A Speech Steganography Privacy Protection Scheme Based on Secret Sharing

Changxiang Zhao, Jianping Cai, Ximeng Liu, Qi Zhong and Zuobin Ying

(Corresponding author: Jianping Cai)

Faculty of Data Science, City University of Macau, Macau 999078, China

(Email:jpcai@cityu.edu.mo)

***Abstract*—Speech data stored in the cloud are vulnerable to attribute inference attacks, which can lead to the leakage of sensitive attribute information. This paper proposes a novel privacy-preserving scheme that combines secret sharing and steganography to enhance Privacy Protection for speech data. Experimental results demonstrate that the proposed method effectively defends against attribute inference attacks and enhances robustness against malicious interference and data corruption. It achieves an optimal balance between privacy protection and data availability by ensuring the successful reconstruction of the original speech without compromising security.**

Keywords: Secret Sharing, steganography, Defends against attribute inference attacks, Enhance Privacy Protection,

I. Introduction

In smart voice services, there is frequently a requirement to employ sensitive features involving private user attributes to ensure a more convenient service. Many enterprises contemplate the utilization of cloud servers for the storage of user voice-sensitive data features [1]. This approach is driven by the desire to circumvent the constraints imposed by limited computing resources and the risk of direct theft of voice data when storing it locally [2]. Adopting this strategy would enable enterprises to offer real-time voice data processing services. Despite the apparent advantages and convenience of using cloud servers to store voice data, the privacy and security of these stored voice features will continue to be enhanced, given that they are directly related to the private information of individual users. In 2019, voice messages of Facebook users were transcribed and stored without user consent. These messages contain sensitive information about users and private information such as users' identities, addresses, and finances and can be easily compromised, leading to various potential security risks. This serious privacy threat has attracted the attention of government, industry, and academia. The introduction of security and privacy compliance obligations such as the EU's General Data Protection Regulation (GDPR) [3] and the California Consumer Privacy Act (CCPA) [4] has given users greater autonomy over their data, including voice recordings. As a result, the Company is expected to enhance privacy protection and control over its critical information.

The primary potential privacy threat associated with the storage of voice data in cloud-based systems pertains to the possibility of malicious attacks initiated by the cloud itself, given the typically semi-trusted nature of cloud servers. Common attacks include eavesdropping [5] and voice reply [6], which can lead to user privacy breaches and identity fraud. Furthermore, the potential exists for stealing sensitive user data using attribute inference attacks [7]. Privacy concerns are becoming increasingly pronounced, with attribute inference attacks representing a particular area of concern. Consequently, enhancing the security of cloud-based voice data storage is imperative. To address these concerns, enterprises can employ various encryption methods, such as the AES encryption method [8], [9], to enhance voice privacy protection. This approach offers a theoretical guarantee of security and privacy for speech. However, overly reliance on a pre-set key renders the system vulnerable to compromise, resulting in the inability to reconstruct original voice data in the event of key loss or damage. In the event of the voices above being encrypted and situated directly in the cloud, it will be easy to detect the state of the cipher information in it and then feel that it is some classified information and thus directly carries out malicious damage, which will cause a lot of losses. Therefore, it is necessary to use robust methods for this potential interference phenomenon. In this regard, a combination of secret sharing [10] and audio steganography can be used, enhancing the security and privacy of speech very well. Specifically, steganography that embeds all processed secret share shares in random audio can hide these secret shares and make them difficult to detect. Furthermore, even if a small number of shares are lost due to mismanagement or damage to cloud storage, the original voice data can still be reconstructed, provided a suitable secret share recovery mechanism has been implemented.

Given the robust outcomes attained by secret sharing and audio steganography in speech data processing in the face of the aforementioned interference scenarios, this paper proposes a novel scheme integrating these two approaches to enhance the privacy of speech-sensitive features. The proposed scheme involves several methods, including lattice coding, Chinese remainder theorem, secret sharing, and audio steganography, to ensure the security of speech-sensitive features. The proposed method integrates all secret shares obtained by randomly

This paper was supported by the National Natural Science Foundation of China under grant 62402111. And was supported by the NSFC-FDCT under its Joint Scientific Research Project Fund (Grant No. 0051/2022/AFJ), China & Macau.

Manuscript received April 19, 2021; revised August 16, 2021.

processing speech-sensitive features into the frequency domain of random carrier audio. This approach ensures that the impact of these shares on the audio is minimized, making them difficult to detect. Consequently, the secret shares are rendered highly covert. Once embedded in the audio, shares are stored and managed separately, and they are no longer all saved in the cloud. The original speech features are only reconstructed once a specified share recovery threshold has been attained. This approach ensures that the effects of attribute inference attacks are circumvented, and even if an attacker were to steal some of the shares at the storage location, it would be secured by the inability to analyze or reconstruct the original speech effectively. The effectiveness of the proposed method is empirically evaluated through experimental validation on four different datasets. The experimental results confirm the method's effectiveness in resisting attribute inference attacks and protecting the privacy and security of sensitive speech features. The highlights of our original contributions are summarized below:

- **Enhanced Privacy Protection**: The proposed method can effectively improve the privacy and security of sensitive part-of-speech data and can effectively resist attribute inference attacks.
- **Robustness against malicious interference**: The solution demonstrates robust and reliable performance during operation, even during unexpected malicious interference or damage. This is achieved by improving random secret sharing applied to speech and using appropriate recovery thresholds.
- **Preservation of Voice Usability**: The proposed enhancement of privacy and speech integrity is evaluated by executing simulation experiments. The objective is to ensure the integrity of speech data while enhancing user privacy.

II. Related Work

In the domain of privacy preservation in speech, there are already established methodologies, including traditional classical encryption techniques, such as homomorphic encryption methods [11], [12] , which can be applied to speech well to encrypt some sensitive information. Still, there are some problems if the encrypted information is uploaded to the upstream server; due to the untrustworthiness of the upstream server, it is possible to infer the corresponding sensitive data due to curiosity. May infer the corresponding sensitive data due to curiosity; meanwhile, the same problem exists in joint learning [13] in the speech domain, but recently, there has been privacy-enhanced joint learning against attribute inference attack, which can achieve the effect of resisting curiosity inference attack to a certain extent [14]. Still, this method is only for single sensitive information. The computational overhead and privacy protection effect are unsuitable for simultaneous sensitive details in multiple speech. Overhead and privacy protection are not reasonable. Some recent encryption methods, such as AES encryption [8], [9], can provide good privacy but often tend to rely too much on the key when used. Meanwhile, some methods directly add noisy, distorted data to conduct inference attacks on sensitive information, such as the differential privacy method, which can ensure personal privacy well. Still, due to the addition of noisy, distorted data, the high precision information will be significantly affected, such as voiceprint data, leading to a significant decrease in the accuracy of voiceprint authentication [15]–[18]. Compared with these methods, the designed scheme can better protect the availability of data and information while resisting attribute inference attacks.

