Joint Timbral and Non-Timbral Speaker Anonymisation

Rayane Bakari[1,2], *Olivier Le Blouch*[1], *Nicolas Gengembre*[1], *Nicholas Evans*[2]

[1]Orange innovation, France
[2]EURECOM, Sophia Antipolis, France

{mohamedrayane.bakari, olivier.leblouch, nicolas.gengembre}@orange.com,
evans@eurecom.fr

Abstract

Voice anonymisation aims to conceal speaker identity while preserving linguistic content. Most approaches target predominantly timbral cues and often overlook non-timbral cues such as prosody, rhythm, speaking style and accent, which may still leak speaker-specific information related to voice identity after anonymisation. With this paper, we propose a speaker anonymisation system that explicitly obfuscates both timbral and non-timbral cues. Extensive experiments conducted within the VoicePrivacy Challenge framework show improved protection against attacks exploiting non-timbral information compared to state-of-the-art systems. For evaluation, we use a pair of complementary automatic speaker verification models to demonstrate improved anonymisation robustness by 32% relative to attacks which target either timbral and non-timbral cues. Results also show stronger anonymisation comes at the cost of only moderate degradation to intelligibility and naturalness.
Index Terms: voice privacy, anonymisation, speech disentanglement, non-timbral

1. Introduction

Voice recordings contain sensitive information that can be used to identify individuals or infer personal traits, a concern that has grown with the rapid expansion of always-listening voice-activated and voice-interactive services [1]. In response, voice anonymisation techniques have been developed to sanitise speech recordings of the voice identity while preserving linguistic and paralinguistic attributes that are essential for communication and downstream tasks [2].

Given the assumption that voice identity is encoded primarily in timbral cues, most anonymisation systems [3, 4, 5, 6] function by substituting timbral characteristics within a speech recording with those of another, possibly pseudo voice. However, recent studies have shown that the obfuscation of timbral cues alone is insufficient. Other voice identity cues, e.g. in the form of non-timbral characteristics, persist after anonymisation and can still be exploited to re-identify the original speaker [7]. As a result, anonymisation methods that focus solely on timbral information often remain vulnerable to re-identification attacks.

The VoicePrivacy Attacker Challenge [8] was launched to study such anonymisation vulnerabilities, to stimulate progress in evaluation, and hence to improve anonymisation robustness. Attack techniques reported in [9, 10, 11, 12] show that previous approaches to evaluation result in exaggerated estimates of anonymisation performance [13]. Results from the Attacker Challenge and other recent work all confirm that non-timbral cues such as prosody, rhythm, speaking style, and accent are among the most dominant source of speaker information after anonymisation [7].

Among these cues, phoneme duration and temporal speech patterns have recently been shown to carry strong speaker-discriminative information [14, 15]. This suggests that effective anonymisation requires a broader strategy that addresses both timbral and non-timbral sources of voice identity information.

We propose a joint timbral and non-timbral speaker anonymisation framework designed to better mitigate identity leakage. Relying on discrete speech units as also proposed in [16, 17] combined with a phoneme-duration predictor and a Unit HiFi-GAN vocoder, our approach supports controlled modification of the temporal speech structure while preserving to some extent linguistic and paralinguistic attributes.

For a more comprehensive evaluation of privacy, we use a pair of complementary automatic speaker verification (ASV) systems designed to capture either timbral or non-timbral cues [18]. In addition, since automatic speech recognition (ASR) performance alone does not reliably reflect perceptual speech quality [19], we complement an objective evaluation of intelligibility using ASR with objective evaluation of naturalness using a pair of approaches to predict the mean opinion score (MOS). This evaluation, using a broader set of metrics, allows for a more nuanced comparison of the privacy–utility trade-off for different anonymisation strategies. To the best of our knowledge, this work is the first to systematically evaluate anonymisation robustness using ASV systems sensitive to both timbral and non-timbral voice characteristics.

The main contributions of this work are as follows:

- we introduce a joint speaker anonymisation framework that reduces identity leakage by explicitly targeting the obfuscation both timbral and non-timbral voice characteristics;
- we show that modifying phoneme durations yields +6–9% absolute EER gains against non-timbral attacks;
- we highlight that conditioning the vocoder on non-timbral (NT) speaker embeddings improves obfuscation of non-timbral cues, enhancing anonymisation robustness while helping preserve perceptual naturalness;
- our extensive evaluation shows stronger protection against non-timbral attacks with only moderate, controlled impact on intelligibility and naturalness.

The remainder of the paper is organised as follows. In Section 2, we review existing anonymisation baselines and current state-of-the art system relevant to our non-timbral obfuscation approach. We describe our approach to joint timbral and non-timbral speaker anonymisation in Section 3. We present our evaluation methodology in Section 4, followed by experimental results, ablation study, and analysis in Section 5, Section 6 and 7. Conclusions are presented in Section 8.

2. Speaker Anonymisation Baselines

In this section, we present a review of representative speaker anonymisation systems used as baselines in our experiments. First, we describe the Voice Privacy Challenge (VPC) baseline systems, and then, the nearest neighbours-based approaches which serve as baseline for comparing our proposed methods.

