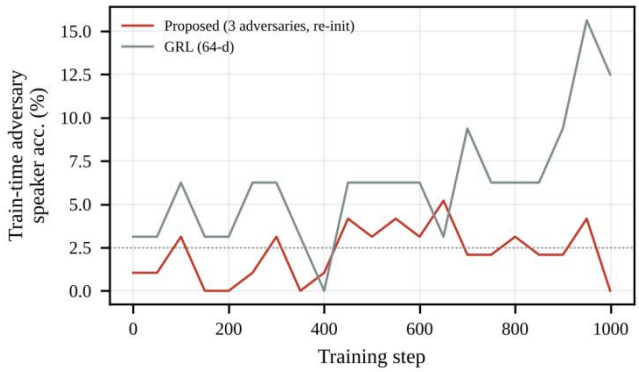

Figure 4 Speaker Accuracy of the Co-Trained Adversaries during Training (40 Training Speakers, Chance 2.5%)

Note: Re-trained attackers recover much more speaker information (Table 1).

Table 2 isolates the contributions of the privacy mechanisms. Comparing the full model with the FSQ bottleneck alone, BCIN and the adversary ensemble jointly reduce SID from 59.2% to 25.5%, raise EER from 25.3% to 37.4% and lower gender BA from 87.9% to 76.1%. Figure 4 shows the speaker accuracy of the co-trained adversaries during training: for PrivTok-XR it stays close to the 2.5% chance level of the 40 training speakers (final value 2.3%), whereas the attacker re-trained after convergence reaches 25.5% on the test speakers. Training-time adversary accuracy is therefore not a privacy metric; all privacy numbers in this paper come from attackers trained from scratch on the frozen representation. The two-stage recipe is equally important: with the same budget, PrivTok-XR trained from scratch reaches only 29.8% clean TER, because the low-rate bottleneck and the adversarial objective slow down the formation of linguistic units, whereas starting from a converged recognizer the privacy modules only have to reshape an existing representation.

6.4 Real-Time Performance

The streaming implementation processes one 40-ms hop (four log-Mel frames) per call and reproduces the full-sequence encoder output up to a maximum absolute difference of 0.0e+00. On one core of a 2.3-GHz Xeon CPU, the PrivTok-XR device encoder takes 0.066 ms per hop on average (95th percentile 0.098 ms), i.e. a real-time factor of 0.0017, and requires about 213 kFLOP per hop; log-Mel extraction adds 0.030 ms. The edge decoder needs 0.058 ms per hop. For comparison, the McAdams transformation implemented frame-by-frame costs 0.96 ms per hop. Table 3 compares sequence mixers of identical width and depth in the device encoder. Compute per hop is independent of the weight values, so all backbones are benchmarked with the same streaming harness. The gated minGRU block keeps a constant state, as LSTM and Mamba do, whereas the causal Transformer must keep a key-value cache that grows with the stream (reported at a 256-frame, i.e. 10.2-s, context). Beyond compute, the decisive latency for interaction is when a command is recognized. Because CTC emits a token as soon as the causal evidence suffices, PrivTok-XR emits commands a median of 225 ms before the end of the spoken token, and 90% of the tokens at least 104 ms before it (measured on correctly recognized clean test strings, including the 25-ms analysis window). Together with the negligible per-hop cost, this leaves the time budget of the headset to rendering and tracking.

Table 3 Streaming Cost of the Device Encoder with Different Sequence Mixers (Single CPU Thread, Batch 1)

Backbone	Params	ms / hop	RTF	State	MFLOP / hop
Gated minGRU (ours)	95.8k	0.067	0.0017	768	0.21
Mamba (S6)	69.7k	0.078	0.0019	3,648	0.17
LSTM	156.3k	0.067	0.0017	384	0.33
Causal Transformer	86.8k	0.148	0.0037	65,536 (grows)	0.35

6.5 Accessibility-Oriented Group Analysis

An accessible interface must work for all users, not only for the majority group of the training data. Figure 5 breaks the 5-dB TER down by gender and by accent group (German-accented versus other accents). For PrivTok-XR the TER is 20.2% for female and 13.8% for male speakers, and 13.3% for German-accented versus 19.1% for other accents; the corresponding values of the unprotected recognizer are 16.8%, 11.2%, 11.1% and 15.3%. The privacy mechanisms therefore do not introduce a new disparity between groups, while the accent gap that already exists in the unprotected recognizer remains a target for data collection with more diverse speakers.

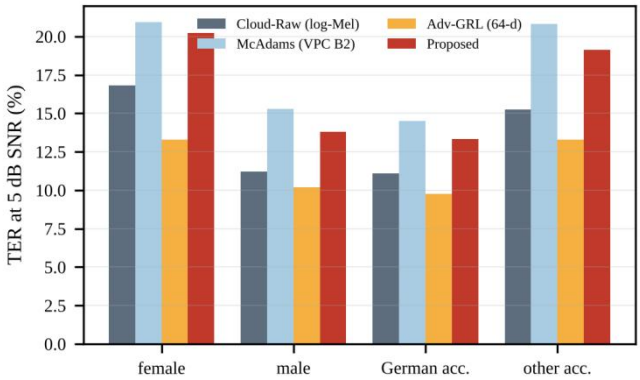

Figure 5 TER at 5 dB SNR per Speaker Group (Female/Male, German-Accented/Other Accents)

6.6 Implications for XR Deployment

Three properties make the design attractive for headsets. First, the uplink is a narrow, fixed-rate symbol stream, so it can be encrypted, rate-limited and audited in a way that raw audio cannot, and its information content is capped by (1) independently of any attacker model. Second, the privacy modules are inserted into an existing streaming recognizer and fine-tuned, so a platform can retrofit them without retraining from scratch. Third, all on-device operations are causal and constant-cost, which keeps the voice channel responsive for users who depend on it. The remaining gender leakage

indicates that coarse attributes are partly carried by linguistic timing and spectral balance at the token level; extending the adversary ensemble to further attributes, combining tokens with local differential privacy, and scaling the vocabulary from commands to sub-word captions are natural next steps.