It is also possible to protect voice privacy by detecting audio noise interference, i.e., using sound sensors to interfere with ambient sound. The INFOMASKER system was developed for voice eavesdropping [19]. The system utilizes acoustic sensors within the environment to inject sound into the climate based on the noise of the mobile phone. The analysis of voice signals captured by a wiretap is rendered unfeasible without interfering with daily life and thus cannot protect an individual's voice privacy. However, the acoustic sensors in this approach possess a constrained operational range, ensuring privacy within a delineated area. Once this range is exceeded, the efficacy of the system diminishes.

Therefore, with development and progress, there are machine learning methods that protect privacy; it is possible to add scrambling to speech, even if a malicious user uses it to synthesize speech, the maliciously synthesized voice sounds no longer the original voice [20] due to the existence of the perturbation to achieve the effect. The complete voice leakage at the hardware microphone level [21] is also a method of countering perturbation. In the case of the malicious collection of microphones, voice conversion will be carried out using maliciously collected signals to achieve the effect of privacy protection. Unlike the protection of complete speech, this study focuses on the privacy of sensitive information in the voice. For the safety of some sensitive information in the voice, there are some machine learning anonymization methods, and anonymization suppresses the individual recognizable elements in the voice signal while retaining other aspects, mainly the user's voice data, which is processed through the anonymization system. The anonymous system hides the speaker's sensitive attributes, does not affect other non-sensitive attributes, and ensures that the anonymous speaker's voice sounds like the same hypothetical speaker spoke it. For the emotional data in speech [22], [23], it is proposed to set up a privacy middle layer between the user and the cloud service to purify the voice input. The cyclic generative adversarial network converts the original speech input into a speech signal without emotion to realize the stylistic transfer of speech emotion and protect the user's speech emotion data. For the identity authentication of voiceprint data in speech, a V-CLOAK real-time voice anonymization system was proposed to ensure the clarity, naturalness, and sound quality of speech conversion [24]. FAPG systems also implement speech conversion for specific goals to protect privacy [25]. However, these methods can only protect sensitive data in a single voice, and it may not work well if you need to protect multiple sensitive attributes simultaneously. Furthermore, a privacy-preserving framework in a crowdsourcing environment has been proposed, which employs a hybrid learning approach

to train the feature extractor and can effectively safeguard multiple sensitive attributes in the data [26]. The proposed approach has been validated in terms of both its privacy and utility on both image and text datasets. However, it should be noted that the attributes between images and text are relatively more independent than those in speech data and that the correlation between various attributes in speech data is not the same [7]. Consequently, the approach is not suitable for the field of speech privacy protection. Another approach for speech is the recently proposed voice fence wall [27], which separates sensitive and non-sensitive speech in speech and uploads non-sensitive speech information. Still, this approach only uploads some non-sensitive information. Conversely, our methodology entails the direct enhancement of privacy safeguards for sensitive elements of speech data. And the present study employs a clandestine steganography scheme grounded in secret sharing. In contradistinction to extant classical speech steganography methods [28], [29] that aspire to safeguard speech security, the pivotal characteristic of this research method is its integration of secret sharing. This mechanism can more efficaciously resist attribute inference attacks and concomitantly enhance speech privacy security.

III. Prieminary

This section reviews some of the main methods used for knowledge and the techniques used to protect privacy and security.

A. Privacy Security Protection Artifact

To start the review, it is necessary to define lattice coding first. Lattice coding maps raw data features into a discrete space using a lattice structure. The mapping process may yield multiple candidate outputs rather than a single result. In lattice coding, the data is not simply encoded as a specific value; instead, it is represented by selecting a suitable point in a high-dimensional space with some inherent coding redundancy. The presence of redundant information renders the relationships between data more complex and ambiguous. The encoding process can be described by finding the lattice point X that is closest to the feature vector of the original data as follows:

$$y = \arg\min_{z \in L} \|X - z\|$$

$\|X - z\|$ represents the encoding error between the original data X and the lattice point z. y is the encoded data point. The encoded speech data has a certain amount of redundancy, making it more difficult for an attacker to decode it.

Next, the Chinese Remainder Theorem (CRT) is a mathematical theorem that solves the problem of satisfying a system of congruent equations under multiple moduli [30]. A unique solution can satisfy all congruence conditions for mutually prime moduli. In addition to enhancing the system's security, this approach effectively alleviates the overdependence on participants. T is the participant threshold required to recover the secret S. If the number of participants is greater than or equal to T, the secret can be recovered as follows:

$$S = \sum_{i \in D \text{ where } D \geq T} r_i M_i y_i \pmod{M}$$

By combining linear secret sharing and the Chinese Remainder Theorem, suppose the secret S distributed to N participants through linear secret sharing is represented using the Chinese Remainder Theorem as follows:

$$S = \sum_{i=1}^{N} s_i M_i y_i \pmod{M}$$

where s_i is the secret part generated by participant i, M_i is the computed quotient, and y_i is the modular inverse of M_i. Each secret part s_i is shared based on different modulus congruences, enhancing the system's security.

Audio steganography can be defined as the process of embedding secret information in the frequency domain of a speech signal without being easily detected. The most suitable location for embedding the secret information for steganographic embedding is identified by combining a convolutional neural network (CNN) to balance the quality of the embedded secret information and the quality of the carrier audio after embedding. Assume that the input of the CNN model is the frequency domain signal $\mathbf{X}(t, f)$, where $\mathbf{X} \in \mathbb{R}^{T \times F}$, and T represents the number of discrete points in the time dimension, and F represents the number of discrete points in the frequency dimension. Each element $\mathbf{X}(t, f)$ represents the signal value at time t and frequency f. It will output a weight matrix W to indicate which positions are most suitable for embedding secret information as follows:

$$W = \text{CNN}(\mathbf{X}(t, f)) \qquad W \in \mathbb{R}^{T \times F}$$

Based on the weight matrix W, the best time-frequency points (t_i, f_i) are selected for embedding secret information. These selected points (t_i, f_i) can be used to adjust the signal $\mathbf{X}(t, f)$ as follows:

$$\mathbf{X}(t_i, f_i) \rightarrow \mathbf{X}(t_i, f_i) + \Delta\mathbf{X}(t_i, f_i, s)$$

$\Delta\mathbf{X}(t_i, f_i, s)$ is the increment added to embed the secret bit s, which is usually a slight adjustment to the phase.

IV. Problem Formulation

Suppose the voice intelligence service provider enterprise collects the voice data features of the users to be stored in the cloud. In this case, the semi-trusted nature of the cloud may lead to malicious speculation by the cloud about the sensitive attributes involved in these speech features, which may result in user privacy leakage.

A. Threat Model

We consider the adversary a third party's semi-trusted cloud, which destroys a portion by analyzing its content or maliciously corrupting it. The adversary may:

1) Speech-sensitive features stored on cloud servers are vulnerable to semi-trusted cloud attribute inference attacks, which can infer sensitive parts of the message design
2) The unauthorized recording or distribution of voice messages transmitted by the user is prohibited;

Once one of the two methods described above has been accomplished, an adversary can extract sensitive information

from it, which may result in the disclosure of sensitive information. The present paper focuses on enhancing the privacy of sensitive part-of-speech information stored in the cloud.