2.1. VPC Baselines

A set of six baseline anonymisation systems were made available to the community for the third VPC edition. Among these, systems B4 and B5 were the most competitive with the state of the art at the time. Baseline B4 leverages the capabilities of neural audio codecs (NACs) to encode speech into discrete, quantised bottleneck representations which provide for the inherent disentanglement of voice identity, linguistic information, and prosody. Anonymisation is performed by substitution of the voice identity component at the bottleneck level, while retaining content and prosodic cues to ensure naturalness and intelligibility in reconstructed, anonymised speech. Baseline B5 incorporates vector quantisation (VQ) to enhance the disentanglement of linguistic and voice information [20]. The system extracts fundamental frequency (F0) features using a Torch-based implementation of YAAPT [21], and acoustic vector-quantised bottleneck (VQ-BN) features from an ASR acoustic model trained to recognise left-biphones. These VQ-BN features, together with F0 and a target voice identity encoded as a one-hot vector corresponding to the voice of one specific speaker observed during training, are fed to a HiFi-GAN vocoder, which synthesises an anonymised speech output.

2.2. kNN-VC

kNN-VC [22] is an any-to-any voice conversion system based on a *k*-nearest neighbour (kNN) retrieval mechanism. Speech representations are first extracted from both source and reference (target) utterances using WavLM [23]. Voice conversion is then performed by replacing each frame of the source representation with the nearest neighbour extracted from the reference representation space. In practice, converted representations are the average of four nearest target representations selected according to cosine similarity. This approach is effective because the representation space of WavLM is structured by phonetic content [24]. A HiFi-GAN vocoder [25] is then used to synthesise an anonymised output from the resulting converted representations. To simulate the kNN conversion process, the vocoder is trained on data where representations are replaced with those from other utterances of the same speaker

Previous studies have shown that continuous representations learned by self-supervised speech models such as WavLM or HuBERT encode both contextual and voice information [26], which is problematic for anonymisation tasks requiring the disentanglement of linguistic content and voice identity. One approach to reduce the capture of voice-related information is the application of k-means clustering to HuBERT representations, thereby obtaining representations of discrete speech units. However, the use of overly discretised representations degrades pronunciation accuracy [27], a phenomenon that we also observe in our experiments.

2.3. Private kNN-VC

Private kNN-VC [15] extends kNN-VC with explicit mechanisms designed to mitigate voice identity leakage through prosodic cues. While the original system replaces spectral characteristics via kNN retrieval in a WavLM feature space, it preserves the temporal structure of the source utterance, including phone durations and phonetic variability, which may encode speaker-specific information.

To address this, Private kNN-VC introduces a phone predictor and a duration predictor, both based on the FastSpeech2 [28] convolutional variance predictor architecture. The phone predictor assigns a phonetic label to each frame of the WavLM feature sequence, while the duration predictor estimates the duration of each phone in a phonetic transcript. The transcript is obtained by removing consecutive occurrences from the sequence of predicted phones.

Anonymisation strength is controlled in two ways. First, predicted durations can partially or fully replace the original ones, allowing control of source speaker temporal characteristics. Second, phonetic variability in retrieved target representations is controlled by clustering of the target speaker representations based on their predicted phones, and then the cluster centroids are used for the conversion instead of the target features. Finally, the remaining processes are identical to those of the original kNN-VC algorithm. In this study, we use the (7-8) configuration with random gender conversion, as proposed in [15], which is representative of realistic anonymisation settings without prior speaker information. Although other configurations may produce better performance, this choice was validated by the author of Private kNN-VC to avoid overestimation, as these configurations from [15] are prone to bias in performance evaluation [29].

However, the effectiveness of these mechanisms for suppressing non-timbral identity cues has not been extensively evaluated using dedicated privacy metrics. Previous studies report ASV performance [15], with no explicit analysis of the contributions of temporal or prosodic cues to residual identity leakage. This motivates our evaluation framework, where complementary ASV systems are used to assess the relative contributions to anonymisation from timbral and non-timbral cues.

3. Proposed Speaker Anonymisation Method

Figure 1 illustrates the complete anonymisation pipeline. The method relies on the independent representation of linguistic content and voice identity using discrete speech units and modification of both the temporal speech structure and speaker conditioning used during waveform synthesis. The system consists of two main components: (1) a phoneme duration predictor; (2) a unit HiFi-GAN vocoder;

3.1. Phoneme duration predictor

The phoneme duration predictor follows the architecture proposed in [17] and is trained using the *LibriSpeech train-clean-100* dataset [30]. Given an input waveform, a pretrained HuBERT model[1] followed by a k-means clustering step is first used to extract discrete speech units (DUs). The vocabulary size corresponds to the number of centroids, with $K = 200$, inspired from [26]. In practice, a unit corresponds to a phonetic segment of 20 ms.