7 CONCLUSION

We presented PrivTok-XR, a privacy-preserving real-time speech processing framework for accessible XR. The headset streams 10-bit FSQ tokens every 40 ms instead of audio, produced by a causal gated-minGRU encoder with Bayesian causal instance normalization and trained with a re-initialized maximum-entropy adversary ensemble. Against informed attackers that re-train on the intercepted tokens, the framework lowers speaker identification from 100.0% to 25.5%, raises the verification EER from 4.2% to 37.4% and reduces the uplink rate by a factor of 257, at 3.2% clean token error rate and 0.066 ms of computation per 40-ms hop. These results show that voice interaction in XR can be made accessible and privacy-preserving at the same time.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Mott M, Cutrell E, Gonzalez Franco M, et al. Accessible by design: An opportunity for virtual reality. In: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), 2019: 451-454.
[2] Dudley J, Yin L, Garaj V, et al. Inclusive Immersion: A review of efforts to improve accessibility in virtual reality, augmented reality and the metaverse. Virtual Reality, 2023, 27(4): 2989-3020.
[3] Creed C, Al-Kalbani M, Theil A, et al. Inclusive AR/VR: Accessibility barriers for immersive technologies. Universal Access in the Information Society, 2024, 23(1): 59-73.
[4] Matsubara Y, Levorato M, Restuccia F. Split computing and early exiting for deep learning applications: Survey and research challenges. ACM Computing Surveys, 2022, 55(5): Art. 90.
[5] Nautsch A, Jiménez A, Treiber A, et al. Preserving privacy in speaker and speech characterisation. Computer Speech & Language, 2019, 58: 441-480.
[6] Nair V, Munilla Garrido G, Song D, et al. Exploring the privacy risks of adversarial VR game design. Proceedings on Privacy Enhancing Technologies, 2023: 238-256.
[7] Munilla Garrido G, Nair V, Song D. SoK: Data privacy in virtual reality. Proceedings on Privacy Enhancing Technologies, 2024: 21-40.
[8] Tomashenko N, Miao X, Champion P, et al. The VoicePrivacy 2024 Challenge evaluation plan. arXiv:2404.02677, 2024.
[9] Tomashenko N, Wang X, Vincent E, et al. The VoicePrivacy 2020 Challenge: Results and findings. Computer Speech & Language, 2022, 74: Art. 101362.
[10] Patino J, Tomashenko N, Todisco M, et al. Speaker anonymisation using the McAdams coefficient. In: Proceedings of Interspeech, 2021: 1099-1103.
[11] Fang F, Wang X, Yamagishi J, et al. Speaker anonymization using x-vector and neural waveform models. In: Proceedings of the 10th ISCA Speech Synthesis Workshop (SSW), 2019: 155-160.
[12] Panariello M, Nespoli F, Todisco M, et al. Speaker anonymization using neural audio codec language models. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024: 4725-4729.
[13] Srivastava B M L, Bellet A, Tommasi M, et al. Privacy-preserving adversarial representation learning in ASR: Reality or illusion? In: Proceedings of Interspeech, 2019: 3700-3704.
[14] Elazar Y, Goldberg Y. Adversarial removal of demographic attributes from text data. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018: 11-21.
[15] Ahmed S, Roy Chowdhury A, Fawaz K, et al. Preech: A system for privacy-preserving speech transcription. In: Proceedings of the 29th USENIX Security Symposium, 2020: 2703-2720.
[16] Aloufi R, Haddadi H, Boyle D. Paralinguistic privacy protection at the edge. ACM Transactions on Privacy and Security, 2023, 26(2): Art. 19.
[17] Feng L, Tung F, Ahmed M O, et al. Were RNNs all we needed? arXiv:2410.01201, 2024.
[18] Mentzer F, Minnen D, Agustsson E, et al. Finite scalar quantization: VQ-VAE made simple. In: Proceedings of the International Conference on Learning Representations (ICLR), 2024.
[19] Champion P, Jouvet D, Larcher A. Are disentangled representations all you need to build speaker anonymization systems? In: Proceedings of Interspeech, 2022.
[20] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks. Journal of Machine Learning Research, 2016, 17(59): 1-35.
[21] Roy P C, Boddeti V N. Mitigating information leakage in image representations: A maximum entropy approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[22] Chou J-C, Yeh C-C, Lee H-Y. One-shot voice conversion by separating speaker and content representations with instance normalization. In: Proceedings of Interspeech, 2019: 664-668.
[23] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780.
[24] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), 2017: 5998-6008.
[25] Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces. In: Proceedings of the Conference on Language Modeling (COLM), 2024.
[26] Dao T, Gu A. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In: Proceedings of the International Conference on Machine Learning (ICML), 2024.
[27] van den Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning. In: Advances in Neural Information Processing Systems (NeurIPS), 2017: 6306-6315.
[28] Graves A, Fernandez S, Gomez F, et al. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the International Conference on Machine Learning (ICML), 2006: 369-376.
[29] Becker S, Vielhaben J, Ackermann M, et al. AudioMNIST: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. Journal of the Franklin Institute, 2024, 361(1): 418-428.
[30] Snyder D, Garcia-Romero D, Sell G, et al. X-vectors: Robust DNN embeddings for speaker recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018: 5329-5333.