B. The workflow for Protecting Voice Privacy

This delineates the privacy-preserving workflow in our scheme. To further enhance the privacy of the user's voice, a steganography-based linear secret-sharing scheme has been designed to address this issue, as illustrated in Figure 1. The scheme comprises three phases to maximize the privacy of speech-sensitive features. This ensures that the privacy of the user's speech is secure in the event of a malicious attack. The scheme also balances voice privacy and usability, ensuring that speech-sensitive features f can be fully recovered.

The workflow of the proposed program entails the initial processing of speech data. The proposed scheme emphasizes enhancing the privacy of speech-sensitive attribute features. This is achieved by performing a series of processing operations on sensitive voice feature f. These operations include feature extraction and data quantification. The purpose of this phase is to prepare the preprocessing for the subsequent privacy enhancement operations. Then, a lattice coding operation is performed to map the processed sensitive feature vectors to the high-dimensional lattice vectors x, thus completing the lattice coding process. To further enhance privacy, noise vectors are also crafted and added to the encoded lattice vectors, and the next privacy enhancement operation is performed for this lattice vector x'.

Preprocessed sensitive speech lattice vectors are encoded and processed as target s to apply a secret sharing scheme. Following a meticulously designed linear secret-sharing operation, the encoded vectors are divided into multiple shares with carefully configured recovery thresholds. This approach enhances speech privacy while ensuring accurate reconstruction of the original speech. Subsequently, these secret sharing shares are embedded in carrier audio for secret and secure transmission. Following the steganography process, the resulting audio files (stego_1, stego_2, ..., stego_n) are not stored solely on a semi-trusted cloud. Conversely, a distinct storage method is employed, whereby a proportion of shares are retained on a local device. These shared shares can be retrieved from their respective storage locations and merged to reconstruct the original data for further use.

To demonstrate the usability of the data recovered from the privacy-preserving scheme, the process is initiated by retrieving audio files from their respective storage locations to reconstruct the speech. The steganography recovery procedure initially extracts the embedded secret information from the steganographic audio. Subsequently, a linear secret-sharing recovery mechanism is employed. Utilizing a predefined recovery threshold, designated as t, collecting a sufficient number of secret shares yields a unique solution to the linear secret sharing equation, thereby reconstructing the original high-dimensional lattice-encoded vector, denoted as x'. Finally, a recovery operation is performed to recover the original speech feature vector.

V. Proposed Method

The proposed method is principally employed to secure the voice-sensitive attributes of users stored in the cloud server, primarily through voice steganography and linear secret sharing, to enhance the privacy of voice-sensitive attributes.

A. Preprocess Module

In our framework, the voice data processing module is the first step in privacy processing. In this module, we need to accurately extract the sensitive information features in the user's voice and perform data preprocessing and privacy enhancement operations to facilitate the next processing step.

This paper focuses on processing sensitive attribute features in speech data used for intelligent speech services. These features are processed based on the sensitive attributes selected by the user, with the emotional and voiceprint attributes being given primary consideration. The frequency representation of the speech data over a specific time period is first obtained using the short-time Fourier transform(STFT). The feature extractor for the corresponding attributes is trained to perform the classification task using the Table I datasets, and a cross-entropy loss function is invoked to measure the classifier's performance [27], thus enabling better extraction of speech features involving sensitive attributes. The amalgamation of these features into feature vectors is instrumental in enhancing the representation of sensitive attribute features in speech.

These feature vectors are represented as sensitive attribute features for the next step in the process. The sensitive features are then subjected to quantization operations to transform the feature vectors f using non-uniform quantization techniques into discrete integer vectors f_q. If multiple data features are extracted, fusion and truncation operations are performed to align the features for subsequent processing.

Then, privacy enhancement operations are performed on these sensitive feature data to counter attribute inference attacks. First, a suitable high-dimensional lattice dimension k is selected for these processed vectors to generate k independent basis vectors $\{b_1, b_2, \ldots, b_k\}$. These vectors define the lattice structure to form lattice points, as shown in Equation (3). The lattice points composed of the linear combination of these basis vectors map the quantized speech feature vector to the nearest lattice point code x found by the neighborhood search algorithm to complete the lattice point coding operation. To complete the lattice point coding operation, the preprocessed sensitive feature vector, designated as f_q, is mapped to the lattice point code, represented by x. This mapping process can be expressed as $f_q \rightarrow x$.

The encoded high-dimensional lattice point vector has a complex data point distribution, which may show coding redundancy in the encoding process. Some feature combinations are repeated in different data points, resulting in the same code. Then, additional redundant information r is embedded in the voice point after lattice coding, as shown in Equation (1). The introduction of coded redundant information facilitates the retention of sufficient data in the event of partial data loss or corruption. This excess information can then be utilized in the decoding process for interpolation and extrapolation, thereby enhancing recovery accuracy.

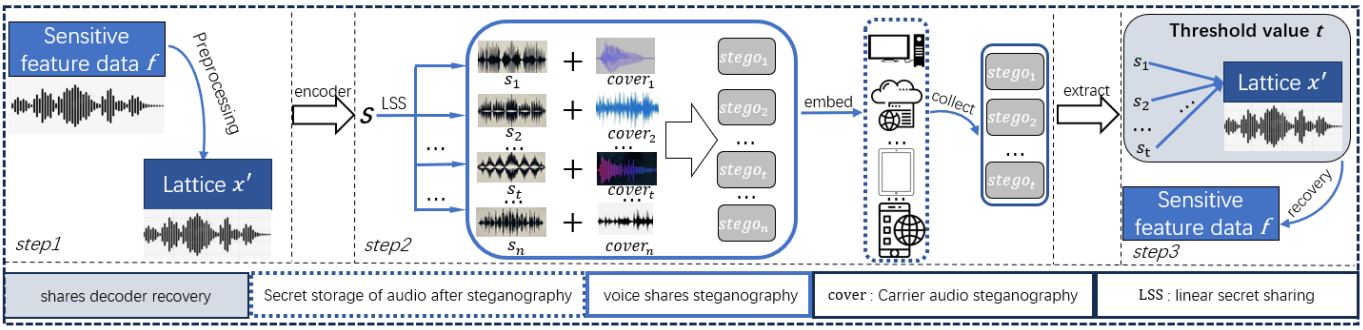

Fig. 1. Workflow for enhanced voice sensitive attribute feature privacy Protection

Next, select an appropriate noise vector $n \sim \mathcal{N}(0, \sigma^2)$ and add it to the encoded lattice vector x shown in Equation (2). This makes the lattice vector more challenging to analyze and infer, thereby protecting the original sensitive features from direct access.

$$x_1 = x + r \quad (1)$$

$$x' = x + n \quad (2)$$

$$s_x = \left\{ \sum_{i=1}^{k} x_i b_i \mid x_i \in \mathbb{Z}, b_i \in \mathbb{R}^k \right\} \quad (3)$$

B. Privacy Enhancement Module

Implementing a secret-sharing-based speech steganography scheme is primarily intended to enhance privacy security.

