SEF-MK: Speaker-Embedding-Free Voice Anonymization through Multi-k-means Quantization

Beilong Tang
Duke Kunshan University, China
beilong.tang@dukekunshan.edu.cn

Xiaoxiao Miao
Duke Kunshan University, China
xiaoxiao.miao@dukekunshan.edu.cn

Xin Wang
National Institute of Informatics, Japan
wangxin@nii.ac.jp

Ming Li
Duke Kunshan University, China
ming.li369@dukekunshan.edu.cn

***Abstract*—Voice anonymization protects speaker privacy by concealing identity while preserving linguistic and paralinguistic content. Self-supervised learning (SSL) representations encode linguistic features but preserve speaker traits. We propose a novel speaker-embedding-free framework called SEF-MK. Instead of using a single k-means model trained on the entire dataset, SEF-MK anonymizes SSL representations for each utterance by randomly selecting one of multiple k-means models, each trained on a different subset of speakers. We explore this approach from both attacker and user perspectives. Extensive experiments show that, compared to a single k-means model, SEF-MK with multiple k-means models better preserves linguistic and emotional content from the user's viewpoint. However, from the attacker's perspective, utilizing multiple k-means models boosts the effectiveness of privacy attacks. These insights can aid users in designing voice anonymization systems to mitigate attacker threats.** [1]

***Index Terms*—Voice anonymization, speaker embedding free, multi-k-means**

I. Introduction

Voice-based human-computer interaction is becoming increasingly prevalent, offering significant convenience in our daily lives. However, uploading raw audio recordings to social media without proper protection may lead to the leakage of personally identifiable information [1]. One form of additional protection is to encrypt utterances, thereby restricting access to unintended recipients and allowing only authorized users to recover the speech using a decryption key. However, this approach effectively prevents open communication and is not suitable for public-oriented applications. In most cases, users wish to convey the content and emotional tone of their speech while concealing their identity. This is where a Voice Anonymization System (VAS) has been proposed [2]–[5]. Specifically, VAS processes the original speech to produce anonymized outputs in which speaker-identifiable information is removed (privacy), while essential attributes, such as linguistic content and emotional expression, are preserved to enable the use of anonymized speech for various downstream tasks (utility). The anonymized speech can be shared openly, mitigating concerns over the misuse of the source speaker's identity.

There are two mainstream approaches to voice anonymization. Digital Signal Processing (DSP)-based methods are training-free approaches that modify speech characteristics from a speech production perspective to conceal the speaker's identity. Techniques include altering formants [6]–[8], changing speech speed [9], and modifying other vocal tract or voice source features [10]. However, DSP-based methods often suffer from content distortion and are generally ineffective against stronger attackers [2], [11], [12].

Deep Neural Network (DNN)-based methods, on the other hand, are generally more effective and draw on techniques from neural voice conversion and speech synthesis. A straightforward solution involves first transcribing speech into text using an automatic speech recognition (ASR) system, followed by re-synthesizing the speech via a text-to-speech (TTS) model [13]. While this ASR+TTS pipeline has been shown to effectively conceal the speaker's identity, it often compromises the utility of the original speech. Specifically, inevitable transcription errors and the loss of important paralinguistic cues, such as emotion, prosody, and accent, can degrade the quality and expressiveness of the anonymized output.

Widely used DNN-based approaches primarily leverage disentangled representation learning and rely on explicit speaker modeling, typically using pretrained speaker encoders. These approaches generally consist of three stages: (i) Speech disentanglement aims to separate speaker-specific information from the input speech. Some methods explicitly disentangle speaker, content, and prosody components [14]–[24], often extracting content information with self-supervised learning (SSL) models [25]–[27] and encoding speaker identity through a pretrained speaker encoder [28], [29]. Others, inspired by speech codec technology, segment speech into acoustic and semantic tokens [30], treating acoustic tokens as the primary carriers of speaker-specific characteristics. (ii) Speaker embedding anonymization involves replacing or transforming speaker-specific features with those of a pseudo speaker. Most approaches rely on an external speaker pool to construct the pseudo speaker [14], [22], [31]–[33]. (iii) Speech genera-

Xiaoxiao Miao is the corresponding author. This research is funded by DKU foundation project "Emerging AI Technologies for Natural Language Processing", and partially supported by JST, PRESTO Grant Number JPMJPR23P9. Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

[1] Code and audio samples can be found at https://github.com/Beilong-Tang/sef-mk

tion [34] synthesizes anonymized speech by combining the anonymized speaker (or acoustic) features with the preserved content and prosody (or semantic) information.

An emerging direction in DNN-based VAS research focuses on speaker-embedding-free neural methods [13], [35]–[37], which avoid the explicit modeling of speaker embeddings. In these approaches, utterances from both source and pseudo speakers typically pass through a shared encoding stage that extracts speech representations using SSL models. Anonymization is then achieved by replacing the SSL feature frames from the source speaker with those from the pseudo speaker's representation selected according to specific strategies, e.g. k-Nearest Neighbor (KNN) [38]. The resulting representations are then used to synthesize anonymized speech. However, since the SSL features contain speaker-related information, the k nearest neighbors and the anonymized waveform may also encode the traits of the source speaker. These systems have been shown to provide poor speaker privacy protection when facing strong attacking [13], [37].

