Secure Audio Reverberation over Cloud

M. Abukari Yakubu  
Department of Applied Computer Science  
University of Winnipeg  
Winnipeg, MB, Canada  
Email: mohammedyakubu-a@webmail.uwinnipeg.ca

Pradeep K. Atrey  
Department of Computer Science  
University at Albany, SUNY  
Albany, NY, USA  
Email: patrey@albany.edu

Namunu C. Maddage  
NextGmultimedia  
Melbourne, Australia  
Email: n.maddage@nextgmultimedia.com

Abstract—Most individuals, governments and companies who outsource audio content to Cloud Data Centers (CDCs) for storage also make use of their high-end computing services. However, security and privacy issues related to CDCs make it difficult for data to be processed without compromising security. In this work, we propose a secure method for artificially adding reverberation effects to an audio secret over cloud with (K,N) Shamir's Secret Sharing (SSS) as the cryptosystem for audio recording, reproduction, editing and enhancement. Our method implements convolution reverb in encrypted domain by convolving a modeled impulse response of an acoustic space with shares of the audio secret. Experimental results show that our proposed method in Encrypted Domain (ED) produces similar results as compared to performing the same operations in Plaintext Domain (PD).

Keywords—Encrypted domain processing, cloud security, audio reverberation effect, convolution reverb, Shamir's secret sharing scheme (SSS).

I. INTRODUCTION

With the advent of cloud computing, which provides a framework of services for data storage, high-end computing and online access to computer resources, companies are saving cost and slashing investment on resources by outsourcing storage and processes to Cloud Service Providers (CSP). Audio content is one type of data outsourced to CDCs for storage and computing services by individuals, governments and companies such as the audio recording and reproduction industries (record labels) and forensic analytical providers, etc. Most of these recordings are confidential and might contain sensitive information like names, credit cards, evidence to be used in a court of law by a jury, information with national security implications, etc.

Because of the security and privacy issues of using third-party servers like CDCs, companies first encrypt confidential audio content before uploading it to CDCs. In such cases, encryption schemes like Advanced Encryption Standard (AES) are used, which suffers from single point vulnerability, meaning that the security of the method lies in securing the encryption key which is usually entrusted to the sender and receiver. Thus an adversary with access to the encryption key can obtain the confidential data. Most companies use not only CSP storage services, but also high-end computing services. When the need arises for some processing to be performed on these audio secrets, the third-party server will first have to decrypt the secret which will expose the confidential information. This makes the confidential data vulnerable to exploitation by an adversary. In the case of sound recording companies (e.g. record labels) millions of dollars might be lost if a record in production is exposed. Hence, secure processing of such confidential audio is of utmost importance.

In order to perform processing in ED over cloud, encrypted signals have to be processed without decryption. This way, the security of the audio secret is not compromised. To this end, fields in signal processing and cryptography have been merged to develop a totally new interdisciplinary framework called Signal Processing in Encrypted Domain (SPED)[1]. Work done so far in this area has applied cryptographic primitives such as Secure Multiparty Computation (SMC), Commitment Schemes, Zero-knowledge Protocols (ZKP), Private Information Retrieval and homomorphic encryption to develop schemes based on the security requirements of the application scenario to make secure signal processing possible. SPED has been applied in applications such as secure processing of medical data (MRI, ECG, DNA) [2], secure digital watermarking [3], Data mining on private databases [4], [5], Protecting Privacy in video surveillance systems [6], etc.

Although some work has been done in SPED for images, videos etc., it has been scarcely explored in audio. Table I details the comparison of work on audio processing in encrypted domain. The authors in [7], [8] propose a technique to identify and verify speakers over encrypted voice over IP (VoIP) conversations. Here, Variable Bit Rate (VBR) and Voice Activity Detector (VAD) used to encode speech over a VoIP channel achieves lower bit rates but produces a packet-length which is dependent on the speaker. Whereas the authors in [7] employ VBR encoding, the authors in [8] use the Voice Activity Detector approach. Both works use this speaker dependent packet-length information extracted from encrypted VoIP signals to build models for identification and verification. The authors of [9] and [10] also present a framework for speaker verification/identification and sound recognition/classification, respectively, using Gaussian Mixture Models (GMM) in encrypted domain. Both methods are based on SMC and homomorphic encryption [like the Paillier cryptosystem and BGN cryptosystem] which enables computation and classification to be performed in a secure manner. The framework in [9] enables computation on encrypted speech data without revealing the actual voice patterns in plaintext. The framework in [10] involves two parties (party A and party B), party A providing the data and party B providing the recognition algorithm (classifier). Party B then applies his classification algorithm to party A’s data in such a way that the data and results are not revealed to party B, thereby maintaining the privacy of the data. The security of these systems lies in protecting the encryption and decryption keys as an adversary with access to the keys can obtain the plaintext data. Moreover,
TABLE I: Comparison of Work on Audio Processing in Encrypted Domain