The preprocessed sensitive speech coding vector x is initially encoded and processed as the target s for applying the secret sharing scheme. Combined with the Chinese Residual Theorem (CRT), the encoded vector x after data preprocessing will be partitioned into multiple shared segments, also called shared shares. Expressly, a set of random seeds is first set to generate a set of random numbers from random seeds $\{m_1, m_2, \ldots, m_k\}$ ensure $k > t$. These numbers should fulfill the following conditions:

1) Ensure that a set of numbers is randomly generated.
2) This set of numbers should have no relationship with each other and are mutually prime. That is, for any m_i and m_j, it is necessary to satisfy $\gcd(m_i, m_j) = 1$ to ensure that the modulus composed of these numbers has a unique solution.

The set of numbers is treated as modulus, while the vector x' is treated for segmentation coding according to Equation (2) and as a shared segment S. This is calculated according to the different moduli and finally partitioned into n segments to obtain n shared shares $\{S_1, S_2, \ldots, S_n\}$ as follows:

$$P(x) = S + a_1 x + a_2 x_2^2 + \cdots + a_{t-1} x^{t-1}$$

Share$_i$ is secret sharing gets a secret share which we denote as S_i and can be expressed as $S_i = (x_i, s_i)$, s_i is the value of each Share is calculated as follows:

$$s_i = \{P(x_i) \mod m_1, \ldots, P(x_i) \mod m_k\}$$

A threshold value t must be set, which is $t \geq 0.8n$, to ensure that a certain number of secret shares must be collected to recover the original secret. Even if the attacker maliciously obtains a small portion of the shares, restoring the secret information is impossible. It is challenging to infer the original secret information from the partial information obtained because the random moduli are not correlated with the fact that there are less than t shared segments, the original coded information cannot be obtained because the unique solution cannot be obtained, effectively preventing data leakage and also dramatically increasing the difficulty of the attribute inference attack. After ensuring that at least t shared fragments are collected, the original coded information can be completely recovered. The underlying principle of this scheme is rooted in the concept of secret sharing, a method of information security predicated on the theoretical principles of information theory [31]. The objective of secret sharing is to guarantee the absolute security of the information in question.

Upon completing the secret-sharing process, the next step is to store the secret shares securely and covertly. A covert transmission process is employed, where the secret shares are transmitted to designated storage locations. This is accomplished through frequency-domain steganography, where the secret shares are first embedded into audio carriers before transmission. Specifically, given the set of n secret shares $\{S_1, S_2, \ldots, S_i, \ldots, S_n\}$, n audio segments are randomly selected as steganographic carriers $\{\text{cover}'_1, \text{cover}'_2, \ldots, \text{cover}'_i, \ldots, \text{cover}'_n\}$. These carriers are processed using frequency-domain steganography, where each segment undergoes Short-Time Fourier Transform (STFT) as Equation (4) to convert the time-domain representation of audio into a form that shows the frequency at a certain point in time, i.e., a time-frequency representation $\{\text{cover}_1, \text{cover}_2, \ldots, \text{cover}_i, \ldots, \text{cover}_n\}$. Applied to the speech data among the secret share splits, the encoded vector is split into multiple shares, and a reasonable threshold is designed to balance privacy protection with data recovery availability. Select the carrier audio cover using Short Time Fourier Transform (STFT): $x(t)$: input the audio signal, $w(t)$: window function used to intercept part of the signal, m: time offset, ω: frequency variable, to decompose the carrier audio to select the high-frequency part of the carrier audio as the carrier for embedding the secret information, and embed the

secret information generated by the secret Share as the secret in the carrier audio.

$$\{x(t)\}(m,\omega) = \int_{-\infty}^{\infty} x(t)w(t-m)e^{-j\omega t}\,dt \quad (4)$$

A trained convolutional neural network (CNN) is employed to determine the optimal hidden embedding locations within the coverage of the high-frequency region of the carrier audio, thus facilitating the execution of the secret implicit embedding operation. The selected coefficients of the high-frequency portion of the STFT are extracted from the frequency domain representation of the STFT. Subsequently, bit flipping is employed to modify the value of the lowest valid bit of the frequency coefficients, which is minimally weighted. It has been demonstrated that modification of this bit hardly affects the overall magnitude of the value. The overall magnitude of the value enables the hiding of information. It ensures that the embedded information is imperceptible to the human ear (i.e., the carrier frequency domain representation) without significantly altering the signal, thereby preserving data privacy. The secret information sharing, denoted by s_i, and the embedding function, f_{embed}, are used to generate the steganographic audio, which contains the secret Share. This is illustrated as follows:

$$\text{stego } i(m,\omega) = \text{cover } i(m,\omega) + f_{\text{embed}}(s_i, \text{cover } i)$$

Each secret Share is embedded in a specific high-frequency location of the carrier audio Fourier coefficients. This strategic placement ensures that the embedded information cannot be detected during transmission, thus effectively enabling covert communication. After embedding the covert shared information, the carrier audio is reconstructed into normal time domain format using Inverse Short Time Fourier Transform (ISTFT). Recover the time-domain audio from the stego frequency domain representation stego $i(m,\omega)$ as follows:

$$\{\text{stego } i(m,\omega)\}(t) = \int_{-\infty}^{\infty} \text{stego } i(m,\omega)e^{j\omega t}w(t-m)\,d\omega$$

The resulting audio contains embedded steganographic information, and the transmission of this audio is less likely to detect the secret information, thus achieving secure and covert transmission. This steganographic transmission method dramatically enhances the privacy of the secret sharing program, ensuring that sensitive data is transmitted securely, covertly, and undetectably. Instead of storing all of the data in a semi-trusted cloud, a portion of the data is stored in local secure storage devices such as local computers and cell phones, reducing the likelihood of an attack being realized.

C. Voice Reconstruction and Recovery

The original usable voice data can still be obtained to reflect the availability of data from the privacy protection program, and the final recovery of voice data is also an essential step in this program. The process commences with steganography recovery, extracting the secret information from the steganographic audio. Secret sharing recovery is then employed, Combining Lagrange interpolation and the Chinese remainder theorem to reconstruct the original high-dimensional lattice coding vectors from the extracted secret information data according to the set recovery threshold by t. The final stage is lattice coding recovery to recover the original speech feature vector.