This paper explores the **s**peaker-**e**mbedding-**f**ree VAS paradigm and anonymizes SSL representations through **m**ultiple **k**-means quantization, referred to as **SEF-MK**. The method comprises three key stages: (i) Encoding, where an SSL model-WavLM [26] generates continuous speech representations that capture linguistic content and (undesirably) speaker identity; (ii) Multi-k-means quantization, where SSL representations are quantized to suppress speaker-specific traits by randomly selecting a k-means model from multiple quantizers; (iii) Decoding, where high-quality anonymized speech is reconstructed using a Conformer-based decoder [39] and a HiFi-GAN vocoder [34], ensuring naturalness while preserving privacy.

Prior work, such as KNN-based VAS [13], [37], has explored speaker-embedding-free approaches, none have employed k-means in this context, despite its demonstrated effectiveness in suppressing speaker identity [21]. Although k-means has been applied in disentanglement-based methods [21], most studies use a single k-means model trained on a broad speaker population and primarily examine how varying the number of centroids affects the privacy-utility trade-off, lacking a thorough investigation of the specific role and effectiveness of k-means in anonymization.

This motivated us to pose two research questions, one from the user's perspective and the other from the attacker's, within the context of the VoicePrivacy Challenge (VPC) [5], which simulates a game-theoretic scenario between users and attackers. In this setting, users apply anonymization techniques to conceal speaker identities before publication, while attackers attempt to recover the original identities from the anonymized data. From the user's perspective, we explore strategies for training k-means models within the proposed SSL-based, speaker-embedding-free VAS paradigm by investigating the following question: *Does the composition of the k-means training data affect anonymization performance?* We construct a pool of k-means quantizers, each trained on a distinct subset of speakers using different speaker grouping strategies, such as training one model per individual speaker or training models on subsets of multiple speakers. We then investigate how these strategies affect anonymization performance from the user/defender's perspective. Through extensive experiments, we found that compared to a single k-means model, from the user's perspective, the use of multiple k-means models in the proposed **SEF-MK** system better preserves linguistic and emotional attributes.

From the attacker's perspective, we investigate *how the attacker's composition of the k-means degrade the anonymization performance*. We found that employing multiple k-means models can enhance the effectiveness of privacy attacks, regardless of which k-means strategy the user adopts. We hope these findings can be helpful for users when designing the VAS against attackers with various attacking strategies.

II. Speaker-Embedding-Free Voice Anonymization

In this section, we first provide an overview of the voice anonymization problem formulation and introduce two recently proposed speaker-embedding-free methods, which serve as baselines for this study. We then present the novel **SEF-MK** architecture, which leverages WavLM, k-means, Conformer, and HiFi-GAN to explore various strategies for training multiple k-means models to conceal the original speaker identity.

A. Voice Anonymization Problem Formulation

An anonymization system transforms an original speech utterance $\mathbf{X} = [x_1, \ldots, x_{\tilde{T}}]$ into an anonymized version $\mathbf{Y} = [y_1, \ldots, y_{\tilde{T}}]$, where $\tilde{T}$ is the length of the utterance. An ideal anonymization system should preserve linguistic content and emotional cues[2] while removing the speaker identity. This objective can be formally defined as:

$$\mathbf{Y} = f(\mathbf{X}), \quad \text{s.t.} \begin{cases} \texttt{ASV}(\mathbf{X}) \neq \texttt{ASV}(\mathbf{Y}) \\ \texttt{ASR}(\mathbf{X}) \approx \texttt{ASR}(\mathbf{Y}) \\ \texttt{SER}(\mathbf{X}) \approx \texttt{SER}(\mathbf{Y}) \end{cases} \tag{1}$$

The transformation function $f(\cdot)$ should simultaneously satisfy three competing objectives: (i) distort speaker identity, as measured by automatic speaker verification (ASV) systems, (ii) preserve linguistic content, as evaluated by an automatic speech recognition (ASR) system, and (iii) maintain emotional content, as assessed by a speech emotion recognition (SER) systems.

B. KNN-based VAS

Among speaker-embedding-free approaches, one popular approach is based on KNN [38] with encoding, anonymization, and decoding stages.

In the encoding stage, both the source and target/pseudo speaker utterances go through the same pretrained WavLM model [26] to extract SSL representations. In the anonymization stage, the goal is to remove speaker-specific information from the source WavLM representation while preserving other

[2] While paralinguistic attributes include not only emotion, this study follows the setup of Voice Privacy Challenge 2024 [5] and only considers emotion.

