Integration of an RSA-2048-bit public key cryptography solution in the development of secure voice recognition processing applications

Nhu-Quynh Luc*, Duc-Huy Quach, Chi-Hung Vu, Hong-Truong Nguyen, Thanh-Long Vo-Khac

Academy of Cryptography Techniques, 141 Chien Thang Road, Tan Trieu Ward, Thanh Tri District, Hanoi, Vietnam

Received 20 October 2022; revised 20 December 2022; accepted 28 December 2022

Abstract:

The authors initially employs the fast Fourier transform (FFT) approach to transforming voice inputs into digital signals before integrating a speech recognition solution (which includes two models: the hidden Markov model (HMM) and the artificial neural network (ANN)). To achieve standard-tone identification of voice signals and digitally store speech, the authors then incorporated a 2048-bit Rivest-Shamir-Adleman (RSA) encryption method to encrypt and decrypt digital speech. The authors' building team constructed the program using a 256-bit advanced encryption standard - Galois counter mode (AES-GCM) encryption method to assure the application's effectiveness. The authors successfully created a voice recognition application according to the HMM of ANN. The collected findings suggest that the authors' secure speech recognition program (named soft voice - RSA) has improved in terms of safety, keeping speech material secret, and speed. It takes roughly 0.2 s to generate a 2048-bit RSA key pair that exceeds the National Institute of Standards and Technology (NIST) standard, 700-1070 ms to process speech, 1-4 ms to encrypt 2048-bit RSA, 6-8 ms to decrypt 2048-bit RSA.

Keywords: **artificial neural network, fast Fourier transform, hidden Markov model, Rivest-Shamir-Adleman.**

Classification number: **1.2**

1. Introduction

Speech recognition research and applications are now being investigated extensively, with practical applications in people's everyday lives. However, little research has been conducted on incorporating security measures to safeguard speech in voice recognition processing. The issue of speech recognition is a new development trend of the times, and numerous research papers on this subject have been developed and implemented in practice [1, 2]. Nevertheless, no security strategy has been put in place, putting speech recognition system users in danger of a number of things. The authors developed an application based on Markov's hidden model [3, 4], ANN [5-7], and the RSA cryptosystem [8] to handle the problem. The HMM will be integrated with ANN to handle the real-time speech recognition difficulty. To guarantee security, the RSA cryptosystem utilised for security must fulfill current NIST security criteria [9].

The authors include a 2048-bit RSA encryption technique in this research to preserve the user's voice data (voice utilises an HMM to identify speech-to-text data). The test team passed the NIST key evaluation criteria for the encryption key for the 2048-bit RSA cryptosystem. Following PKCS#1 (Public-Key Cryptography Standards-Version 2.1), an author-integrated 2048-bit RSA encryption and decryption solution assures security with today's real deployed applications. The specifics of cryptographic modules and the findings gained in this investigation are explained by the authors in the following parts of the publication.

2. Subject and methodology

2.1. Solution for converting and processing voice via FFT

Several voice processing algorithms have been used in practical applications [10, 11], most notably the audio-visual toolbox [12], the group delay [13], and the FFT. The FFT technique evolved from the discrete Fourier transform (DFT) approach, which overcomes the drawback that processing speech with a high sample length N takes a long time, and the complexity of the FFT is lowered to just N/2 log 2N. Although FFT has worse output information dependability than DFT in analogue-to-digital conversion, it is substantially faster, assuring real-time.

The authors selected FFT as the key approach in this research for converting speech to digital form and vice versa. The outputs of this FFT technique are essentially spectrum tables of the phonemes, which may be recognised on the computer using a hidden Markov-based machine learning model and an ANN [7]. ANN will investigate the possibility of any phoneme happening after another in this scenario, precisely identifying the phonemes reviewed by

*Corresponding author: Email: quynhln@actvn.edu.vn; lucnhuquynh69@gmail.com

HMM, and will then review and analyse whether the final output is accurate or not, based on a dictionary consisting of words and phrase meanings. Continue this approach until the output is quick and accurate.

2.2. Security solution for speech recognition applications

Following the conversion of voice to digital, the next phase in the study is to identify an encryption solution for digital data to assure the security of the received data against existing assaults. The authors have used the 2048-bit RSA public key cryptosystem in this work to conduct encryption and decryption for voice signal protection after converting it to digital form [14]. Furthermore, to test the speech recognition software module, the authors used an AES-256 cipher system to safeguard the voice in digital form for comparison. It follows that the efficiency in the speed of security implementation for speech communications is improved. The authors have analysed key parameter generation for the RSA cryptosystem to ensure that it meets NIST key quality testing criteria. The RSA cryptosystem is included in a software module that conforms to the PKCS#1 [15] standard and is certified to be safe against existing threats.

According to NIST standards, the RSA cryptosystem is considered safe with a length of nLen=1661 by 2022. At the time, the 2048-bit RSA cryptosystem was still guaranteed to be safe. The authors evaluate generating key parameters for use in the program for the AES 256 cryptosystem to surpass current NIST criteria. This demonstrates that the authors' method, which employs a 2048-bit RSA cryptosystem (or AES 256-bit), is sufficient to safeguard the speech signal against various assaults.

