# Framework for VoIP Authentication using Session ID based on Modified Vector Quantization

Yazid Jaafar, Azman Samsudin, Alfin Syafalni and Mohd Adib Omar

*School of Computer Sciences, Universiti Sains Malaysia*
*11800 USM, Penang, Malaysia*
*mymj10_com116@student.usm.my,*
*azman@cs.usm.my,as10_com083@student.usm.my, adib@cs.usm.my*

***Abstract***

*The Session Initiation Protocol (SIP) is the main protocol behind Voice over IP (VoIP). However, it does not provide authentication, which may lead to possible impersonation and eavesdropping threats. Negotiating keys using digital certificates may help secure the channel, but this method incurs extra maintenance cost. Verbal authentication utilizes the real-time nature of VoIP but it requires users to read and manually compare the authentication code. This study proposes a session identifier framework based on the Modified Vector Quantization (MVQ) method on real-time video frames for video communication in VoIP. After the selected image frames are averaged and quantized, the output of the MVQ process is a set of image metrics that serves as pre-shared keys for key agreement. The framework is certificate-less and users do not need to read the authentication code to the other user. The implementation is evaluated for accuracy and robustness towards network noise and frame conditions.*

***Keywords:*** *Multimedia Security, VoIP, Eavesdropping, Cryptography*

## 1. Introduction

Voice over IP (VoIP) is a telephony system that transmits audio and video data using an IP network. Video call and multimedia conference are some of its applications. VoIP is not geographically constrained; thus its user has full flexibility and maintains a single ID across the globe. All real-time communication data are transmitted through the same channel as a local network. This reduces the overall cost and optimizes the bandwidth. In comparison, a traditional landline phone requires a unique number for a particular area and therefore requires greater expense to have the same flexibility as VoIP.

In most cases, VoIP is implemented on top of an existing IP network. This causes VoIP to inherit all its vulnerabilities [1]. In contrast, landline phones secure the analog data up to the physical layer, considering all analog data are transmitted using a circuit-switched network. Session Initiation Protocol (SIP) is a text-based signalling protocol developed primarily to establish, manage, and terminate the VoIP session. It provides simplicity by following challenge-response mechanisms but does not provide any peer authentication. A digital certificate from a trusted third party, known as a Public Key Infrastructure (PKI), is a popular alternative to address the issue. However, this introduces extra renewal costs [2]. Other methods need a complex exchange of key materials to achieve mutual authentication, which could cause communication overhead [3–5].

Verbal authentication is an alternative verification method, which depends on human intelligence to perform manual authentication. ZRTP is the first authentication protocol based on this method [6], followed by voice interactive personalized security (VIPSec) [7]. Verbal

authentication does not use a digital certificate and is independent of a trusted third party. The authentication is performed manually by session participants where both speak and validate the authentication code. The code is a string of characters that is not exchanged during the key agreement phase but is "interlocked" with the key materials. The authentication code in ZRTP is the hash of Diffie-Hellman's shared key, whereas VIPSec uses the hash of random objects selected by the user prior to the VoIP session.

In this paper, we introduce a framework of session identifier based on the Modified Vector Quantization (MVQ) for VoIP authentication. The proposed framework is developed specifically for video call sessions, the output of which will be a unique Session Identifier ($SID$) that acts as a pre-shared key for VoIP authentication. We also propose a new MVQ approach that enhances the simplicity and robustness of the original method in creating keys from image [8]. Our contribution is twofold: First, we forgo the need to speak the authentication code; and second, mutual authentication is achieved without using a digital certificate while eliminating the threat of a Man-in-the-Middle (MITM) attack.

## 2. Related Works

VoIP is prone to security threats from the underlying network [1]. Session eavesdropping and identity impersonation are two common issues [9]. Eavesdropping is a passive attack where the adversary taps and listens into the VoIP session, thus enabling information theft. Impersonation on the other hand, is an active attack where the adversary uses the stolen identity to initiate the session and appear as a legal caller.