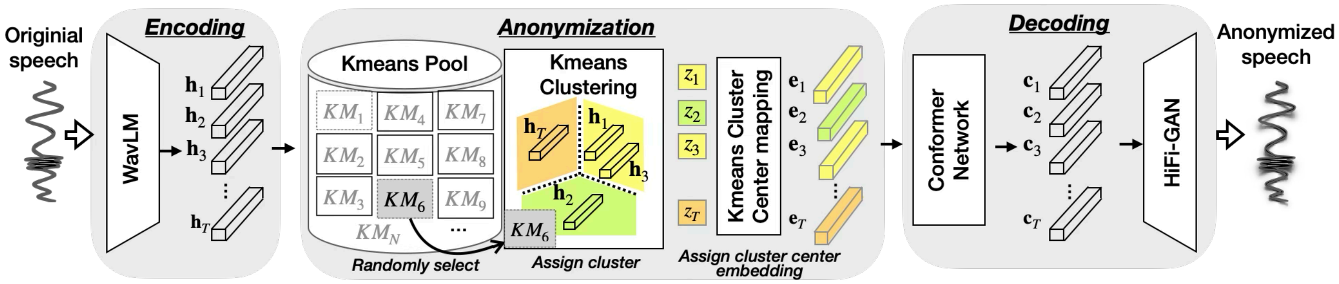


Fig. 1. Framework of multiple k-means-based speaker embedding-free voice anonymization.

speech attributes. This is achieved by replacing each frame of the source representation with the most similar frames from the target speaker. Specifically, for each frame in the source utterance, a KNN search is performed to identify the top-k most similar frames in the target speaker's representation. In the decoding stage, the average of these top-k target frames, ideally devoid of the original speaker identity but still conveying the speech content and emotion, is passed to a HiFi-GAN vocoder [34], which synthesizes anonymized speech that retains the linguistic content of the source while adopting the vocal characteristics of the target speaker.

However, these systems have been shown to provide poor speaker privacy protection when facing stronger attacks [13], [37]. One possible reason is that SSL features are known to encode both content and speaker information. Consequently, although KNN retrieves the most similar frames from the target speaker, aiming to remove the timbre of the source speaker, the retrieved frames may still preserve speaker-specific cues to reflect the source speaker's habitual speaking style, thereby leaking information about the source speaker under stronger attacks [13], [37]. A variant of the KNN-based model [40] extends the previously described KNN approach by introducing two interpretable components that anonymize the duration and variation of phonemes to enhance privacy. However, this improvement in privacy comes at the cost of reduced utility.

C. Proposed SEF-MK VAS

As explained in Section II-B, KNN-based speaker-embedding-free methods have not utilized k-means, despite its proven ability to suppress speaker identity [21]. While k-means has been applied in disentanglement-based approaches, prior work typically uses a single k-means model trained on the full dataset and focuses mainly on the number of centroids, without thoroughly exploring its role in voice anonymization. To address this gap, we propose a novel speaker-embedding-free framework that employs multiple k-means quantizers, as shown in Figure 1.

In the encoding stage, the input speech waveform X is processed by WavLM [26] to extract SSL representations $h(X) = \mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_T]$, where each $\mathbf{h}_t \in \mathbb{R}^{1024}$ captures comprehensive speech characteristics, and T is the number of frames in the input utterance. These SSL-based features inherently encode linguistic content, speaker identity, and emotional state [21], [41].

In the anonymization stage, the goal is to transform the SSL features to suppress speaker identity while preserving linguistic content and emotional expression. We observe that existing approaches [21] typically train a single k-means model on the entire data with a large set of speakers. However, because this clustering captures both phonetic and inter-speaker variations, the resulting centroids can unintentionally encode speaker-specific information, potentially leaking source speaker identity. This raises an important but underexplored question: *Does the composition of the k-means training data affect anonymization performance?* Intuitively, when k-means is trained on a single speaker's data, the clustering focuses solely on intra-speaker phonetic variation, inherently avoiding the encoding of speaker identity. However, applying such a speaker-specific quantizer to features from a different speaker may introduce a mismatch that degrade the linguistic content. Using multiple k-means models may either improve the the extraction of linguistic content (i.e., better utility) or leaking speaker information (i.e., worse privacy).

To verify the above assumptions, the anonymization stage introduces a pool of N specialized k-means models $\{KM_1, \ldots, KM_N\}$, where each model is trained on a distinct subset of speakers to encourage diverse clustering behavior. Each model uses $K = 1024$ clusters. During inference, a model KM_n is randomly selected from the pool and applied to the SSL features $\mathbf{H}$ to assign each frame to a cluster, producing discrete assignments $\mathbf{Z} = [z_1, \ldots, z_T]$ with $z_t \in \{1, \ldots, K\}$, $\forall t \in \{1, \ldots, T\}$. Each assignment z_t is then mapped to its corresponding cluster center embedding vector, yielding $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_T]$, where $\mathbf{e}_t \in \mathbb{R}^{1024}$ represents the 1024-dimensional centroid of cluster z_t. This operation is denoted as $\mathbf{E} = \text{Center}(\mathbf{Z})$.