Let the resulting sequence of DUs be denoted by $d = (d_1, \ldots, d_T)$ with $d_t \in \{1, \ldots, K\}$ and T the number of frames. To estimate phoneme durations, the sequence

[1]Available at: `https://huggingface.co/facebook/hubert-base-ls960`

is deduplicated by collapsing consecutive repetitions, e.g., $(20, 20, 20, 11, 25, 25) \rightarrow (20, 11, 25)$. This deduplicated sequence is denoted by $d_c = (d_1^c, \ldots, d_L^c)$ where $d_i^c \in \{1, \ldots, K\}$ and L is the sequence length. The number of repetitions removed during deduplication reflects the rhythmic structure of the speech signal.

The duration predictor estimates the duration associated with each unit in d_c, conditioned on speaker embeddings. Conditioning the duration predictor on a target speaker embedding enables rhythm conversion consistent with the speaking style of the target speaker. Two types of embeddings are considered: non-timbral (NT) embeddings from [18] and ECAPA embeddings [31]. The model is trained as a regression task using a Mean Squared Error (MSE) loss between predicted and ground-truth durations. Because discrete units do not perfectly align with phonemes, neighbouring units may still correspond to the same phoneme after deduplication. To account for this mismatch, an additional MSE loss is applied to the total duration of four neighbouring units, encouraging local errors to compensate for each other [17].

During inference, predicted durations are rounded to obtain integer repetitions of each unit. To reduce rounding bias, rounding remainders are accumulated and propagated to subsequent predictions.

3.2. Unit HiFi-GAN Vocoder

We follow the approaches in [27, 32] to train a variation of the HiFi-GAN vocoder. The generator is adapted to accept a sequence of discrete units (DUs) together with a speaker embedding used for speaker conditioning during waveform synthesis.

Each unit index d_t is first mapped to a continuous vector through a learned embedding table, producing a sequence of unit embeddings. To condition the vocoder on speaker identity, the speaker embedding z_{spk} is concatenated with the unit embedding at each frame of the upsampled DU sequence, yielding the conditioning representation:

$$e_t = [d_t, z_{\text{spk}}]. \quad (1)$$

The resulting sequence $\{e_t\}_{t=1}^{T}$ is provided as input to the HiFi-GAN generator, which produces the speech waveform through a stack of convolutional layers.

During training, the vocoder is trained independently for waveform reconstruction using DUs extracted from real speech. At inference time, the input consists of the deduplicated unit sequence d_c expanded according to the predicted durations, together with a target speaker embedding.

To support the two types of speaker representations used in this work, we train two separate Unit HiFi-GAN models: (i) a vocoder conditioned on ECAPA embeddings [31] (192-dimensional), which capture global speaker characteristics, and (ii) a vocoder conditioned on NT embeddings [18] (250-dimensional), which encode prosodic and stylistic traits while being less sensitive to timbral information.

Each vocoder is trained using the corresponding speaker representation. Both Unit HiFi-GAN models are trained on the LibriTTS *train-clean-100* subset using two RTX 4090 GPUs.

3.3. Speaker anonymisation technique

To perform anonymisation with the system illustrated in Figure 1, an input utterance from a source speaker is first converted into a sequence of discrete units (DUs). The resulting unit sequence is then deduplicated to obtain a compact representation of the phonetic content. This deduplicated sequence

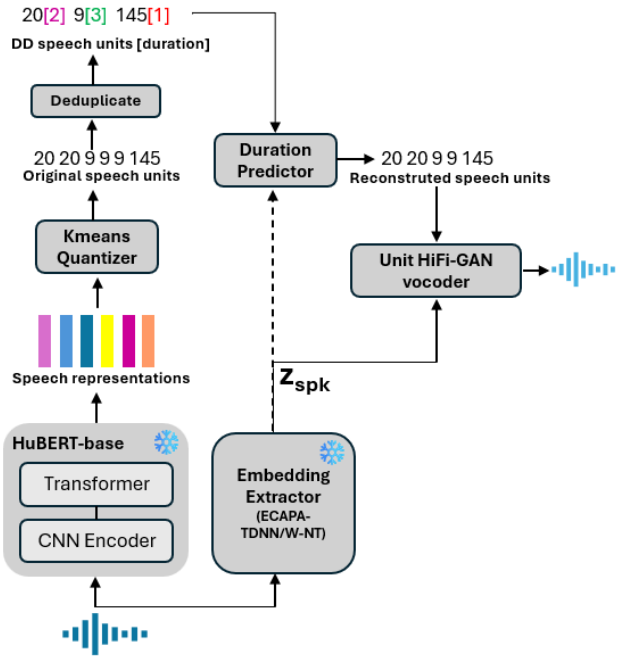

Figure 1: *Overview of the proposed speaker anonymisation system based on discrete speech units, duration prediction, and a Unit HiFi-GAN vocoder.*

d_c is provided to the phoneme duration predictor together with a randomly sampled target speaker embedding (either ECAPA or NT) drawn from *LibriTTS-train-other-500* [33]. Conditioned on this embedding, the duration predictor generates new durations for each unit that reflect the rhythmic characteristics of the target speaker. The unit sequence is subsequently expanded according to the predicted durations and passed to the corresponding Unit HiFi-GAN vocoder, which synthesises the anonymised speech waveform while being conditioned on the same target speaker embedding.