The following notations are used in this paper:

| | |
|---|---|
| *Alice*: | Session initiator (caller) |
| *Bob*: | Session receiver (callee) |
| *Mallory*: | Malicious adversary |
| *SSK*: | Session shared key |
| *U*: | Public key |
| *R*: | Private key |
| *SID*: | Session identifier |
| $SID_{as}$or$SID_{bs}$: | *SID* created locally by Alice or Bob |
| $SID_{ar}$or$SID_{br}$: | *SID* created by Alice or Bob based on the received video stream |
| *EncU*: | Encrypted public key |

### 2.1. Man-in-the-Middle (MITM) Attack

Consider that Alice calls Bob, and Mallory wants to eavesdrop on their VoIP session. Key agreement is based on a cryptographic primitive that has been proven hard to be solved. However, Mallory does not need to break the cryptographic primitive to breach the session. Instead, she performs an MITM attack to bluff Alice and Bob [10]. Mallory intercepts the public key and replaces it with her key before re-transmitting to both of them [11].

Figure 1 illustrates the scenario of the MITM attack between Alice and Bob. Alice and Bob generate a key pair $(U_a, R_a)$ and $(U_b, R_b)$, respectively, and exchange $U_a$ and $U_b$ through an open network to create the $SSK$. Mallory also generates her key pair $(U_m, R_m)$ and intercepts $U_a$ and $U_b$. Then, she sends $U_m$ to Alice and Bob. Alice assumes that $U_m$ is Bob's public key and Bob is unaware that $U_m$ is coming from Mallory. Alice creates $SSK_{am}$ from $U_m \times R_a$, while Bob creates $SSK_{bm}$ from $U_m \times R_b$. Mallory generates $SSK_{am}$ and $SSK_{bm}$ using

$U_a \times R_m$ and $U_b \times R_m$ respectively. Finally, Alice successfully establishes a secure connection to Mallory using $SSK_{am}$, and Bob to Mallory using $SSK_{bm}$. In this case, since Alice and Bob cannot determine the authenticity of the received public key, the MITM attack is thus successful. Therefore, every public key needs to be authenticated to prevent Mallory from sending her key.

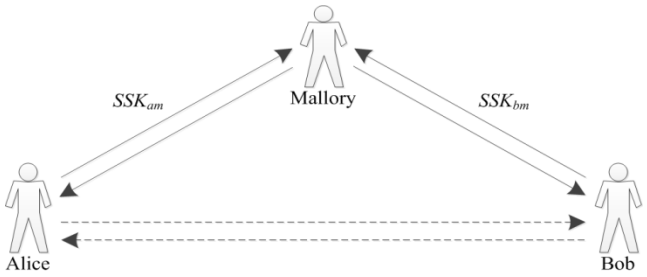


**Figure 1. MITM Attack between Alice and Bob**

### 2.2. Verbal Authentication in VoIP

Verbal authentication requires users to read out and compare the authentication code over the phone. If the code is exactly the same, then the $SSK$ is genuine, and the VoIP session is free from MITM attack. ZRTP is an example of a verbal authentication protocol for VoIP [6]. It takes full advantage of real-time communication between Alice and Bob. ZRTP uses the Diffie-Hellman (DH) key exchange to negotiate the keys. The resulting $SSK$ is hashed by using a hash-based message authentication code (HMAC) to derive the Short Authentication String ($SAS$). Alice reads her $SAS$ and compares it with Bob's $SAS$. If $SAS_A = SAS_B$, then the session is secure. ZRTP uses key continuity, an approach that generates a subsequent $SAS$ from the hash of a previous $SAS$.

However, the ZRTP identifier in the implementation of ZRTP can be manipulated by Mallory to ensure that every session will create a new $SAS$ value [12]. The author also presents certain situations where relay attacks on ZRTP are possible. Moreover, perfect forward secrecy is not guaranteed if the user chooses to ignore the $SAS$ authentication phase during the first session [13].

VIPSec is another authentication protocol for VoIP that is based on verbal authentication [7]. Similar to ZRTP, VIPSec requires Alice to read out the authentication code and verify with Bob over the phone. The authentication protocol is based on an asymmetric key encryption where a hash of random objects selected by the user serves as an "interlocking" mechanism for the exchanged public key. If the hash is exactly the same, then the session is fully secure. However, a relay attack, as observed in ZRTP [12], can still happen in VIPSec. If Mallory can somehow bypass the verbal authentication phase, the whole session is still vulnerable to an MITM attack.