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Task in Encrypted Domain</th>
<th>Cryptographic Primitive</th>
<th>Techniques Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>[7]</td>
<td>Speaker identification and verification over encrypted voice over IP (VoIP) conversations</td>
<td>AES</td>
<td>Variable Bit Rate (VBR) and hidden Markov models (HMMs)</td>
</tr>
<tr>
<td>[8]</td>
<td>Speaker identification and verification over encrypted voice over IP (VoIP) conversations</td>
<td>AES</td>
<td>Voice Activity Detector (VAD)</td>
</tr>
<tr>
<td>[9]</td>
<td>Speaker verification / identification</td>
<td>SMC and homomorphic encryption</td>
<td>Gaussian Mixture Models (GMM)</td>
</tr>
<tr>
<td>[10]</td>
<td>Sound recognition / classification</td>
<td>SMC and homomorphic encryption</td>
<td>Gaussian Mixture Models (GMM)</td>
</tr>
<tr>
<td>Proposed Scheme</td>
<td>Addition of reverb effect</td>
<td>Homomorphic encryption (SSS)</td>
<td>Convolution with modeled reverb impulse response</td>
</tr>
</tbody>
</table>

these methods are computationally expensive as a result of the large message space (1024 bits) and exponentiation operation of the Paillier cryptosystem.

In this work, we focus on the addition of reverb effects to an audio recording in ED over cloud. This is one of the most widely used delay effects, amongst others like flanging, phasing, chorus effects etc., for audio recording, reproduction, editing and enhancement. This effect adds an acoustic environment to an audio recording to make it sound realistic. The resulting reverb affected audio inherits characteristics from that acoustic environment and sounds as if the recording was created in that environment. Reverberation is a series of delayed and attenuated sound waves reflected within an acoustic environment which is perceived by the human ear in less than 0.1 seconds after the original sound wave. The human auditory system is unable to perceive the 0.1 second delay and interprets the original sound wave and delayed reflections as one prolonged sound. This effect is different from echo where delays are more than 0.1 seconds and the delayed sounds are perceived distinctly as decaying copies of the original sound. Key areas of applications of reverb effects in audio are:

1) The recording industry for audio production, editing and enhancement.
2) Audio forensics for analyzing/simulating an audio recording in different acoustic environments.
3) Simulation of acoustic reverberation for dereverberation algorithms, etc.

In order to protect an audio secret to 1) eliminate single point vulnerability of widely used AES, 2) avoid large message space and high computational complexity of homomorphic schemes like the Paillier cryptosystem and 3) make processing on the encrypted audio secret possible, we employ Shamir’s Secret sharing (SSS) to encrypt the secret into a number of shares that can be distributed among a number of CDCs such that only more than a certain number of shares can be retrieved by an authorized user to reconstruct the secret; individual shares are of no use on their own. The homomorphic property (addition and multiplication) of SSS makes ED processing possible. Furthermore, SSS is information theoretically secure and $(K,N)$ threshold.

In this work, we propose the implementation of convolution reverb to artificially add reverb effects to an encrypted audio secret over cloud to ensure information assurance of client’s data from a privacy and security perspective. To the best of our knowledge, this is the first work that applies reverb effect to an audio signal in the encrypted domain.

The rest of this paper is organized as follows. In Section II, we discuss Shamir’s secret sharing scheme. We discuss artificial reverberation techniques in Section III. The proposed method for encrypted domain convolution reverb is detailed in Section IV, and Section V discusses the experimental results. We conclude the paper in Section VI.