To reflect the availability of voice data and continue the efficient use of innovative voice services, obtaining the original data from the privacy-preserving scheme and verifying its availability is necessary. This process starts with collecting steganographic audio from different storage locations first to check if it is usable, followed by selecting the usable steganographic audio starting from steganography recovery and extracting the secret information embedded in the previous step, i.e., extracting the embedded shared shares from the steganographic audio. These secret shares are used to solve the unique solution of the equation to recover the original feature information, collect at least t shared segments, and recover encoded high-dimensional vectors x' using the Chinese remainder theorem. The recovery process includes:

- (1) For each modulus, the collected shares$\{S_1, S_2, \ldots, S_t\}$ were utilized and computed using Lagrange interpolation, as expressed below:

$$P(x) \mod m_i = \sum_{j=1}^{t} s_j[i] \cdot \prod_{k=1}^{t} \frac{x_j - x_k}{x - x_k} \mod m_i$$

- (2) The polynomial value $p(x)$ can be recovered by combining the Chinese Remainder Theorem (CRT) as follows:

$$P(x) = \sum_{i=1}^{k} (P(x) \cdot w \mod m_i) \mod M$$

$$w = m_i \cdot M_i \cdot M_i^{-1}$$

Each segment $s_j[i]$ represents the value of the polynomial $P(x)$ at a specific point x_j is expressed as for $s_j[i] = P(x_j)$.The product of all moduli, denoted by M is given by $M = \prod_{i=1}^{k} m_i$. The result of dividing the total modulus M by the current modulus m_i, denoted as M_i satisfies the congruence condition:

$$M_i \times M_i^{-1} \equiv 1 \pmod{m_i}$$

Finally, the polynomial $P(x)$ is successfully reconstructed, and its constant term, corresponding to $P(x)$, yields the secret S

Subsequently, a series of processes are conducted, including noise elimination decoding, recovery, and verification. In the noise decoding stage, The variance of the preceding additive noise is σ^2. Wiener filtering, utilizing a linear filter, enables noise removal by minimizing the mean square error [32]. The redundancy introduced during the lattice encoding stage ensures that even if some shares are compromised due to noise, the remaining redundant shares provide sufficient information for recovery. This enhances the overall data usability and significantly improves the accuracy of data recovery.

Following the decoding and recovery of the noise, the lattice-coded vectors can be decoded and recovered from the discrete lattice space. This is achieved by combining redundant

information with an inverse transform of the mesh coding operation. The result is the original feature vector before mapping, designated as f_q. To obtain the original feature vector, it is necessary to map the quantized feature values back to the original values as far as is feasible. Following acquiring the original features, must be restored to the initial time-domain speech signal utilizing an inverse STFT transform. Successive iterations of the Griffin-Lim algorithm are then used to recover the forfeited phase information during the privacy-enhancing processing phase [33]. The iterations are optimized to determine the best phase information for the speech signal to efficiently reconstruct the time-domain signal and ensure the quality and naturalness of the speech. Experimental validation is necessary for applications to determine the difference between the final recovered speech feature and the initial speech.

VI. Evaluation

This section comprehensively conducts experiments to evaluate our voice steganography based on linear secret-sharing performance. We first introduce the Experiment setup, dataset, performance metrics, and experimental evolution. Then, we report and analyze the experimental results from various perspectives.

A. Experimental Setup

We evaluate the impact of linear steganography sharing, particularly the proposed framework, on model accuracy. In our experiments, we trained and evaluated the steganographic models using the TensorFlow 2.10 framework on a server equipped with an NVIDIA RTX 4060 GPU. Parameter Configuration: The batch sizes used for the speech emotion extraction and voiceprint feature extraction modules are 16 and 64, respectively. Both modules use an Adam optimizer with an adaptive learning rate, an initial learning rate of 1e-6, a sampling rate of 16KHZ, and a quantization level 1024 for the lattice encoding dimensions of 128. A standard deviation of the added Gaussian noise of 0.001σ and a secret share of 10 for the secret Share. The Share of secret sharing is set to 10, and the corresponding threshold t to be recovered is 8, and the number of Griffin-Lim iterations in the recovery process is set to 500. After setting up the parameters, the experiment is started.

B. Dataset

We present a list of the five datasets considered for use:

RAVDESS: The dataset comprises 24 actors uttering a sentence in a North American accent while displaying a range of emotions. We utilize the speech data from this dataset, labeled with five emotion categories (angry, happy, sad, fearful, and surprised), to extract features to validate the privacy associated with multiple sensitive features.

LibriSpeech: is a large open-source English speech recognition dataset consisting of about 1,000 hours of English speech read aloud from audiobooks from the LibriVox project, carefully segmented and aligned, with about 2,000 speakers participating that can be used for voiceprint privacy testing

IEMOCAP: The IEMOCAP dataset comprises 12 hours of audiovisual data from 10 actors. It encompasses both scripted and improvised dialogue between male and female actors in the English language. The data were segmented by speaker turn, resulting in 5,255 scripted and 4,784 improvised recordings. These recordings contribute substantially to the study of multimodal emotion recognition tasks.

VoxCeleb1: The VoxCeleb1 dataset is a widely used resource for speaker verification. It comprises over 100,000 quotes from more than 1,251 celebrities extracted from YouTube videos, with audio from multiple speakers in various acoustic environments, including outdoor stadiums and red carpets. The dataset is relatively balanced concerning gender and is also employed in voiceprint privacy testing tasks.

GTZAN: The GTZAN dataset contains 1000 30-second-long audio tracks divided into 10 genres of 100 tracks each. This is a standard dataset for music genre classification and music information retrieval, and the dataset was used as the carrier audio in the steganography experiments.

The initial four datasets were used as the confidential data set for evaluating privacy and eventual recovery. In contrast, the final music dataset was utilized as the carrier audio for the steganography process as the confidential dataset. Table I delineates the distinctive attributes of each dataset. For each dataset, partitioning was implemented, whereby 80% of the dataset was utilized for training and 20% was allocated for testing.

TABLE I
DIFFERENCE VOICE DATASET

Dataset	Emotion	Users	Gender
Voxceleb1	-	1000	Male,Female
LibriSpeech	-	1000	Male,Female
REVADESS	5	24	Male,Famale
IEMOCAP	4	10	Male,Famale

C. Metrics

Before the experimental evaluation, it is necessary to introduce the relevant terminology and evaluation metrics. In the initial assessment phase, the system's ability to resist privacy leakage is assessed by implementing a series of attacks that potentially leak the user's privacy. The most critical of these is the ability to withstand membership inference attacks, and the accuracy of attack algorithms (ACC) is used to assess the privacy situation under a range of attack algorithms. In addition, the final recovered data will be evaluated using Associated Task Accuracy (ACC) and Equivalent Error Rate (EER) as metrics to assess the efficacy of speech emotion recognition and voiceprint verification.

In a binary classification problem, EER denotes the error rate obtained by the classifier with equal true and false negative error rates. It is a metric used to test the performance of speech data, the smaller the value, the lower the error rate and the stronger the model performance.