## 3. Proposed Framework

We propose a framework of a Session Identifier ($SID$) based on the MVQ and specifically developed for video call in VoIP. A set of feature descriptors is derived from the image subspaces as a $SID$ to encrypt the public key. Only the receiver who receives the actual video can digest the same $SID$ and decrypt the received public key to eliminate eavesdropping and, ultimately, the MITM attack. This removes the need for verbal authentication and digital certificates. Figure 2 illustrates the proposed framework in detail.

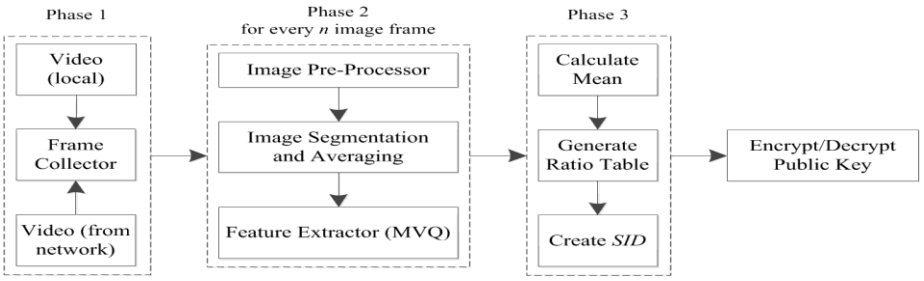


**Figure 2. The proposed framework**

### 3.1. Framework Description

Alice calls Bob and initiates the video call session. She clones the video source (hers and Bob's video) and channels it to the framework's engine at the moment she starts sending and receiving the RTP video transmission from Bob. Two separate threads are running concurrently, one for each video source. During the first five seconds, *n* frames are grabbed for the second phase. However, importantly, more frames can give better accuracy but lead to processing overhead and reduced robustness.

In the second phase, the image frames will be pre-processed if too much noise is present, given this can disrupt the value of the feature vector. Although each frame is visually similar, the feature vector can vary. An image averaging technique is needed to ensure every image is "smoothed," thereby reducing the effect of variation. Each frame is segmented into multiple quadrants called feature subspaces. In every subspace, all RGB values from each pixel are averaged to replace the original value based on Equation 1.

$$\overline{(r,g,b)} = \frac{\sum_{i=0}^{n}(r,g,b)_i}{n} \quad (1)$$

After the image is averaged, a set of feature vectors is extracted from the frame based on our MVQ approach. Figure 3 shows the difference of the original VQ approach [8] and the MVQ. In VQ, the image is segmented into multiple subspaces. Each subspace contributes a unique identity and becomes part of the concatenated string, $\delta = (a||b||c||d||e||f||g||h||i)$ where $\delta$ is the authentication code and $(a, b, c, d, e, f, g, h, i)$ is the feature descriptor. This method has a higher sensitivity towards packet drop, considering a disturbance in a single subspace will disrupt the accuracy of $\delta$. In contrast, MVQ uses the subspace only for averaging the pixel by taking advantage of parallel processing. We use the ratio of red $(r)$, green $(g)$, and blue $(b)$ levels as the feature descriptor, considering they are easy and fast to calculate. The [*r*,*g*,*b*] level is extracted from the whole frame for the third phase of the framework.

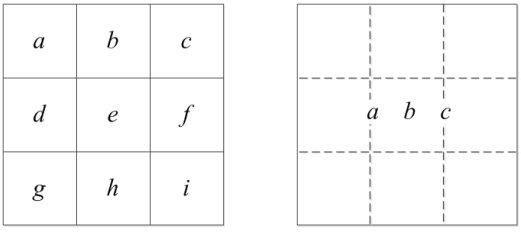


**Figure 3. Comparison of (a) Original Vector Quantization (VQ) and (b) Modified Vector Quantization (MVQ)**

| (x/y) | *r* | *g* | *b* |
|---|---|---|---|
| *r* | | *rg* | *rb* |
| *g* | *gr* | | *gb* |
| *b* | *br* | *bg* | |