3. Results

3.1. Design and implement a secure voice recognition application

In this research, the secure voice recognition module (Soft voice-RSA) built by the authors incorporates the following modules: 2048-bit RSA key generation (according to PKCS#1 v2.1 standard) and has been tested to pass through NIST standards; the module converts speech into text and vice versa using the FFT method that applies HMM and ANN models; text-to-speech encryption module with RSA 2048 cryptography and software-generated public key; 2048-bit RSA decryption module with software generated RSA private key. Fig. 1 depicts the operational concept of 2048-bit RSA key generation, voice conversion, and protection utilising a 2048-bit RSA cryptosystem.

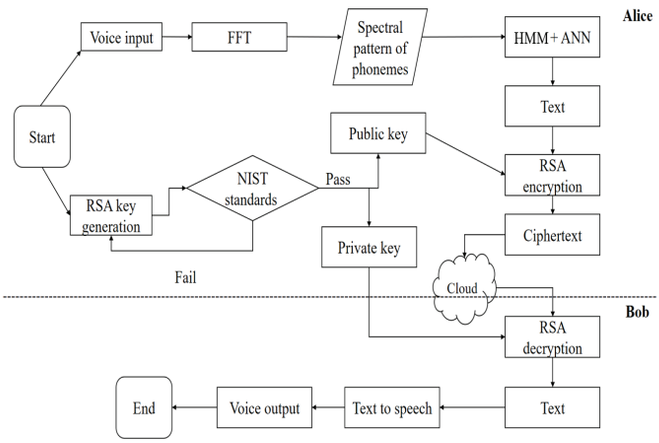

Fig. 1. Model of operation, conversion, and protection of voice using the RSA cryptosystem.

First, the speech recorded from the microphone is turned into signals that are transferred within the computer and processed into discrete samples. In the computer, the aforesaid samples travel via the HMM and ANN to compare and produce the standard text form matching the speech. Next, the resulting text is encrypted using the 2048-bit RSA cryptosystem. Then, transmit the freshly produced ciphertext to the recipient. At this moment, the receiver decodes the ciphertext and retrieves the plaintext, matching the original voice content. The text is transformed back to speech using a normal voice set.

3.2. Interface design for the soft voice-RSA module

In this research, the program module is created with two primary interfaces: Fig. 2A is a design detail for the implementation of 2048-bit RSA key generation (PKCS #1 v2.1 standard) to ensure that the key generated meets the NIST key quality assessment standard; Fig. 2B is an interface for performing speech-to-text and vice versa, as well as encrypting/decrypting speech signals in text using 2048-bit RSA cryptography.

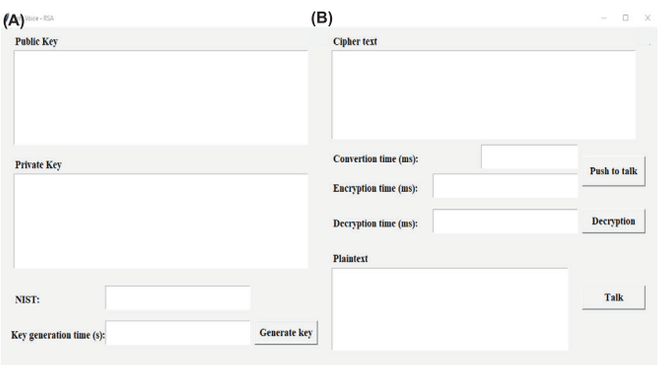

Fig. 2. (A) RSA key modulus generator interface; (B) Modular interface encrypts and decrypts voice using RSA cryptosystem.

3.2.1. The interface for key generation: To use the program, the user needs to prepare the private key and the public key of the RSA cryptosystem or use the key generated by the program with a minimum modulo length of 2048 bits by clicking the button "Generate key". After clicking, the machine will automatically generate a set of 2048-bit RSA keys. These keys will be checked through the NIST standard set. If they meet the standards, they will be printed on the screen and displayed throughout the entire processing time to produce the key.

3.2.2. The voice converter interface is secure: Once the needed security key is accessible, the user delivers his message via the microphone attached to the computer by clicking the "Push to talk" button. After pressing, the application will automatically record the portion of the voice we emit and stop when we finish talking. The analogue signal collected here from vibrations in the microphone is transformed into electrical impulses in the computer. Through the FFT transformation, electrical impulses are turned into spectral patterns. These spectral samples are processed using a language model (a mix of HMM and ANN) against the available sample sets to create the original spoken material in the form of text. The text here will be encrypted using the RSA cryptosystem using the public key created or prepared above. The ciphertext may then be communicated to the desired audience. When processing is complete, the ciphertext component will be displayed on the screen along with the conversion and encoding times. Then we hit "Decryption" to read the newly decoded segment and output the contents to the screen. The "Talk" button is used to play back the translated text.