The Mean Square Error (MSE) is a metric employed in the domain of privacy security to evaluate the impact of steganography and privacy-preserving schemes on the quality of speech data. The MSE is determined by comparing the

original data with the speech data processed by the privacy schemes. A smaller MSE thus indicates a closer match to the original speech after processing. MSE quantifies the impact of privacy protection on data availability. Let N be the total number of samples, y_i be the actual value, and $\hat{y}_i$ be the predicted value. The formula is:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Several data points can assist in verifying the effectiveness of the recovered data. Signal-to-Noise Ratio (SNR) is used to measure the signal quality of the recovered audio relative to the noise. x_i are the original audio samples, and $\hat{x_i}$ is the sample value of the restored audio, The formula is:

$$\text{SNR} = 10 \cdot \log_{10}\left(\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{N} x_i^2}\right)$$

Meanwhile, Short-Time Objective Intelligibility (STOI) is used to evaluate the intelligibility of the recovered audio, especially in speech-related tasks, to assess the effect of the final reconstruction of the recovered obtained speech, where T is the number of time frames, F is the number of frequency components, $x_{t,f}$ and $hatx_{t,f}$ are the values of the original and recovered audio at the t moment and f frequency component, The formula is:

$$\text{STOI} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\sum_{f=1}^{F} x_{t,f}^2}{\sum_{f=1}^{F}\hat{x}_{t,f}^2} \cdot \sum_{f=1}^{F}(x_{t,f} \cdot \hat{x}_{t,f})\right)$$

D. Evaluation

We will first verify that our program protects users' sensitive attributes while maintaining good data utility. The scheme's privacy security is verified by mimicking the effect under the influence of just the environment and then several attribute inference attacks. The usability of the final recovered data is obtained by the effect of the final reply data and the signal-to-noise ratio of the data.

To demonstrate the adequacy of our experiments, we focus on the emotional and identity attributes contained in the speech data in our experiments. Therefore, we simulate the following two types of users in use with our scheme for privacy enhancement.

- Bob: Bob doesn't want to reveal emotional attributes. The choice is to enhance privacy protection of emotional attribute features.
- Alice: Alice does not want to reveal identity attributes. The choice is to enhance privacy protection of voiceprint attribute features.

Table I lists several types of speech data and their characteristics. Our scheme considers privacy and utility and then performs an analysis based on these two features. We evaluated the utility and privacy-preserving capability of speech fencing using RAVDESS, IEMOCAP, LibriSpeech, and VoxCeleb1 datasets.

The following two user selection scenarios will be simulated in isolation, and the privacy of the proposed scheme will be evaluated. To this end, the study will simulate half-trusted upstream servers that may exist in an actual physical network environment. Attack scenarios involving the inference of additional attributes are then performed using four commonly utilized algorithms on data necessitating privacy protection. The ACC(%) of the aforementioned attacking algorithms is then used to evaluate the privacy of our scheme's secret-sharing-based speech steganography scheme.

- Support vector machine (SVM): SVM classifiers find a hyperplane in N-dimensional space (N: number of features) to accurately classify data points. Use the Radial Basis Function (RBF) to simulate an attribute inference attack on the processed data.
- Random forest (RF): The RF algorithm is an integrated learning algorithm that improves the stability and accuracy of the model by combining multiple decision trees. In our experiments, we set the number of decision trees to 200. For simulation to perform attribute inference.
- Statistical Analysis-Based Attacks (SABA): Hidden attributes can be inferred indirectly by analyzing statistical correlations and distribution patterns between attributes in a dataset. It is a simulated attribute inference attack.
- Reconstruction Attacks (RA): Reconstruction attacks can use known parts of the information, dependencies between attributes, and parts of the data to infer the sensitive characteristics of the data, which is also an attribute inference attack to simulate a semi-trusted cloud server.

Two standard machine learning inference attack algorithms were first simulated, as demonstrated in Table II and Table III, as soon as Figure 2 and Figure 3. Specifically, two user-selected requirements were simulated, and then an attack on sensitive attribute features in speech data was inferred by comparing them with several selected SOTA methods. Specifically, wav2rac [34] is utilized as the baseline for experimental comparison in the speech domain, while Privacy Enhanced Federated Learning (PE-FL) [14] and speech data processed by our scenario are employed for attribute inference attack validation. Next, we conducted experimental tests on Bob and Alice users by simulating two other attacks on wav2vec and voice fence wall [27] and on our scheme. Subsequently, the needs of Bob's users are simulated to perform sensitive attribute inference on the data processed by wav2rac and Purifier [35] as well as the approach under investigation. As can be seen from the table and graphical results, the data processed with the proposed scheme effectively reduces the accuracy of sensitive attribute inference. This result suggests that the protection of users' voice attribute features has been enhanced. Furthermore, this outcome corroborates the finding that the secret-sharing-based speech steganography scheme demonstrates superior performance in terms of privacy preservation.

The steganography module was evaluated after secret sharing to facilitate a more comprehensive assessment of the privacy performance of the modules examined in this study. The primary objective of the evaluation was to ascertain the steganography module's efficacy in facilitating covert trans-

TABLE II
EMOTION FEATURE INFERRING ATTACKS USING DIFFERENT ACOUSTIC SYSTEMS FOR EXTRACTING PROPERTIES OF REPRESENTATIONS ACC (%)

	wav2vec		PE-FL		ours	
	REVADESS	IEMOCAP	REVAEDSS	IEMOCAP	REVADESS	IEMOCAP
SVM	77.3	75.7	48.6	45.5	44	42
RF	72.5	72.2	46.7	44.2	37.6	38.9

TABLE III
VOICEPRINT FEATURE INFERRING ATTACKS USING DIFFERENT ACOUSTIC SYSTEMS FOR EXTRACTING PROPERTIES OF REPRESENTATIONS ACC (%)

	wav2vec		Purifer		ours	
	VoxCeleb1	LibriSpeech	VoxCeleb1	LibriSpeech	VoxCeleb1	LibriSpeech
SVM	73.2	71.4	45.2	43.3	43	41.8
RF	69.8	68.6	44.2	43.4	40.5	39.6

TABLE IV
PERFORMANCE METRICS OF STEGANOGRAPHIC MODULES UNDER EXTERNAL NOISE INTERFERENCE

noise amplitude	SNR	MSE	Audio hash changes
0	35.3dB	0.0002	unchanged
0.005	31.5dB	0.0014	unchanged
0.01	28.8dB	0.0139	unchanged

TABLE V
OUR PROGRAM OF THE RECOVERY ORIGINAL VOICE DATA

	REVADESS	LibriSpeech	VoxCeleb1	IEMOCAP
MSE	0.00043	0.00042	0.00032	0.00027
SNR	19.8dB	20.5dB	22.3dB	23.6dB
STOI	0.78	0.81	0.83	0.84