**Figure 4. The generated ratio table after feature extraction**

In the third phase, the mean arithmetic of [$r$,$g$,$b$] levels from all frames are calculated to generate the ratio table. Figure 4 describes the ratio table generated and how it is calculated. The output is a set of feature vectors that consists of $\delta = (rg||\, rb||\, gr||\, gb||\, br||\, bg)$. Alice hashes the feature vectors to form the final eight characters of the authentication code to create the local session identifier, $SID_{as}$. Finally, Alice uses $SID_{as}$ to encrypt her public key using the Advance Encryption Standard (AES) algorithm, $EncU_a = E_{SIDas}(U_a)$, and sends it to Bob. At the same time, Bob performs the same operation and sends his public key, $EncU_b = E_{SIDbs}(U_b)$, to Alice.

Alice and Bob generate new session identifiers based on the received video stream, creating $SID_{ar}$ and $SID_{br}$, respectively. At this stage, the value of $SID$ being calculated locally should be same as the value calculated by the peer, considering the calculation is performed on the same video. For instance, if Alice gets $y = f(x)$, where $x$ is her video, Bob should get the same $y = f(x)$, considering he calculated the value of $y$ based on the video stream $x$ that he received from Alice. Therefore, $SID_{as} = SID_{ar}$ and $SID_{bs} = SID_{br}$. Finally, Alice and Bob decrypt the encrypted public key with $SID_{ar}$ and $SID_{br}$, respectively. If the generated $SSK_{ab}$ can be used to encrypt and decrypt the subsequent video stream, they can safely assume that the $U_a$ and $U_b$ are genuine, and the session is secured from MITM attack.

### 3.2. Security Model

During the MITM attack, getting the public key alone is not enough. Mallory still needs Alice and Bob to receive her public key for the MITM attack to work [10]. Usual key agreement schemes allow Mallory to intercept and replace the public key before it reaches the intended receiver. In the proposed framework, the encrypted public key, *EncU*, is the only message that is exchanged between Alice and Bob. The rest is real-time video stream.

Given that $EncU = E(U)_{SID}$ and $SID = f(\delta)$, where $\delta$ is the operation performed on the video stream, the only way for Mallory to send her $U_m$ is by transmitting her own video stream. Consider that Alice and Bob initiate the video session and begin generating $SID_{as}$ and $SID_{bs}$, respectively. At the same time, Mallory also prepares $SID_{ms}$. When the video transmission starts, Mallory intercepts both transmissions and $EncU_a$ and $EncU_b$ from Alice and Bob, respectively. Mallory should be able to decrypt $EncU_a$ and $EncU_b$, given that she can create $SID_{ar}$ and $SID_{br}$ based on the received video stream. However, although Mallory can communicate directly with the sender, impersonation is nearly impossible since they can visually verify that they are not talking to each other, but to Mallory instead.

In order to eavesdrop, Mallory needs to send her $EncU_m$ to Alice and Bob. Given that $U = E_{SID}^{-1}(EncU)$, both Alice and Bob cannot create $SID_{mr}$ and decrypt the public key unless Mallory transmits her video. Therefore, Mallory cannot send the fake public key without

Mallory exposing herself to Alice and Bob. The $SID$ works as an "interlocking" mechanism that prevents Mallory's MITM attack from succeeding.

## 4. Evaluation

The proposed framework is evaluated to find the relationship between the number of frames and the consistency of the $\delta$ value. Our hypothesis is that the level of consistency is parallel to the number of frames. A large variance in $\delta$ causes $SID_{ar} \neq SID_{as}$ and $SID_{br} \neq SID_{bs}$. In such a situation, Alice and Bob cannot decrypt the received public key; thus, SSK cannot be created. Hence, a consistent $\delta$ value is important. The difference between averaged and non-averaged image frames is also examined. The experiment is conducted on two identical desktops with Intel Dual Core 2.6 GHz processor and 2 GB RAM equipped with webcam and an Ethernet LAN connection. Table 1 describes the result of the experiment.