By jointly modifying the rhythmic structure through the duration predictor and conditioning the vocoder on a target speaker representation, the system may alter speaker-specific characteristics while preserving the linguistic content of the original utterance.

4. Evaluation

We evaluate the proposed anonymisation system using the official 2024 Voice Privacy Challenge (VPC) framework [2]. To obtain a comprehensive assessment, we employ complementary automatic speaker verification (ASV) systems targeting timbral and non-timbral cues, along with utility metrics capturing both intelligibility and perceptual naturalness.

4.1. Attack models

We report experiments performed using the usual three VPC-defined attack models described in [2, 34, 35], also illustrated in Figure 2. All three models involve comparisons between a pair of utterances, an original or anonymised enrolment utterance and an anonymised trial utterance.

For the *ignorant* (**I**) attack model [34], the adversary compares original enrolment and anonymised trial utterances with-

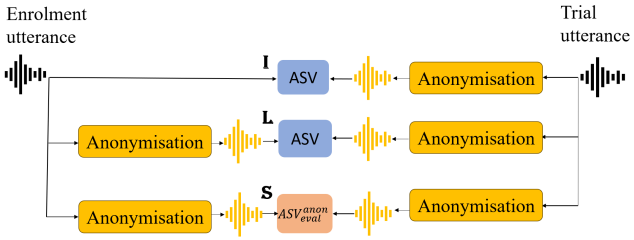

Figure 2: *Privacy evaluation in ignorant (I), lazy-informed (L) and semi-informed (S) attack models. In the S attack model, the attacker uses a retrained ASV system, ASV_{eval}^{anon} which is fine-tuned on anonymised data.*

out any compensation for anonymisation. Under the *lazy-informed* (**L**) attack model, the adversary makes a nominal effort to compensate for the use of anonymisation. To reduce domain mismatch, the adversary anonymises the enrolment utterance so that comparisons are made between anonymised enrolment and anonymised trial utterances. The adversary is assumed to have access to the same anonymisation system, but not the specific configuration [2]. The third, strongest model is the *semi-informed* (**S**) attack. In an effort to further reduce domain mismatch, the adversary now uses the same anonymisation system to produce anonymised data with which to retrain the ASV system, referred as $ASV_{\text{eval}}^{\text{anon}}$ in [2].

Results for **I** and **L** attack models are included to show the benefit to the adversary of focusing on non-timbral cues or, more specifically, information that remains after anonymisation is applied to remove predominantly timbral information. These two attack models are particularly informative because, unlike for the semi-informed **S** model, they support the evaluation of model robustness without the need for costly retraining using anonymised data.

4.2. Experimental Setup

We adopt the usual VPC metrics. These include the *equal error rate* (EER) as the *privacy* metric which reflects the potential to re-identify the original speaker using an ASV system. The *utility* metric is the *word error rate* (WER) which reflects the ability of the anonymisation system to preserve linguistic content. An effective anonymisation system should maintain WER close to that of original (non-anonymised) speech.

The EER is computed with two complementary ASV systems: (i) *E-VPC* which is the standard VPC evaluation system based on ECAPA-TDNN [31] and which captures primarily global *speaker* characteristics; and (ii) *W-NT* [2] [7], based on WavLM [23], which is designed to capture predominantly non-timbral speaker cues. The use of both systems allows us to determine the degree to which anonymisation removes timbral and residual non-timbral information.

The ASR model used to assess WER is the official fine-tuned *wav2vec2-large-960h-lv60-self* model [3] implemented using a SpeechBrain recipe.

In addition to the standard VPC-defined WER utility metric which might better reflect intelligibility, we adopt an additional metric with which to estimate the perceived naturalness of the anonymised speech. Instead of conducting large-scale listening tests, which are costly, and time-consuming, we estimate perceptual quality using two Mean Opinion Score (MOS) prediction models: WhiSQA [36] and UTMOS [37]. WhiSQA estimates MOS from feature representations extracted using Whisper [38]. UTMOS adopts a self-supervised framework that fine-tunes pretrained speech representations for direct MOS regression using large-scale subjective datasets.

Using both predictors allows cross-validated estimates of perceptual quality, which provides complementary insights to those provided by use of the standard EER and WER. This multi-metric evaluation enables a more comprehensive analysis of the privacy–utility trade-off.

5. Results & Analysis

Each proposed system, denoted as **DU-X-Y**, operates discrete speech units (DU) from HuBERT, refers to a speaker anonymisation system that jointly modifies speaker identity and rhythm. X indicates the speaker embedding for the phoneme duration predictor (NT or ECAPA), and Y the embedding for the Unit HiFi-GAN vocoder. When X = orig, no duration prediction is applied and the original phoneme durations are preserved and when X = pred, no speaker embedding is used for duration prediction. For example, DU-pred-ECAPA applies no conditioning in duration prediction and uses ECAPA for waveform synthesis.