We explore several strategies for constructing the pool of k-means models based on different speaker groupings in the training dataset. Let the dataset D contain speech from S speakers, and let L be the number of speakers per group, with $L < S$. We define the following strategies:

- D**-all**: All utterances from all S speakers are used to train a single k-means model.
- D-L**-sep**: The dataset is partitioned into groups of L speakers, with each group used to train a separate k-means model, resulting in $\left\lfloor \frac{S}{L} \right\rfloor$ number of k-means mod-

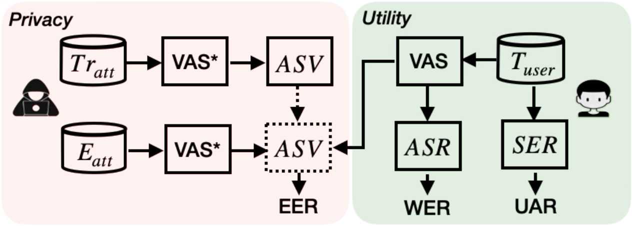


Fig. 2. Evaluation protocol following the VoicePrivacy 2024 guidelines. The subscript *att* means attacker. The databases are Tr_{att}: *libri-train-360*; E_{att}: *libri-dev-enroll* and *libri-test-enroll*; T_{user}: *libri-dev-trial-f*, *libri-dev-trial-m*, *libri-test-trial-f*, *libri-test-trial-m*, *IEMOCAP-dev*, and *IEMOCAP-test*. If VAS* = VAS, this represents a full attacker; if VAS* $\cap$ VAS $\neq \emptyset$, this represents a semi-attacker with partial knowledge of the user's VAS.

els. Note that D-1-sep denotes training a single k-means model for each speaker.

In our experiments, we evaluate the effects of different datasets and speaker group sizes. Specifically, we use either LibriSpeech-train-clean-460[3] with 1,172 speakers ($S = 1,172$) or VoxCeleb2 [43] with 5,994 speakers ($S = 5,994$) as the dataset for D, and we conduct experiments using speaker group sizes $L \in \{1, 10, 20\}$.

The decoding stage reconstructs natural speech from the anonymized embeddings E using a Conformer-based sequence model followed by a HiFi-GAN vocoder. The Conformer architecture combines self-attention mechanisms for capturing long-range dependencies with convolutional operations for local pattern modeling, transforming the input embeddings $\mathbf{E}$ into continuous representations $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_T]$, where each $\mathbf{c}_t \in \mathbb{R}^{1024}$ corresponds to a 1024-dimensional frame-level feature vector. Note that during Conformer training, instead of using a single general k-means model trained on the entire training set, we use the matched k-means model trained on the same speaker as the input feature generator for the Conformer, aiming to best resynthesize the original speech.

We utilize the discrete representations (i.e., centroids of k-means models) as the input and the continuous representation as the target for the Conformer. The HiFi-GAN vocoder subsequently converts these intermediate representations into waveform samples, completing the anonymization pipeline while preserving natural speech characteristics.

III. Experiments

A. Evaluation Protocol

We follow the VPC 2024 evaluation protocol [5] to assess the effectiveness of our proposed VAS as plotted in Figure 2.

1) Privacy Evaluation: The attackers are assumed to have access to a few original or anonymized utterances for each speaker, referred to as *enrollment* utterances and denoted as E_{att}, as well as some knowledge of the VAS. In the case of **SEF-MK**, the attackers are assumed to use a different k-means pool from the one used by the users during anonymization. This setup is referred to as a *semi-attacker*. If the attacker and the user share the same k-means pool, we refer to it as a *full attacker* scenario. Note that the VPC 2024 focuses on the *semi-attacker* scenario. In the following experiments, we also examine the *full attacker* scenario from the user's point of view to understand the worst-case situation.[4]

The attackers employ an ASV model, specifically, the ECAPA-TDNN model [29], trained on anonymized speech, denoted as Tr_{att}, to reduce the mismatch between original and anonymized utterances and infer the speaker's identity. The Equal Error Rate (EER) is used to evaluate the privacy protection capability of the VAS, with an EER close to 50% indicating perfect privacy protection.

2) Utility Evaluation: The utility evaluation depends on the downstream tasks. We follow VPC 2024 and consider ASR and emotion analysis. The evaluation of speech content and emotion preservation of anonymized *test trial* speech, denoted as T_{user}, is straightforward. The speech content preservation ability in anonymized speech is assessed by the word error rate (WER) computed using an ASR evaluation model[5]. A lower WER, similar to that of the original speech, indicates good speech content preservation ability. The emotion preservation ability in anonymized speech is assessed by unweighted average recall (UAR) produced by a pre-trained speech emotion recognizer (SER), which is a wav2vec2-based system [5]. A higher UAR, similar to that of the original speech, indicates better emotional content preservation ability.

B. Datasets and System Configurations

The **SEF-MK** VAS is built using the publicly available WavLM-large[6]. The Conformer encoder consists of 6 layers with a kernel size of 31, each with 4 attention heads, and an MLP with a hidden dimension of 2048, trained on *LibriSpeech-train-clean-460* [42]. The HiFi-GAN is trained on *LibriTTS-train-clean-100* [44]. This training setup is compatible with the VPC 2024 evaluation protocol. VASs are evaluated on the official VPC 2024 dev and test sets [3]. It contains English utterances by several female and male speakers from the *LibriSpeech* corpora, split into dev and test sets.