**Table 1. Comparison of averaged and non-averaged image frames**

| No. of frames ($n$) | Averaged | | | | Non-averaged | | | |
|---|---|---|---|---|---|---|---|---|
| | $r$ | $g$ | $b$ | Time ($ms$) | $r$ | $g$ | $b$ | Time($ms$) |
| 2 | $\alpha+1$ | $\alpha+1$ | $\alpha+1$ | 94 | $\alpha+2$ | $\alpha+3$ | $\alpha+3$ | 101 |
| 3 | $\alpha+1$ | $\alpha+1$ | $\alpha+1$ | 141 | $\alpha+2$ | $\alpha+2$ | $\alpha+3$ | 143 |
| 4 | $\alpha$ | $\alpha+1$ | $\alpha+1$ | 179 | $\alpha+1$ | $\alpha+2$ | $\alpha+2$ | 178 |
| 5 | $\alpha$ | $\alpha$ | $\alpha+1$ | 216 | $\alpha+1$ | $\alpha+1$ | $\alpha+2$ | 218 |
| 6 | $\alpha$ | $\alpha$ | $\alpha$ | 267 | $\alpha+1$ | $\alpha+1$ | $\alpha+2$ | 268 |
| 7 | $\alpha$ | $\alpha$ | $\alpha$ | 287 | $\alpha$ | $\alpha+1$ | $\alpha+1$ | 289 |
| 8 | $\alpha$ | $\alpha$ | $\alpha$ | 323 | $\alpha$ | $\alpha$ | $\alpha+1$ | 325 |
| 9 | $\alpha$ | $\alpha$ | $\alpha$ | 352 | $\alpha$ | $\alpha$ | $\alpha+1$ | 355 |
| 10 | $\alpha$ | $\alpha$ | $\alpha$ | 393 | $\alpha$ | $\alpha$ | $\alpha$ | 396 |
| 11 | $\alpha$ | $\alpha$ | $\alpha$ | 424 | $\alpha$ | $\alpha$ | $\alpha$ | 478 |
| 12 | $\alpha$ | $\alpha$ | $\alpha$ | 448 | $\alpha$ | $\alpha$ | $\alpha$ | 451 |

$\alpha$= value of respective color level

The results of the experiment show that although an increase in the number of frames can improve the $\delta$ consistency, it also increases the processing time. This can be improved further by optimizing the parallel processing on the Graphic Processing Unit (GPU). Other than that, it takes the averaged images only 6 frames to reach full consistency, whereas the non-averaged images takes 10 frames. Thus, an image averaging technique can reduce the processing overhead by "smoothing" the explicit details on images.

## 5. Conclusion

We present a Session Identifier ($SID$) framework based on the MVQ approach to digest a unique authentication code. A number of image frames are captured from the video stream before they are segmented and averaged. Subsequently, a set of feature descriptors is extracted to generate a unique $SID$ for the caller and callee. The $SID$ is then used to encrypt and decrypt the public key during the key agreement. Only the original sender can generate the same $SID$, decrypt the public key, and generate the *SSK*.

Unlike verbal authentication, this framework allows Alice and Bob to authenticate the received public key automatically without having to read out the authentication code. The framework also does not rely on a digital certificate to eliminate an MITM attack. The results of the experiment show that the generated $SID$ becomes more accurate if more image frames

are collected. The image averaging technique used in the framework has significantly improved the processing overhead and time consumption. For future improvement, a large part of segmentation and averaging operation will be optimized on a GPU, which can further reduce the time taken in generating the $SID$.

## Acknowledgements

The authors would like to thank the anonymous reviewer for their useful comments. This research is supported by Short Term Grant Scheme (grant number = 304/PKOMP/6312091) from Universiti Sains Malaysia, and government of Malaysia.