Table 1 reports anonymisation performance for the most competitive VPC baselines and our proposed DU systems. Privacy is measured using the EER (the higher the better; an EER of 50% corresponds to chance-level speaker verification), while utility is assessed using the WER (the lower the better) and the two MOS predictors, UTMOS and WhiSQA (the higher the better, maximum 5).

5.1. VPC baselines and kNN Variants

E-VPC EERs (first results block) are substantially higher for all anonymisation systems than for original data (4.6%). Most systems achieve 44–49% EER for I/L attacks, approaching the 50% perfect privacy target. For W-NT (second results block), which emphasises non-timbral cues, EERs are generally lower and system differences become more apparent.

The kNN-VC baseline achieves high E-VPC EERs (42–44%) for I/L attacks while maintaining low WER (2.77%) and strong MOS scores (3.80–4.05). However, for W-NT, EERs drop to 17–19%, indicating that residual non-timbral information remains exploitable. Private kNN-VC improves W-NT EERs to 41–44%, confirming that privacy can be enhanced by explicitly targeting temporal and prosodic features. This improvement comes at the cost of a higher WER (5.00%) and the lowest MOS scores (2.69–3.10) among all models, reflecting the expected trade-off between stronger privacy and degraded utility.

5.2. Proposed DU Systems

Under I/L attack models, which show how well residual speaker information can be exploited without retraining, the proposed DU systems (second block of four rows) achieve consistently high EERs. For E-VPC, EERs remain around 46–49%, comparable to the strongest baseline systems (B4, B5, and Private kNN-VC), Under S attack model, DU-pred-nt surpasses Private kNN-VC by 7 absolute points in E-VPC EER.

When considering W-NT, EERs of between 33–43% are substantially higher than for kNN-VC (17–19%) and lower than Private kNN-VC (42–44%) in I/L attacks. However, In S at-

[2] https://huggingface.co/Orange/Speaker-wavLM-pro

[3] https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

Table 1: *EER (%) for ignorant (I), lazy-informed (L) and semi-informed (S) attack models under E-VPC and W-NT, and WER (%) on LibriSpeech test set, and MOS evaluation (UTMOS, WhisQA). Higher EER and lower WER indicate better anonymisation performance.* ***DU-X-Y*** *is a speaker anonymisation system based on discrete speech units, X is the speaker embedding used for the phoneme duration predictor (NT or ECAPA), while Y for the Unit HiFi-GAN synthesis.*

Anon. system	E-VPC (EER ↑)			W-NT (EER ↑)			WER ↓	MOS ↑	
	I	L	S	I	L	S		UTMOS	WhiSQA
Original	4.6			7.8			1.8	4.11	4.08
B4	47.8	49.5	27.3	38.2	34.7	17.4	5.90	3.42	3.87
B5	49.1	48.7	34.3	42.5	42.0	23.2	4.37	3.31	3.65
kNN-VC	42.1	43.8	11.0	19.0	17.6	5.9	2.77	3.80	4.05
Private kNN-VC	47.4	49.7	27.9	41.6	43.9	18.6	5.63	2.62	2.92
DU-orig-nt	47.6	47.2	35.9	33.1	33.6	22.6	4.48	3.59	3.70
DU-orig-ecapa	47.3	46.6	24.3	34.0	33.9	11.5	4.66	3.30	3.17
DU-pred-nt	47.2	47.9	34.8	39.9	42.3	24.6	7.41	3.63	3.69
DU-pred-ecapa	47.1	47.6	31.6	39.9	42.4	20.5	8.10	3.32	3.18
DU-nt-nt	48.2	48.1	34.1	40.5	42.5	23.7	7.73	3.63	3.69
DU-ecapa-ecapa	46.9	48.9	41.1	40.4	40.7	22.2	7.63	3.27	3.19
DU-ecapa-nt	47.6	47.9	36.6	40.3	43.0	24.1	7.59	3.60	3.70
DU-nt-ecapa	47.1	47.6	35.1	40.1	42.4	22.9	7.55	3.32	3.16

tack model, DU-pred-nt exceeds Private kNN-VC by 6 absolute points (32% relative). Such results highlight the DU systems' ability to generalise privacy protection beyond fully I/L attacks, a key consideration for real-world deployment.

Among DU variants, DU-pred-nt provides the strongest privacy protection for W-NT, suggesting that the prediction of rhythmic structure without conditioning on the original speaker embedding contributes to the removal of identity-related temporal cues. Variants using NT embeddings for vocoder synthesis tend to maintain slightly better perceptual quality, suggesting that embedding selection in this module critically impacts the balance between privacy and utility.

5.3. Privacy–Utility Trade-off

The improvements in non-timbral privacy come with expected trade-offs in utility. DU systems show higher WERs (7.41–8.10%) compared to Private kNN-VC (5.63%), and perceptual quality (MOS) is moderately reduced (UTMOS: 3.32–3.63, WhiSQA: 3.16–3.69 vs. 2.62–2.92 for Private kNN-VC). Notably, the gap in MOS between DU and Private kNN-VC is considerable, indicating that DU systems better preserve naturalness even when stronger anonymisation is applied. This illustrates that DU systems strike a favourable privacy–utility balance: they achieve higher W-NT EERs under S attack while maintaining reasonable intelligibility and perceptual quality, outperforming Private kNN-VC on both fronts. The results confirm that discrete speech unit-based anonymisation is particularly effective at mitigating non-timbral privacy risks, making it a strong candidate for applications where moderate utility reduction is acceptable in exchange for robust privacy.