II. SSS Scheme

Shamir introduced his scheme in 1979 [11]. His scheme is based on polynomial interpolation. The goal of this scheme is to divide data into $N$ shares such that:

1) Any $K$ or more shares can reconstruct the secret.
2) $K-1$ or fewer shares cannot reconstruct the secret.

Such a scheme is called a $(K,N)$ threshold scheme where $2 \leq K \leq N$, $N$ is the number of shares and $K$ is the least number of shares required to reconstruct the secret. To share a secret $S$ among $N$ participants, a polynomial function $f(x)$ is constructed with degree of $K - 1$ using $K$ random coefficients $a_1, a_2 \ldots a_{K-1}$ in a finite field $GF(q)$ where $a_0$ is $S$, and $q$ is a prime number greater than $a_0$.

$$f(x) = (a_0 + a_1 x + \cdots + a_{K-1} x^{K-1}) \mod q \quad (1)$$

Any $K$ out of $N$ shares can reconstruct the secret using Lagrange interpolation to reconstruct the polynomial $f(x)$; the secret can be obtained at $f(0)$ i.e. $f(0) = a_0 = S$

$$f(x) = \sum_{j=1}^{K} y_j \prod_{i=1,i \neq j}^{K} \frac{x-x_i}{x_j-x_i} \mod q \quad (2)$$

A. Homomorphic Encryption

A cryptosystem is homomorphic if computation on its ciphertext yields an encrypted result, which when decrypted will match the result of the computation on its plaintext. SSS is homomorphic on addition and multiplication. Let $m_1$ and $m_2$ belong to the plaintext space of some cryptosystem, and $E(.)$ and $D(.)$ denote the encryption and decryption functions respectively. Then SSS satisfies the conditions below:

$$D(E(m_1) + E(m_2)) = m_1 + m_2 \quad (3)$$
$$D(E(m_1) \times E(m_2)) = m_1 \times m_2 \quad (4)$$

Examples of additive homomorphic cryptosystems are Paillier [12] and Benaloh [13], and multiplicative homomorphic cryptosystems are RSA [14] and ElGamal [15].
III. ARTIFICIAL REVERBERATION TECHNIQUES

There are many techniques for applying reverb effects to audio. The most common techniques for artificial addition of reverb are:

1) Filter banks and delay line: this approach involves connecting filters (comb, all-pass, lowpass filters) in parallel and series and adjusting their parameters to produce the desired reverb effect, e.g. Schroeders Reverberator [16], Moorers Reverberator [17] etc.

2) Convolution reverb: acoustic space is a Linear Time-Invariant (LTI) system, and like any LTI system its impulse response can be modeled and convolved with an audio signal to produce the effects of that space. Modeling of the impulse response depends on the application scenario and needs of the designer which is beyond the scope of this paper. Some examples of modeled impulse responses are room, concert hall, cathedral, bottle hall, conic long echo hall, deep space, etc. In this work, we apply this approach to artificially add reverb effect to an audio secret in ED. Let \( h[n] \) be the modeled impulse response and \( x[n] \) be the audio signal. Then, below discrete convolution will yield the reverb effected signal \( y[n] \),

\[
y[n] = x[n] * h[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k]
\]

IV. PROPOSED WORK

The audio secret to be reverb effected is encrypted by creating shares with the SSS \((K,N)\) threshold scheme on the client’s system. The client then uploads each of the \( N \) shares to \( N \) different non-colluding CDCs. The CDC then applies the reverberation effect to their hosted shares by convolving with the desired impulse response. Reverb impulse responses are public signals in plaintext. The client does not need to transmit or upload impulses response to the CDC since the CSP can obtain a library of all impulse responses. An authorized user then reconstructs the reverb effected secret by combining at least \( K \) out of \( N \) processed shares.

Signal processing operations often involve real-valued signal amplitudes; however, using real numbers in a cryptosystem means excluding the modular prime operation, which in the case of SSS, degrades security. Therefore, we have to preprocess the real valued samples of the audio secret to positive integer values. Steps 1 and 2 below preprocesses the original signal before creating shares. The below steps details our method in ED.