## References

[1] R. Dantu, S. Fahmy, H. Schulzrinne and J. Cangussu, "Issues and challenges in securing VoIP", Computers & Security, vol. 28, **(2009),** pp. 743-753.
[2] M. Spalding, "Deciding whether or not to use a third party certificate authority", Network Security, vol. 2000, **(2000)**, pp. 7-8.
[3] C. -H. Wang and Y. -S. Liu, "A dependable privacy protection for end-to-end VoIP via Elliptic-Curve Diffie-Hellman and dynamic key changes", Journal of Network and Computer Applications, In Press, Corrected Proof, **(2010)**.
[4] Y.-P. Liao and S.-S. Wang, "A new secure password authenticated key agreement scheme for SIP using self-certified public keys on elliptic curves", Computer Communications, vol. 33, **(2010),** pp. 372-380.
[5] E. -J. Yoon, K. -Y. Yoo, C. Kim, Y. -S. Hong, M. Jo and H. -H. Chen, "A secure and efficient SIP authentication scheme for converged VoIP networks", Computer Communications, vol. 33, **(2010)**, pp. 1674-1681.
[6] P. Zimmerman, "ZRTP: Extensions to RTP for Diffie-Hellman key agreement for SRTP, IETF Internet Draft", **(2006)**.
[7] D. Zisiadis, S. Kopsidas and L. Tassiulas, "VIPSec defined", Computer Networks, vol. 52, **(2008)**, pp. 2518-2528.
[8] S. Hoque, M. Fairhurst, G. Howells and F. Deravi, "Feasibility of generating biometric encryption keys", Electronics Letters, vol. 41, **(2005)**, pp. 309-311.
[9] D. Butcher, L. Xiangyang and G. Jinhua, "Security Challenge and Defense in VoIP Infrastructures", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 37, **(2007)**, pp. 1152-1162.
[10] C. Zhe, G. Shize, Z. Kangfeng and L. Haitao, "Research on Man-in-the-Middle Denial of Service Attack in SIP VoIP", International Conference on Networks Security, Wireless Communications and Trusted Computing, NSWCTC '09, **(2009)**, pp. 263-266.
[11] S. Prowell, R. Kraus and M. Borkin, "CHAPTER 6 - Man-in-the-Middle", in Seven Deadliest Network Attacks, Editted Boston: Syngress, **(2010)**, pp. 101-120.
[12] M. Petraschek, T. Hoeher, O. Jung, H. Hlavacs and W. Gansterer, "Security and usability aspects of man-in-the-middle attacks on ZRTP", Journal of Universal Computer Science, vol. 14, **(2008)**, pp. 673-692.
[13] R. Bresciani and A. Butterfield, "A formal security proof for the ZRTP Protocol", International Conference for Internet Technology and Secured Transactions, ICITST '09, **(2009)**, pp. 1-6.

## Authors

**Yazid Jaafar** is currently pursuing his M.Sc at School of Computer Sciences, Universiti Sains Malaysia (USM), Malaysia. He earned his B.Sc. in Computer Science from USM in 2010. His research interest includes Applied Cryptography, Network security and Parallel Computing.

**Azman Samsudin** is an Associate Professor at the School of Computer Sciences, Universiti Sains Malaysia (USM). Recently, he serves as Deputy Dean at School of Computer Sciences, USM. He earned his B.Sc. in Computer Science from University of Rochester, New York, USA, in 1989. Later, he received his M.Sc. in Computer Sciences and his Ph.D. in Computer Science, in 1993 and 1998, respectively, both from the University of Denver, Colorado, USA. He has been with Universiti Sains Malaysia since 1998. He has published articles in various professional journals and conference proceedings and has held a series of grants in the fields of Cryptography, Switching Networks and Parallel Computing.

**Alfin Syafalni** received the B.Sc. degree in Computer Science from Universiti Sains Malaysia in 2010 and currently pursuing a M.Sc. degree at the same university. In 2009, he underwent the internship program with the Malaysian Institute of Microelectronics Systems (MIMOS) working on IMS implementation and its application. His research interests are on Networking, Multimedia, and Information Security.

**Mohd Adib Omar** completed his B.Sc. (Artificial Intelligence) and M.Sc. (Computer Networks) in Computer Science from American University, Washington DC, USA in 1996 and 1997 respectively. He received his PhD in Collaborative Computing from Universiti Sains Malaysia, in 2009. He is currently a senior lecturer at School of Computer Sciences, Universiti Sains Malaysia. His research interests include Wireless Networks, Collaborative and Service Computing, Distributed, Parallel Computing and Information Security.