Methods with higher W-NT EERs, reflecting stronger suppression of non-timbral cues, generally show increased WER and slightly lower MOS, while systems with low WER and high perceptual quality (e.g., kNN-VC) remain vulnerable to non-timbral attacks. Overall, DU systems deliver a consistent balance between privacy and utility, making them well-suited for applications demanding stronger anonymisation.

6. Ablation study

To better evaluate the relative contributions of each system module, we perform an ablation study focusing on two key components: (i) rhythm conversion through duration prediction, and (ii) the type of speaker embeddings used to condition the duration predictor and Unit HiFi-GAN vocoder.

6.1. Impact of duration modification

Temporal characteristics such as speaking rate, rhythm, and phonetic variability encode non-timbral speaker identity cues. Their modification is hence expected to enhance robustness to W-NT attacks. To quantify this effect, we compare the full system with variants that retain original temporal characteristics during synthesis. These correspond to DU-orig-nt and DU-orig-ecapa in Table 1, for which deduplicated discrete units are expanded using original, instead of predicted durations.

Results show that the use of original durations reduces privacy in the case of W-NT. For instance, DU-orig-nt achieves EERs of 33.1% and 33.6% for the I and L attacks, respectively, while the rhythm-converted system DU-pred-nt reaches 39.9% and 42.4%. This represents an improvement of 6–9% in absolute EER, demonstrating that duration modification effectively suppresses residual non-timbral speaker information. A similar trend is observed for ECAPA-conditioned systems, for which DU-orig-ecapa (≈34%) is outperformed by DU-pred-ecapa (≈40%) for the L attack, confirming the importance of rhythm conversion to the suppression of speaker-dependent temporal patterns. The use of original durations yields lower WERs compared to systems with predicted durations, illustrating a clear privacy–utility trade-off: rhythm conversion strengthens privacy at the cost of slightly degraded utility. These results confirm that duration prediction is critical to improve robustness to attacks exploiting non-timbral cues.

6.2. Effect of speaker embeddings

We analysed four embedding combinations: NT–NT, ECAPA–ECAPA, ECAPA–NT, and NT–ECAPA. Results in

Table 1 (last four rows) show that all configurations achieve similar privacy for E-VPC (47–49% EER), indicating effective removal of global speaker information regardless of embedding type. Systems using NT embeddings for the vocoder tend to achieve slightly higher W-NT EERs. In particular, DU-pred-nt reaches and EER of 42.3% for the L attack, which is among the best results for embedding-conditioned variants. This confirms again that NT embeddings better capture non-timbral characteristics and facilitate stronger obfuscation of rhythmic and prosodic identity cues during synthesis. However, the impact of the embedding used for the duration predictor appears relatively limited. Systems conditioned with ECAPA or NT embeddings show comparable W-NT EERs, and the differences are smaller than those observed when no speaker embedding is used.

In terms of speech utility, slightly higher MOS scores are observed when conditioning the vocoder with NT embeddings, suggesting NT embeddings may help preserve perceptual naturalness, implicitly improve subjective intelligibility while modifying speaker identity. Overall, results shows that performance of the proposed systems is agnostic to the choice of embedding. Nevertheless, NT embeddings provide a modest advantage when targeting non-timbral speaker cues, particularly when used to condition the vocoder.

7. Discussion

A central motivation of this work is the knowledge that voice identity is not encoded solely in timbral characteristics, but also in non-timbral attributes such as rhythm, speaking rate, and prosodic patterns. Results presented in Table 1 confirm this. While most anonymisation systems achieve EERs close to the 50% privacy target for E-VPC, which primarily captures global speaker characteristics, performance drops substantially for W-NT, which emphasises non-timbral cues.

Our approach addresses this drop by explicitly modifying the temporal structure of speech via duration prediction and modification. By predicting phoneme durations, the system performs rhythm conversion, replacing the temporal characteristics of the source speaker with those of a randomly selected target speaker. Our ablation study demonstrates that this component is critical to anonymisation performance. Systems using duration conversion consistently achieve higher W-NT EERs (40–43%) compared to variants using original durations (33–34%), confirming that temporal modification effectively obfuscates speaker-specific information encoded in rhythm and speaking style.

Depending on the target embedding used during inference, the duration predictor is expected to generate phoneme durations that follow different rhythmic speaking styles. In practice, however, our results indicate that the influence of the speaker embedding on rhythm generation remains limited, as predicted durations show only minor differences across embedding types. This behaviour may stem partly from the limited variability of speaking styles in the LibriSpeech test set. When no speaker embedding is used, the duration predictor produces an average rhythm learned from multispeaker training data, consistent with prior observations [15] that training on a single-speaker dataset reproduces that speaker's characteristic rhythm.