IV. Results and Discussion

In this section, we begin from the user's point of view to find the best configurations for the proposed **SEF-MK**, focusing on how the composition of the k-means training data affects performance under the *full-attacker* scenario, where the attacker shares the same k-means pool as the user, i.e., a worst-case (but unrealistic) setting. After determining the optimal user settings, we shift to the attacker's point of view to investigate how the attacker can exploit the composition of the k-means training data. When the attacker isolates the k-means pool, this corresponds to evaluating **SEF-MK** under the standard VPC scenario, i.e., the *semi-attacker* setting. Finally, we compare the results with those of other VASs under both fully and semi-attacking scenarios.

[3] *LibriSpeech-train-clean-360* and *LibriSpeech-train-clean-100* datasets [42]

[4] It is useful to measure system performance in the worst-cast situation, but it is unlikely to happen in applications if the users or vendors use a k-means pool constructed upon undisclosed data that the attacker cannot access.

[5] https://huggingface.co/speechbrain/asr-wav2vec2-librispeech

[6] https://huggingface.co/microsoft/wavlm-large

TABLE I

RESULTS (IN %) FOR VARIOUS K-MEANS TRAINING STRATEGIES APPLIED TO THE USER ON DEVELOPMENT SETS UNDER THE FULL-ATTACKER SCENARIO. **LIBRI-ALL** DENOTES A SINGLE K-MEANS MODEL TRAINED ON THE ENTIRE DATASET. **D-L-SEP** DENOTES TRAINING ONE K-MEANS MODEL PER **L** SPEAKERS. **LIBRI** REFERS TO THE *LibriSpeech-train-clean-460* DATASET WITH 1,172 SPEAKERS, AND **VOX** REFERS TO THE *VoxCeleb2* DEV DATASET WITH 5,994 SPEAKERS.

		EER ↑		WER ↓	UAR ↑
	#k-means	dev-f	dev-m	dev	dev
Original	-	10.51	0.93	1.80	69.08
Resyn	-	17.33	6.55	3.47	59.99
Libri-all	1	22.59	13.34	6.06	49.94
Libri-20-sep	58	**23.86**	11.36	**3.33**	**56.89**
Libri-10-sep	117	21.13	11.80	3.37	55.50
Libri-1-sep	1,172	21.71	15.84	3.99	48.01
Vox-all	1	19.04	**16.00**	4.98	48.99
Vox-20-sep	299	18.89	14.29	4.58	49.23
Vox-10-sep	599	19.32	12.27	4.65	48.66
Vox-1-sep	5,994	22.73	15.96	7.64	45.67

TABLE II

EER RESULTS (%) FOR VARIOUS K-MEANS TRAINING STRATEGIES APPLIED TO THE ATTACKER ON TEST AND DEVELOPMENT SETS UNDER THE *semi-attacker* SCENARIO.

		EER (%) ↑			
	#k-means	dev-f	dev-m	test-f	test-m
Original		10.51	0.93	8.76	0.42
User uses 58 k-means models: Libri-20-sep					
Libri-all	1	42.90	35.22	38.53	35.38
Vox-all	1	28.95	25.48	27.19	28.06
Vox-1-sep	5,994	25.98	14.44	15.16	15.59
User uses 1,172 k-means models: Libri-1-sep					
Libri-all	1	44.62	41.17	42.93	39.68
Vox-all	1	37.20	37.57	38.14	37.17
Vox-1-sep	5,994	26.56	19.56	16.42	17.60
User uses one k-means model: Libri-all					
Vox-all	1	34.20	36.01	37.03	37.39
Libri-1-sep	1,172	23.31	16.62	16.44	17.87
Vox-1-sep	5,994	26.14	20.34	18.40	19.12

A. How can users configure SEF-MK under the full-attacker scenario

Under this scenario, assuming the worst case where the VAS is fully shared with the attacker, the user aims to find a configuration that maximizes privacy (high EER) while maintaining good utility (low WER and high UAR).

We experiment with different k-means training partition strategies on the development sets, with the results summarized in Table I. At the top of the table, we evaluate the performance of the resynthesized **SEF-MK**, denoted as 'resyn', where each utterance uses its own data to train the corresponding k-means model and generate speech. Performance that closely matches the original speech indicates better generation quality. For 'resyn', the EER is close to that of the original speech, while both WER and UAR show reasonable degradations, reflecting the trade-offs introduced by the waveform resynthesis process. This result is consistent with those observed on other mainstream anonymization systems [51].

The middle and bottom sections of the table evaluate different training partition strategies using two datasets: the smaller *LibriSpeech-train-clean-460* dataset with 1,172 speakers, and the larger *VoxCeleb2* dev dataset with 5,994 speakers. For both datasets, we did not observe significant changes in EER when switching from a single k-means model (Libri-all, Vox-all) to multiple models (Libri-20/10/1-sep, Vox-20/10/1-sep). However, WER and UAR consistently improve when using multiple k-means models. The best WER and UAR results are achieved when the k-means models are trained on data from 20 speakers per model, for both the Libri and Vox datasets.

When comparing results between the Libri and Vox datasets under the same configuration, we observe that using the larger dataset (Vox) slightly increases the EER, particularly for male speech, but the utility metrics drop more significantly. One potential reason could be that the VoxCeleb2 data, which was sourced from Celebrities' interviews in the wild, is considerably more noisy than the audiobook-based Librispeech data. Relatively clean data may be preferred for building the k-means models that can encode the linguistic contents. Overall, the Libri-20-sep configuration provides a better trade-off between privacy and utility.