**Step 1**: Scale real-valued signal amplitudes with constant factor \(10^d\) where \( d \) is an integer value. Roundoff error is bounded by:

\[
-\frac{1}{2} \times 10^{1-d} \leq \epsilon \leq \frac{1}{2} \times 10^{1-d}
\]

where \( \epsilon \) is the rounding error and \( d \) is the rounding precision.

**Step 2**: Shift signal obtained from equation (6) to first quadrant by a constant additive shift \( \gamma \) to avoid negative numbers.

\[
A' = ((A + \gamma) \times 10^d) + \gamma
\]

**Step 3**: Create \( N \) shares \((S_1, S_2 \ldots S_N \in S)\) of preprocessed signal \( A' \) using Equation (1) of SSS in the finite field of \( q \)

where \( q > max(A') \). Upload the \( N \) shares to \( N \) non-colluding CDCs.

**Step 4**: Convolution of each share on the CDC. The modeled impulse response is real-valued and there are instances where some samples are negative. This might result in errors while performing convolution reverberation. In order to avoid errors, perform the following steps:

1) Scale impulse response \( h \) by \(10^t\) that is \( h' = h \times 10^t\); where \( t \) is an integer value.

2) Modify Equation (5) by adding a constant additive shift \( \beta \) to avoid negative samples. \( \beta = \lceil \frac{q}{2} \rceil \) if \( h' \) has a negative sample else \( \beta = 0 \), where \( \lceil \cdot \rceil \) is a ceiling function.

Convolve each share \((S_1, S_2 \ldots S_N \in S)\) with \( h' \) using Equation (8) to obtain \((S_1', S_2' \ldots S_N' \in S')\).

\[
S_i'[n] = \left( S_i[n] * h'[n] + \beta \right) \mod q
\]

\[
= \left( \sum_{k=0}^{L-1} S_i[k]h'[n-k] + \beta \right) \mod q , \ i \in \{1,2\ldots N\}
\]

Where \( L \) is the number of samples of each share.

**Step 5**: An authorized user can reconstruct the reverb effected secret by putting together \( K \) processed shares from any of the \( N \) CDCs using Lagrange interpolation from Equation (2).

**Step 6**: Postprocess the reconstructed secret to reverse engineer the preprocessing done in the above steps. 1) Subtract the additive shift \( \beta \) and divide by the scale factor of \( h' \) which is \(10^t\). 2) Subtract the additive shift \( \gamma \) and divide by the scale factor \(10^d\) from equations (7) and (6) respectively.

A. Data Overhead

Our proposed scheme introduces some data overhead due to the preprocessing steps before creating shares. This data overhead is the number of bits used to represent the maximum preprocessed audio sample which is the same as the transmission overhead of each share to the CDC. Since shares are generated under a finite field bounded by the modulo prime \( q \), we can conclude that the data overhead to transmit a share to each CDC is also bounded by the number of bits representing \( q \).

B. Security Analysis

The proposed method is based on the SSS \((K,N)\) threshold scheme which is proven to be information theoretically secure [18]. Since the generation of shares is bounded by \( q \), it follows that samples of each share are within the set \( \{0,1,2,...,q-1\} \) and, referring to equation (1) of SSS, each sample of a share is a unique polynomial. In this case, an adversary will have to guess with a probability of \( \frac{1}{q} \) making it highly unlikely to infer a secret sample from its share.

Audio signals, by nature, have correlating adjacent samples. This correlation reduces the entropy (degree of uncertainty) of the entire signal; i.e. a sample can be predicted from

ASIA ’15
<table>
<thead>
<tr>
<th>Test File (.wav)</th>
<th>Length (secs)</th>
<th>Bits/Sample</th>
<th>Sampling Frequency (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>audio1</td>
<td>2</td>
<td>16</td>
<td>16000</td>
</tr>
<tr>
<td>audio2</td>
<td>43</td>
<td>16</td>
<td>8000</td>
</tr>
<tr>
<td>audio3</td>
<td>13</td>
<td>16</td>
<td>11025</td>
</tr>
<tr>
<td>audio4</td>
<td>14</td>
<td>8</td>
<td>44100</td>
</tr>
<tr>
<td>audio5</td>
<td>4</td>
<td>8</td>
<td>8000</td>
</tr>
<tr>
<td>audio6</td>
<td>2</td>
<td>32</td>
<td>8000</td>
</tr>
</tbody>
</table>