Results also illustrate the inherent trade-off between privacy and utility in voice anonymisation. Methods that provide better privacy by strongly modifying speech characteristics may also degrade intelligibility and naturalness. Results show that stronger protection against non-timbral attacks achieved through rhythm-conversion, are accompanied by increases to the WER, from $\approx$4.5% for systems using original durations to $\approx$7.6% using predicted durations. This may likely arise from temporal misalignments between discrete units and the reconstructed waveform. Even so, WERs remains within a range acceptable for most speech processing tasks.

Interestingly, despite higher WERs, objective MOS predictions suggest that perceptual quality remains relatively stable. This indicates that ASR-based intelligibility metrics and perceptual quality metrics are complementary and that moderate temporal distortions may affect automatic recognition without substantially degrading perceived naturalness. The higher naturalness reflected in MOS scores when conditioning the vocoder on NT embeddings rather than ECAPA embeddings may stem from the ability of NT embeddings to capture fine-grained phoneme-level information, including duration, stress, and intonation patterns. By preserving these subtle speech characteristics, the vocoder can generate smoother transitions and more natural speech contours, directly contributing to the observed increase in perceived quality.

We also acknowledge some limitations of our work. First, the duration predictor relies on discrete speech units obtained through k-means clustering. These units may not correspond perfectly to phonemes. As a consequence, predicted durations may occasionally produce imperfect temporal alignments, potentially contributing to increases in the WER. Second, our focus upon rhythm-based anonymisation targets only a subset of non-timbral speaker cues. Other non-timbral cues may remain and still reveal the speaker identity. Third, while perceptual quality appears to be preserved, we did not evaluate the preservation of other higher-level expressive attributes, such as emotion. However, preliminary observations suggest that emotional content is retained at levels comparable to Private kNN-VC. Fourth, W-NT–based privacy evaluation may disadvantage our anonymisation framework due to shared NT features; however, this constitutes a conservative setting, and the observed EER results still indicate effective non-timbral obfuscation. Anonymisation behaviour for spontaneous or more diverse speaking styles remains to be explored. In these cases, non-timbral speaker cues may be more prominent, more revealing of identity, and also more challenging to suppress.

8. Conclusion

In this paper, we introduce a joint speaker anonymisation framework that addresses a critical and often overlooked vulnerability in voice privacy: identity leakage through non-timbral cues. While timbral obfuscation is necessary, it is fundamentally incomplete if non-timbral remain unchanged. Our results show that rhythm conversion via phoneme duration prediction is essential, providing a substantial 6–9% absolute EER gains under non-timbral attacks compared to systems that retain original durations. The proposed approach also achieves competitive performance under standard evaluation protocols, even outperforming Private kNN-VC by 6% absolute EER points when the focus is on non-timbral cues under the strongest attack model. Importantly, these privacy gains are achieved with only moderate reductions in intelligibility and perceptual naturalness, as confirmed by MOS predictors. Future work will focus on further reducing the impact on intelligibility while maintaining strong privacy guarantees. Extending the framework to the exploration of privacy leaks from alternative non-timbral characteristics, such as accent or higher-level prosodic patterns, are also promising directions for further research.

9. Acknowledgment

This study was partly supported by the ANR-23-CE23-0018 EVA project.

10. References

[1] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, M. A. Abdelraheem, A. Abad, F. Teixeira, D. Matrouf, M. Gomez-Barrero, D. Petrovska-Delacrétaz, G. Chollet, N. Evans, T. Schneider, J.-F. Bonastre, B. Raj, I. Trancoso, and C. Busch, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441–480, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230818303875

[2] 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," 2024.

[3] H. L. Xinyuan, Z. Cai, A. Garg, K. Duh, L. P. García-Perera, S. Khudanpur, N. Andrews, and M. Wiesner, "Hltcoe jhu submission to the voice privacy challenge 2024," in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 61–66.

[4] 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," in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 67–71.

[5] N. Kuzmin, H.-T. Luong, J. Yao, L. Xie, K. A. Lee, and E.-S. Chng, "Ntu-npu system for voice privacy 2024 challenge," in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 72–79.

[6] W. Gu, Z. Liu, L. Chen, R. Wang, C. Guo, W. Guo, K. A. Lee, and Z.-H. Ling, "Ustc-polyu system for the voiceprivacy 2024 challenge," in *SPSC 2024*, 2024.

[7] R. Bakari, O. L. Blouch, N. Evans, N. Gengembre, M. Panariello, and M. Todisco, "The influence of non-timbral cues in voice anonymisation and evaluation," in *5th Symposium on Security and Privacy in Speech Communication*, 2025.

[8] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi, "The first voiceprivacy attacker challenge evaluation plan," *arXiv preprint arXiv:2410.07428*, 2024.

[9] Y. Zhang, Z. Bi, F. Xiao, X. Yang, Q. Zhu, and J. Guan, "Attacking voice anonymization systems with augmented feature and speaker identity difference," in *ICASSP*, 2025, pp. 1–2.