B. How can attacker configure SEF-MK under the semi-attacker scenario

After exploring the configuration of **SEF-MK** from the user's perspective, it is also important to examine the attacker's configuration. The attacker's goal is to use a similar VAS to anonymize speech in a way that mimics the user's approach, optimize the ASV model trained on anonymized speech, and attempt to trace the original speaker identity from the anonymized speech. This only impacts the privacy aspect of the system. Hence, Table II presents the EER results on both development and test sets for various attacker configurations, assuming fixed user settings.

We first examine the attackers' performance, assuming users using 'Libri-20-sep' (the best configuration for users as reported in the section IV-A). At the top of the table, we list various attacker configurations. Interestingly, we observe that when the attacker uses only a single k-means model to generate anonymized speech, the resulting EER is significantly higher, e.g., over 35% for 'Libri-all' and over 25% for 'Vox-all', indicating poor attacking effectiveness (and good privacy protection from the user's point of view). However, when the attacker employs multiple k-means models, even with different training data (e.g., Vox-1-sep), the attack becomes much more effective, reducing the EER to around 15%. We confirm this trend by testing under different user settings, as shown in the middle and bottom sections of the table, and consistently observe the same pattern.

One possible reason is that using randomly selected multiple k-means models produces more diverse anonymized utterances compared to using a single k-means model, which benefits the attacker. The attacker's ASV model trained on more diverse anonymized utterances may perform better in discriminating

TABLE III
RESULTS (%) ON VARIOUS VASS. B1-B6 DENOTES THE SIX BASELINE MODELS IN VPC 2024.

		EER ↑					WER ↓			UAR ↑		
		dev-f	dev-m	test-f	test-m	avg	dev	test	avg	dev	test	avg
	Original	10.51	0.93	8.76	0.42	5.66	1.80	1.85	1.83	69.08	71.06	70.07
Disentangle	B1 [45]	10.94	7.45	7.47	4.68	7.64	3.07	2.91	2.99	42.71	42.78	42.75
	B3 [46]	28.43	22.04	27.92	26.72	26.78	4.29	4.35	4.32	38.09	37.57	37.83
	B4 [47]	34.37	31.06	29.37	31.16	31.99	6.15	5.90	6.03	41.97	42.78	42.38
	B5 [48]	35.82	32.92	33.95	34.73	34.36	4.73	4.37	4.55	38.08	38.17	38.13
	B6 [48]	25.14	20.96	21.15	21.14	22.10	9.69	9.09	9.39	36.39	36.13	36.26
	OH [49]	44.89	34.74	39.26	37.64	38.54	2.36	2.48	2.42	47.01	47.37	47.19
DSP	B2 [50]	12.91	2.05	7.48	1.56	6.00	10.44	9.95	10.20	55.61	53.49	54.55
SEF	KNN* [35]	18.35	13.66	16.24	12.50	15.19	2.99	2.96	**2.98**	47.70	50.61	49.16
	Private KNN* [40]	-	-	-	-	**49.40**	-	-	4.80	-	-	49.40
	SEF-MK	42.90	35.22	38.53	35.38	38.01	3.33	3.30	3.31	56.89	58.31	**57.60**

*Note that these systems use the same target speaker pool for both the user and the attacker, more likely a *full attacker* scenario.

incoming anonymized utterances and exploit more speaker-specific attributes that are not fully removed by **SEF-MK**.

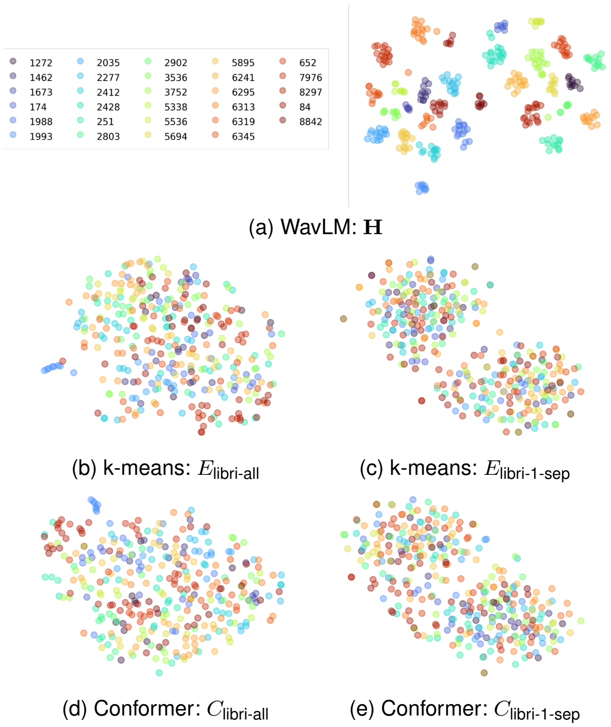
C. Comparasion with other VASs

Finally, we select the best semi-attacker configuration of our system where the user uses 'libri-20-sep', and the attacker uses 'Libri-all' as the k-means pool, and compare it with other VASs, as shown in Table III. Among the disentanglement- and DSP-based VASs, speaker-embedding-free (SEF) KNN-based methods achieve decent performance even under full attacker scenarios. However, **SEF-MK** achieves the highest UAR while maintaining a good balance between privacy protection and content preservation.