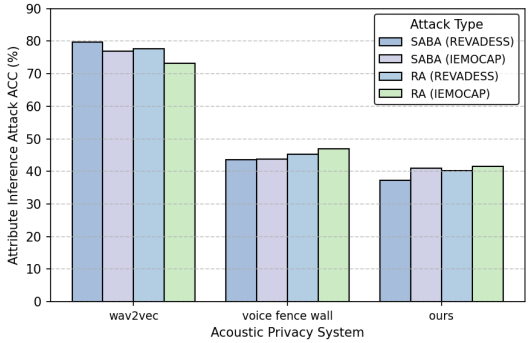

Fig. 2. Voice emotion attributes datasets SABA and RA inference attack ACC%

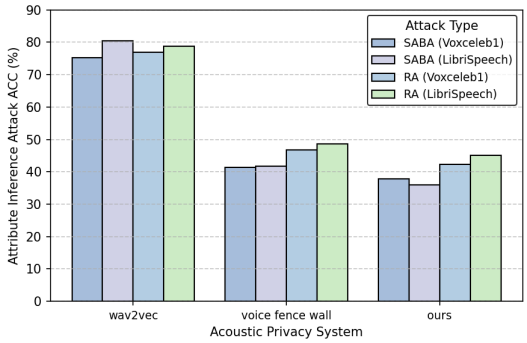

Fig. 3. Voiceprint attributes datasets SABA and RA inference attack ACC%

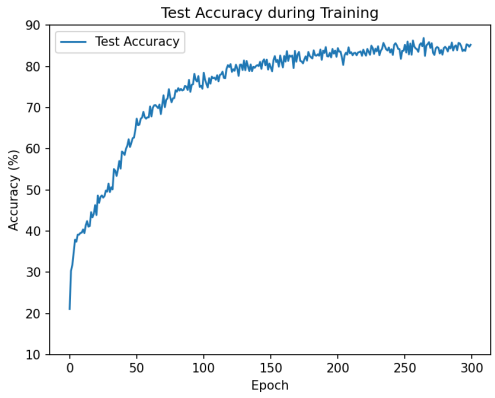

Fig. 4. REVADESS test for recovery voice emotion feature ACC%

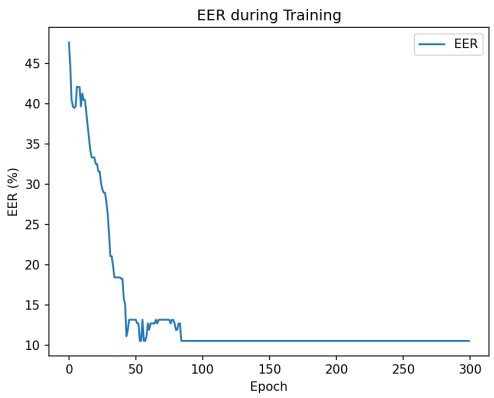

Fig. 5. REVADESS test for recovery voice emotion feature EER%

missions. This is achieved by simulating a man-in-the-middle attack for malicious damage during transmission and testing the change in the audio hash value in the case of added noise interference.

As shown in Table IV, no alteration was observed during simulated attacks on the audio hash value. To achieve superior concealment, the steganography module was simulated, with two additional noise amplitude interferences being introduced under the audio effect. Despite slight noise interference in the external environment, superior audio quality was maintained during the simulation of the attack on the audio hash value. This indicates that the steganography module can function ro-

bustly in the presence of noise interference, thereby facilitating covert transmission and enhancing privacy and security.

To verify this, reconstructing and recovering the original speech feature after privacy protection constitutes an essential element of the proposed solution. The following experimental tests have been conducted on the voice performance of the final data obtained after voice data feature reconstruction and recovery.

As illustrated in Table 5 for the four datasets, the testing effect, superior reconstruction of the original speech data, and normal audibility, there are some slight distortion phenomena. However, a very small MSE can be observed, the original data has been recovered very well, they do not affect the data feature itself. The experiment successfully reconstructed the original speech using only $t = 8$, while maintaining the original speech features even when the remaining $n - t = 2$ shares were damaged. In Figure 5 and Figure 6, a very small EER and can be observed, indicating that the voice data has not been affected by the larger one and can continue to be used.

The experimental voice data demonstrate that our scheme substantially enhances privacy and resilience against specific malicious attacks, including attribute inference attacks. Additionally, the scheme exhibits some robustness against external noise interference. While the final recovery data may not be optimal, they can be fully recovered to retrieve the original speech data.

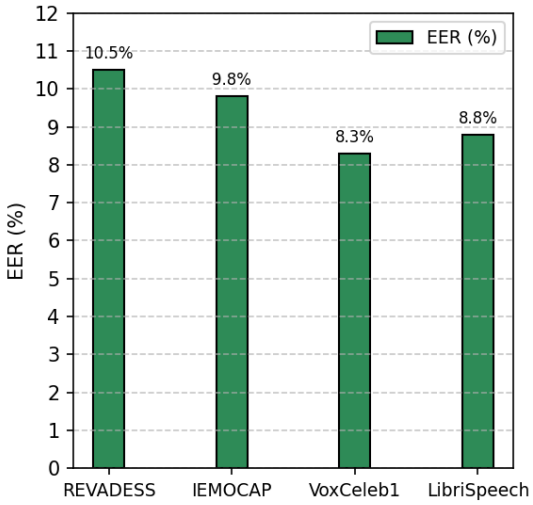

Fig. 6. EER% of all voice recovery data feature

VII. Conclusion

The storage of speech data in the cloud is vulnerable to an attack known as attribute inference, which can result in the leakage of sensitive information due to the semi-trusted nature of the cloud. The proposed solution addresses this vulnerability by enhancing the privacy protection of sensitive information contained within speech data. Specifically, the solution involves implementing separation and privacy enhancement operations on speech data-sensitive features. The resilience of this operation against potential malicious attacks is demonstrated through experimental validation on multiple speech datasets. Voice protection is ensured, and data recovery in normal listening environments is facilitated by better recovery of intact speech data, as no significant differences are detected in such environments. Consequently, users can continue to use similar innovative voice services with confidence, and the eventual recovery of speech data may require optimization of the performance of high-precision intelligent speech data, which will be the focus of our continued research.

References

[1] Shireen Fathi Mallo, Dildar Masood Abdulqader, Rozin Majeed Abdullah, Halbast Rasheed Ismael, Zryan Najat Rashid, and Teba Mohammed Ghazi Sami. A review on feasibility of web technology and cloud computing for sustainable es: Leveraging ai, iot, and security for green operations. *Journal of Information Technology and Informatics*, 3(2), 2024.

[2] Nivedita Singh, Rajkumar Buyya, and Hyoungshick Kim. Securing cloud-based internet of things: challenges and mitigations. *Sensors*, 25(1):79, 2024.

[3] Razvan Viorescu. 2018 reform of eu data protection rules. *European Journal of Law and Public Administration*, 4(2):27–39, 2017.

[4] Rajeev Shrivastava, Mangal Singh, Kalluri Saidatta Subrahmanya Ravi Teja, et al. A real-time implementation for the speech steganography using short-time fourier transformior secured mobile communication. In *Journal of Physics: Conference Series*, volume 2089, page 012066. IOP Publishing, 2021.

[5] Wenbin Huang, Wenjuan Tang, Kuan Zhang, Haojin Zhu, and Yaoxue Zhang. Thwarting unauthorized voice eavesdropping via touch sensing in mobile systems. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 31–40. IEEE, 2022.

[6] Jincheng Zhou, Tao Hai, Dayang NA Jawawi, Dan Wang, Ebuka Ibeke, and Cresantus Biamba. Voice spoofing countermeasure for voice replay attacks using deep learning. *Journal of Cloud Computing*, 11(1):51, 2022.