[10] X. Lyu, Y. Wang, T. Zhao, and H. Liu, "Fast adaptation of pre-trained speaker verification system for source speaker tracking," in *ICASSP*, 2025, pp. 1–2.

[11] C. O. Mawalim, A. Adila, and M. Unoki, "Fine-tuning titanet-large model for speaker anonymization attacker systems," in *ICASSP*, 2025, pp. 1–2.

[12] H. L. Xinyuan, A. Garg, Z. Cai, K. Duh, L. P. García-Perera, S. Khudanpur, N. Andrews, and M. Wiesner, "Hltcoe submission to the voiceprivacy attacker challenge," in *ICASSP*, 2025, pp. 1–2.

[13] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi, "The first voiceprivacy attacker challenge," in *ICASSP*, 2025, pp. 1–2.

[14] N. Tomashenko, E. Vincent, and M. Tommasi, "Exploiting Context-dependent Duration Features for Voice Anonymization Attack Systems," in *Interspeech*, 2025, pp. 5128–5132.

[15] C. Franzreb, A. Das, T. Polzehl, and S. Möller, "Private kNN-VC: Interpretable Anonymization of Converted Speech," in *Interspeech 2025*, 2025, pp. 3224–3228.

[16] O. L. Blouch, R. Bakari, and N. Gengembre, "Tuning DISSC for Voice Privacy Challenge 2024," in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 111–115.

[17] G. Maimon and Y. Adi, "Speaking style conversion in the waveform domain using discrete self-supervised units," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 8048–8061. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.541

[18] N. Gengembre, O. Le Blouch, and C. Gendrot, "Disentangling prosody and timbre embeddings via voice conversion," in *Interspeech 2024*. International Speech Communication Association, 2024.

[19] M. Panariello, N. Tomashenko, X. Wang, X. Miao, P. Champion, H. Nourtel, M. Todisco, N. Evans, E. Vincent, and J. Yamagishi, "The voiceprivacy 2022 challenge: Progress and perspectives in voice anonymisation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 32, p. 3477–3491, Jul. 2024. [Online]. Available: https://doi.org/10.1109/TASLP.2024.3430530

[20] P. Champion, "Anonymizing speech: Evaluating and designing speaker anonymization techniques," *Ph.D. dissertation, Université de Lorraine,*, 2023.

[21] H. Su, H. Zhang, X. Zhang, and G. Gao, "Convolutional neural network for robust pitch determination," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE Press, 2016, p. 579–583. [Online]. Available: https://doi.org/10.1109/ICASSP.2016.7471741

[22] M. Baas, B. van Niekerk, and H. Kamper, "Voice Conversion With Just Nearest Neighbors," in *Interspeech 2023*, 2023, pp. 2053–2057.

[23] 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*, vol. 16, no. 6, pp. 1505–1518, 2022.

[24] A. Sicherman and Y. Adi, "Analysing discrete self supervised speech representation for spoken language modeling," 06 2023, pp. 1–5.

[25] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 17 022–17 033. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/c5d736809766d46260d816d8dbc9eb44-Paper.pdf

[26] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, "Language-Independent Speaker Anonymization Approach Using Self-Supervised Pre-Trained Models," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 279–286.

[27] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," in *INTERSPEECH 2021 - Annual Conference of the International Speech Communication Association*, Brno, Czech Republic, Aug. 2021, in Proceedings of Interspeech 2021. [Online]. Available: https://inria.hal.science/hal-03329245

[28] Y. R. 0006, C. Hu, X. T. 0003, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA

[29] C. Franzreb, A. Das, T. Polzehl, and S. Möller, "Improving the speaker anonymization evaluation's robustness to target speakers with adversarial learning," 2026, preprint, accepted at ICASSP 2026. [Online]. Available: https://arxiv.org/abs/2508.09803

[30] J. Ni, T. Young, V. Pandelea, F. Xue, and E. Cambria, "Recent advances in deep learning based dialogue systems: a systematic survey," *Artif. Intell. Rev.*, vol. 56, no. 4, p. 3055–3155, Aug. 2022. [Online]. Available: https://doi.org/10.1007/s10462-022-10248-8

[31] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 3830–3834.

[32] F. Kreuk, A. Polyak, J. Copet, E. Kharitonov, T. A. Nguyen, M. Rivière, W.-N. Hsu, A. Mohamed, E. Dupoux, and Y. Adi, "Textless speech emotion conversion using discrete & decomposed representations," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 11 200–11 214. [Online]. Available: https://aclanthology.org/2022.emnlp-main.769/

[33] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Interspeech 2019*, 2019, pp. 1526–1530.

[34] 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.

[35] 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.

[36] G. Close, K. Hong, T. Hain, and S. Goetze, "Whisqa: Non-intrusive speech quality prediction using whisper encoder features," in *Speech and Computer*, A. Karpov and G. Gosztolya, Eds. Cham: Springer Nature Switzerland, 2026, pp. 39–51.

[37] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Interspeech 2022*, 2022, pp. 4521–4525.

[38] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.