Note that the private KNN system [40] obtained the highest EER likely because of the additional component to anonymize the duration of the phones, but it may also have affected the utility of the anonymized speech. The anonymization of the durations can also be incorporated in our proposed system, but this is left to future work.

D. Furthur analysis

To further visualize the effectiveness of the proposed **SEF-MK**, we use t-SNE [52] to visualize the k-means and Conformer output embeddings for both the single-model setting (*libri-all*) and the multi-k-means model setting (*libri-1-sep*) as shown in Figure 3. The speaker embeddings are extracted from 29 speakers in the *libri-dev-enroll* dataset, with each speaker represented by a different color. In Figure 3(a), the WavLM features form clear and distinct speaker clusters, indicating that they encode a significant amount of speaker identity-related information. However, after applying k-means and the Conformer, these clusters become indistinct, and speaker separability is no longer clearly visible. Additionally, the embedding distributions before and after the Conformer remain similar, demonstrating its strong feature reconstruction capability. Furthermore, compared to the single k-means model, the multiple k-means models also result in embeddings with less speaker information, but they form two clusters (which probably represent the genders). This echoes with the results in Sec. IV-A that *libri-1-sep* degraded the privacy protection (i.e., higher EER) for one of the gender.



(a) WavLM: $\mathbf{H}$

(b) k-means: $E_{\text{libri-all}}$

(c) k-means: $E_{\text{libri-1-sep}}$

(d) Conformer: $C_{\text{libri-all}}$

(e) Conformer: $C_{\text{libri-1-sep}}$

Fig. 3. *Libri-dev-enroll:* Comparison of WavLM representations and clustering-based embeddings across speaker and separation settings. Different speaker identities are marked using different colors.

V. CONCLUSION

In this work, we propose a speaker-embedding-free voice anonymization system that leverages multi-k-means quantization on SSL features. Specifically, our method utilizes a pool of k-means models, with one randomly selected at either the frame level or the utterance level during inference. We systematically analyze the impact of different k-means anonymization strategies from both the user's and the attacker's perspectives. Experimental results demonstrate that our approach achieves superior utility preservation while providing strong privacy protection, outperforming various existing VASs.

References

[1] "General data protection regulation (GDPR)," https://gdpr.eu/what-is-gdpr.

[2] 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*, 2022.

[3] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J. F. Bonastre, "The VoicePrivacy 2022 Challenge evaluation plan," *arXiv preprint arXiv:2203.12468*, 2022.

[4] 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*, pp. 1–14, 2024.

[5] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and M. Todisco, "The VoicePrivacy 2024 challenge evaluation plan," *arXiv preprint arXiv:2404.02677*, 2024.

[6] P. Gupta, G. P. Prajapati, S. Singh, M. R. Kamble, and H. A. Patil, "Design of voice privacy system using linear prediction," in *Proc. APSIPA ASC*. IEEE, 2020, pp. 543–549.

[7] S. P. Dubagunta, R. Van Son, and M. M. Doss, "Adjustable deterministic pseudonymisation of speech: Idiap-NKI's submission to VoicePrivacy 2020 challenge," *URL: https://www. voiceprivacychallenge. org/docs/Idiap-NKI. pdf*, 2020.

[8] S. E. McAdams, *Spectral fusion, spectral parsing and the formation of auditory images*. Stanford university, 1984.

[9] C. O. Mawalim, S. Okada, and M. Unoki. (2022) System description: Speaker anonymization by pitch shifting based on time-scale modification (PV-TSM). [Online]. Available: https://www.voiceprivacychallenge.org/results-2022/docs/1___T32.pdf

[10] L. Tavi, T. Kinnunen, and R. G. Hautamäki, "Improving speaker de-identification with functional data analysis of f0 trajectories," *Speech Communication*, vol. 140, pp. 1–10, 2022.

[11] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the VoicePrivacy Initiative," in *Proc. Interspeech*, 2020, pp. 1693–1697.

[12] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," in *Proc. ICASSP*. IEEE, 2020, pp. 2802–2806.

[13] 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 *Proc. 4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 61–66.

[14] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," *Proc. 10th ISCA Speech Synthesis Workshop*, pp. 155–160, 9 2019.

[15] S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and N. T. Vu, "Prosody is not identity: A speaker anonymization approach using prosody cloning," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[16] A. S. Shamsabadi, B. M. L. Srivastava, A. Bellet, N. Vauquier, E. Vincent, M. Maouche, M. Tommasi, and N. Papernot, "Differentially private speaker anonymization," *Proceedings on Privacy Enhancing Technologies*, vol. 2023, no. 1, Jan. 2023. [Online]. Available: https://hal.inria.fr/hal-03588932

[17] C. O. Mawalim, K. Galajit, J. Karnjana, S. Kidani, and M. Unoki, "Speaker anonymization by modifying fundamental frequency and x-vector singular value," *Computer Speech & Language*, vol. 73, p. 101326, 2022.