TABLE III: Average Processing Time (ms)

<table>
<thead>
<tr>
<th>Test file</th>
<th>Share Creation</th>
<th>ED Processing</th>
<th>Reverb Effected Secret Reconstruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>audio1</td>
<td>205.18</td>
<td>97.77</td>
<td>17.81</td>
</tr>
<tr>
<td>audio2</td>
<td>1620.27</td>
<td>631.44</td>
<td>80.32</td>
</tr>
<tr>
<td>audio3</td>
<td>747.46</td>
<td>276.57</td>
<td>39.58</td>
</tr>
<tr>
<td>audio4</td>
<td>2802.95</td>
<td>1068.85</td>
<td>115.98</td>
</tr>
<tr>
<td>audio5</td>
<td>219.09</td>
<td>73.21</td>
<td>18.64</td>
</tr>
<tr>
<td>audio6</td>
<td>149.20</td>
<td>44.26</td>
<td>15.02</td>
</tr>
</tbody>
</table>

its adjacent samples as in Linear Predictive Coding (LPC). However, the use of random coefficients as a blinding factor in Equation (1) to generate shares eliminates this correlation. Thus, individual shares do not reveal information about the secret audio. In the future we hope to examine the information theoretical security implications of our scheme with respect to the preprocessing steps.

V. EXPERIMENTAL RESULTS

Table II details the test audio files obtained from [19] that we used for experimenting with our proposed method. We test our method with a modeled impulse response [20] from the acoustic environment of a living room. That is, we add the effects of a living room to our audio test files in ED. In the $(K,N)$ threshold SSS scheme, we set $K = 2$ and $N = 3$, implying that we created three shares of the audio secret such that at least two shares will be required to reconstruct the reverb effected audio secret.

We implemented the proposed method using MATLAB14 on a 2.53GHz i5 CPU with 4GB RAM. Table III details the processing time for creating secret shares, applying the reverberation effect in encrypted domain and reconstructing the audio secret. The table suggests that the complexity of reconstructing the secret is relatively lower than creating shares and ED processing. Our method is applied on a sample basis. As a result, the processing time is directly proportional to the audio bit rate, which is associated with the sampling frequency and number of bits per sample. Therefore, the greater the length of the signal, the greater the complexity. This is evident for audio2.wav and audio4.wav with the greatest complexities.

Fig. 1 shows similarity scores between the PD and ED reverb effected signals. We computed the similarities using Pearson’s correlation method. Results suggest the reverb effected signal, after applying our method, correlates about 99.99% with normal PD processing. Thus, our method yields identical results to PD processing while maintaining security and privacy. The 0.01% difference can be accounted for by the rounding-off during the preprocessing steps for both original audio secret $A$ and impulse response $h$. We hope to optimize the solution in the future to further minimize round-off errors.

Fig. 2: Modeled room impulse response

Fig. 3: Time domain plots of audio1.wav
Time domain plots of audio1.wav are represented in fig. 3 showing the audio secret, one of its shares and the reverb effected reconstructed secret. The time series reveals that 1) the share is noise and likely to have equal power across all frequencies and 2) the amplitude series of the reverb effected reconstructed secret shows some low amplitude regions as compared to the audio secret. This is as a result of the delay and decay effect of the impulse response shown in fig. 2, which verifies that the audio secret has been reverb effected.

VI. CONCLUSION

The use of Cloud computing is growing on a continuous basis. In order for governments and companies to entrust computation of their data to Cloud Service Providers (CSP), security issues should be paid much attention. Though there are policies governing the operations of CSPs, this is not enough to guarantee the security and privacy of data. Researchers in the fields of mathematics, computer science and engineering are currently developing encryption protocols and computational tasks possible in Encrypted Domain (ED). Only then will CSP services be fully embraced without fear of security or privacy issues. As a contribution to realize this, we have proposed in this work a secure artificial addition of reverb effect to an audio secret over cloud by using $(K,N)$ Shamir’s secret sharing (SSS) scheme as our cryptosystem. Our method implements convolution reverb and can be applied to any reverb impulse response and an audio secret in ED over cloud. Experimental results reveal that our proposed method is efficient and yields similar results to Plaintext Domain (PD).

REFERENCES