[7] Ranya Aloufi, Hamed Haddadi, and David Boyle. Privacy-preserving voice analysis via disentangled representations. In *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, pages 1–14, 2020.

[8] Keshav Kumar, KR Ramkumar, and Amanpreet Kaur. A lightweight aes algorithm implementation for encrypting voice messages using field programmable gate arrays. *Journal of King Saud University-Computer and Information Sciences*, 34(6):3878–3885, 2022.

[9] Sabah Salih Hussein. Secure voice by using aes algorithm. Master's thesis, Aksaray Üniversitesi Fen Bilimleri Enstitüsü, 2017.

[10] Arup Kumar Chattopadhyay, Sanchita Saha, Amitava Nag, and Sukumar Nandi. Secret sharing: A comprehensive survey, taxonomy and applications. *Computer Science Review*, 51:100608, 2024.

[11] Ruopan Lai, Xiongjie Fang, Peijia Zheng, Hongmei Liu, Wei Lu, and Weiqi Luo. Efficient fragile privacy-preserving audio watermarking using homomorphic encryption. In *International Conference on Artificial Intelligence and Security*, pages 373–385. Springer, 2022.

[12] Shi-Xiong Zhang, Yifan Gong, and Dong Yu. Encrypted speech recognition using deep polynomial networks. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5691–5695. IEEE, 2019.

[13] Meng Feng, Chieh-Chi Kao, Qingming Tang, Ming Sun, Viktor Rozgic, Spyros Matsoukas, and Chao Wang. Federated self-supervised learning for acoustic event classification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 481–485. IEEE, 2022.

[14] Huan Zhao, Haijiao Chen, Yufeng Xiao, and Zixing Zhang. Privacy-enhanced federated learning against attribute inference attack for speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[15] Yaowei Han, Yang Cao, Sheng Li, Qiang Ma, and Masatoshi Yoshikawa. Voice-indistinguishability–protecting voiceprint with differential privacy under an untrusted server. In *Proceedings of the 2020 ACM SIGSAC*

Conference on Computer and Communications Security, pages 2125–2127, 2020.

[16] Ying Zhao and Jinjun Chen. A survey on differential privacy for unstructured data content. *ACM Computing Surveys (CSUR)*, 54(10s):1–28, 2022.

[17] Yaowei Han, Sheng Li, Yang Cao, Qiang Ma, and Masatoshi Yoshikawa. Voice-indistinguishability: Protecting voiceprint in privacy-preserving speech data release. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.

[18] Alexandru Nelus and Rainer Martin. Privacy-preserving audio classification using variational information feature extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2864–2877, 2021.

[19] Peng Huang, Yao Wei, Peng Cheng, Zhongjie Ba, Li Lu, Feng Lin, Fan Zhang, and Kui Ren. Infomasker: Preventing eavesdropping using phoneme-based noise. In *NDSS*, 2023.

[20] Zhiyuan Yu, Shixuan Zhai, and Ning Zhang. Antifake: Using adversarial audio to prevent unauthorized speech synthesis. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 460–474, 2023.

[21] Peng Cheng, Yuexin Wu, Yuan Hong, Zhongjie Ba, Feng Lin, Li Lu, and Kui Ren. Uniap: Protecting speech privacy with non-targeted universal adversarial perturbations. *IEEE Transactions on Dependable and Secure Computing*, 21(1):31–46, 2023.

[22] Zhuo Ma, Yang Liu, Ximeng Liu, Jianfeng Ma, and Feifei Li. Privacy-preserving outsourced speech recognition for smart iot devices. *IEEE Internet of Things Journal*, 6(5):8406–8420, 2019.

[23] Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng. Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion. *arXiv preprint arXiv:2106.10132*, 2021.

[24] Jiangyi Deng, Fei Teng, Yanjiao Chen, Xiaofu Chen, Zhaohui Wang, and Wenyuan Xu. {V-Cloak}: Intelligibility-, naturalness-& {Timbre-Preserving}{Real-Time} voice anonymization. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5181–5198, 2023.

[25] Yi Xie, Zhuohang Li, Cong Shi, Jian Liu, Yingying Chen, and Bo Yuan. Enabling fast and universal audio adversarial attack using generative model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14129–14137, 2021.

[26] Ang Li, Yixiao Duan, Huanrui Yang, Yiran Chen, and Jianlei Yang. Tiprdc: task-independent privacy-respecting data crowdsourcing framework for deep learning with anonymized intermediate representations. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 824–832, 2020.

[27] Li Luo and Yining Liu. Voice fence wall: User-optional voice privacy transmission. *Journal of Information and Intelligence*, 2(2):116–129, 2024.

[28] Sattar B Sadkhan, Abbas A Mahdi, and Rana S Mohammed. Recent audio steganography trails and its quality measures. In *2019 First International Conference of Computer and Applied Sciences (CAS)*, pages 238–243. IEEE, 2019.

[29] Zhaopin Su, Guofu Zhang, Zhiyuan Shi, Donghui Hu, and Weiming Zhang. Message-driven generative music steganography using midi-gan. *IEEE Transactions on Dependable and Secure Computing*, 2024.

[30] Chien-Chang Chen and Jian-Ying Huang. Progressive share of secret audio by chinese remainder theorem and integer wavelet transform. *International Journal of Electronic Commerce Studies*, 5(2):219–232, 2014.

[31] Yingbin Liang, H Vincent Poor, Shlomo Shamai, et al. Information theoretic security. *Foundations and Trends® in Communications and Information Theory*, 5(4–5):355–580, 2009.

[32] Deborah Pelacani Cruz, George Strong, Oscar Bates, Carlos Cueto, Jiashun Yao, and Lluis Guasch. Convolve and conquer: Data comparison with wiener filters. *arXiv preprint arXiv:2311.06558*, 2023.

[33] Naofumi Aoki. Technique of improving speech restoration from spectrograms of short windows for the griffin–lim algorithm. *Acoustical Science and Technology*, 44(3):186–188, 2023.

[34] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[35] Ziqi Yang, Lijin Wang, Da Yang, Jie Wan, Ziming Zhao, Ee-Chien Chang, Fan Zhang, and Kui Ren. Purifier: Defending data inference attacks via transforming confidence scores. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10871–10879, 2023.

BIOGRAPHY

Changxiang Zhao is a master student at City University of Macau. His research field is voice privacy protection. E-mail: D23091100445@cityu.edu.mo

Jianping Cai is an assistant professor and a supervisor of Master student at City University of Macau . His major research field is differential privacy, federated learning, matrix analysis and optimization theory.

Ximeng Liu is a professor and a supervisor of Doctor student at City University of Macau . His major research field is big data security, AI security, Mobile Crowdsensing, and cloud security.

Qi Zhong is an assistant professor and a supervisor of Master student at City University of Macau . Her major research field is copyright protection for machine learning models, backdoor in machine learning , adversarial machine learning, and data privacy protection.

Zuobin Ying is an associate professor and a supervisor of Doctor student at City University of Macau . His major research field is privacy-preserving computation, blockchain, cryptography and cloud-edge-end security.