[18] P. Champion, "Anonymizing speech: Evaluating and designing speaker anonymization techniques," *arXiv preprint arXiv:2308.04455*, 2023.

[19] J. Yao, Q. Wang, P. Guo, Z. Ning, Y. Yang, Y. Pan, and L. Xie, "Musa: Multi-lingual speaker anonymization via serial disentanglement," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 1664–1674, 2025.

[20] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, "Language-independent speaker anonymization approach using self-supervised pre-trained models," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 279–286.

[21] ——, "Speaker anonymization using orthogonal Householder neural network," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 31, pp. 3681–3695, 2023.

[22] J. Yao, Q. Wang, P. Guo, Z. Ning, and L. Xie, "Distinctive and natural speaker anonymization via singular value transformation-assisted matrix," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[23] S. Meyer, F. Lux, and N. T. Vu, "Probing the feasibility of multilingual speaker anonymization," *Proc. Interspeech*, 2024.

[24] X. Miao, R. Tao, C. Zeng, and X. Wang, "A benchmark for multi-speaker anonymization," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 3819–3833, 2025.

[25] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[26] 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.

[27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, vol. 33, 2020, pp. 12 449–12 460.

[28] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 5329–5333.

[29] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[30] M. Panariello, F. Nespoli, M. Todisco, and N. Evans, "Speaker anonymization using neural audio codec language models," in *Proc. ICASSP*, 2024, pp. 4725–4729.

[31] B. M. L. Srivastava, N. A. Tomashenko, X. Wang, E. Vincent, J. Yamagishi, M. Maouche, A. Bellet, and M. Tommasi, "Design choices for x-vector based speaker anonymization," in *Proc. Interspeech*, 2020, pp. 1713–1717.

[32] B. M. L. Srivastava, M. Maouche, M. Sahidullah, E. Vincent, A. Bellet, M. Tommasi, N. Tomashenko, X. Wang, and J. Yamagishi, "Privacy and utility of x-vector based speaker anonymization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2383–2395, 2022.

[33] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, "Analyzing language-independent speaker anonymization framework under unseen conditions," in *Proc. Interspeech*, 2022, pp. 4426–4430.

[34] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, 2020, pp. 17 022–17 033.

[35] Z. Cai, H. L. Xinyuan, A. Garg, L. P. García-Perera, K. Duh, S. Khudanpur, N. Andrews, and M. Wiesner, "Privacy versus emotion preservation trade-offs in emotion-preserving speaker anonymization," in *Proc. SLT*. IEEE, 2024, pp. 409–414.

[36] S. Ghosh, M. Jouaiti, A. Das, Y. Sinha, T. Polzehl, I. Siegert, and S. Stober, "Anonymising elderly and pathological speech: Voice conversion using ddsp and query-by-example," *Proc. Interspeech*, 2024.

[37] A. Das, C. Franzreb, T. Herzig, P. Pirlet, and T. Polzehl, "Comparing speech anonymization efficacy by voice conversion using knn and disentangled speaker feature representations," in *Proc. SPSC 2024*, 2024, pp. 121–126.

[38] M. Baas, B. van Niekerk, and H. Kamper, "Voice conversion with just nearest neighbors," in *Proc. Interspeech*, 2023, pp. 2053–2057.

[39] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *Proc. Interspeech*, 2020.

[40] C. Franzreb, A. Das, T. Polzehl, and S. Möller, "Private knn-vc: Interpretable anonymization of converted speech," in *Proc. Interspeech*, 2025.

[41] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K.-t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[42] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP*. IEEE, 2015, pp. 5206–5210.

[43] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[44] 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 *Proc. Interspeech*, 2019, pp. 1526–1530.

[45] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019, pp. 155–160.

[46] S. Meyer, P. Tilli, P. Denisov, F. Lux, J. Koch, and N. T. Vu, "Anonymizing speech with generative adversarial networks to preserve speaker privacy," in *Proc. SLT*, 2023, pp. 912–919.

[47] M. Panariello, F. Nespoli, M. Todisco, and N. Evans, "Speaker anonymization using neural audio codec language models," in *Proc. ICASSP*, 2024, pp. 4725–4729.

[48] P. Champion, "Anonymizing speech: Evaluating and designing speaker anonymization techniques," *arXiv preprint arXiv:2308.04455*, 2023.

[49] X. Miao, Y. Zhang, X. Wang, N. Tomashenko, D. C. L. Soh, and I. Mcloughlin, "Adapting general disentanglement-based speaker anonymization for enhanced emotion preservation," *Computer Speech & Language*, p. 101810, 2025.

[50] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the mcadams coefficient," in *Proc. Interspeech*, 2021, pp. 1099–1103.

[51] M. Panariello, M. Todisco, and N. Evans, "Vocoder drift compensation by x-vector alignment in speaker anonymisation," in *Proc. 3rd Symposium on Security and Privacy in Speech Communication*, 2023, pp. 16–20.

[52] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of machine learning research*, vol. 9, no. 11, 2008.