4. Discussion

4.1. Analysis, assessment, and testing of the soft voice-RSA protected speech recognition module

In this research, to assess the operation of the soft voice-RSA software module, the authors did it on a computer with a configuration utilising Intel(R) Core i5-4200U, CPU @1.60GHz, up to 2.30 GHz; RAM: 8.00 GB. Fig. 3A displays the uptime results of the soft voice-RSA program module with 2048-bit RSA key generation that exceeds key assessment requirements. Fig. 3B displays the real-time results of speech conversion and voice encryption/decryption using 2048-bit RSA cryptography. Fig. 3C shows the execution time result of the soft voice program module (secure voice recognition with 256-bit AES-GCM encryption) (secure speech recognition with 256-bit AES-GCM cipher). To assess the performance, the process of creating 2048-bit RSA public and private key pairs; speech recognition by HMM and ANN models; execution time for speech conversion and encryption/decryption using 2048-bit RSA and AES-256-bit ciphers. The authors have run the program numerous times with varied input data for the software to run. Table 1 illustrates the results of running the soft voice-RSA program with varied input data and matching to the produced key set that is guaranteed to pass the key quality evaluation criteria of the NIST.

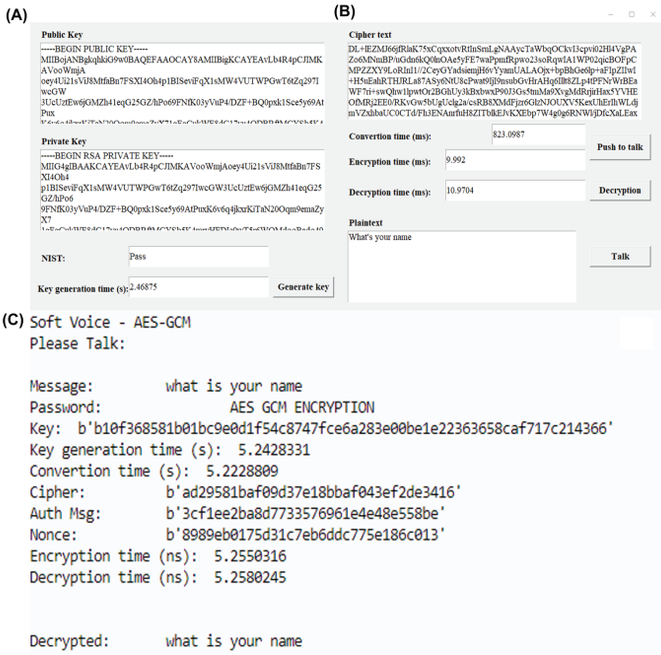

Fig. 3. Soft voice-RSA execution-time results. (A) NIST pass test and key generation; **(B)** Voice reception/conversion, speech encryption, and decryption using a 2048-bit RSA cryptosystem; **(C)** Voice capture/conversion, speech encryption, and decryption using a 256-bit AES-GCM cryptosystem.

Table 1. Execution time findings of soft voice-RSA and soft voice-AES-GCM.

The number of characters	Processing time (ms)	Encryption time (ms)	Decryption time (ms)	Key generation time (s)
Speech recognition using a 2048-bit RSA cryptosystem				
12	746.7266	3.991	7.0148	0.187500
35	1063.2608	1.9937	6.9827	0.328125
37	683.639	0.9993	5.9748	1.781250
40	830.3528	2.0288	7.9833	0.625000
Speech recognition uses 256-bit AES-GCM cryptography				
21	0.6907975	1.3045635	1.3075628	1.3005715
27	0.8660744	1.4963576	1.4993501	1.4933993
35	0.9356286	1.55198	1.5549542	1.5479727

The results show that the execution time of the soft voice-RSA seed when using the 2048-bit RSA cryptosystem is as follows: The process of generating a 2048-bit RSA key pair surpassing NIST's standards takes about 0.2-2 s; speech processing time ranges from 700-1070 ms; 2048-bit RSA encryption time takes between 1-4 ms; and the 2048-bit RSA decryption time takes about 6-8 ms. The execution time of the soft voice-RSA seed while utilising the 256-bit AES-GCM cryptosystem is as follows: the key generation process takes around 1-2 s; speech processing time is about

0.7-1 s; AES-GCM 256-bit encryption time takes about 1.5 s; decoding time takes about 1.6 s.

This enables the authors to notice that while increasing the length of the input, the processing pace varies differently. Because when we say it, while it may be a sentence of the same length, the pronunciation of phrases in it is different. Many clusters will grasp it, but some clusters have to be analysed into sections. So, while the length varies, the processing speed is often many times faster, and the encryption and decryption times are always in the range of 1-4 ms (for RSA 2048 encryption implementation. bit) and 6-8 s (for RSA 2048 decryption time). thus demonstrating that, even with a very large input processing, the software can process encoding and decoding in the time required to meet the user's demands.

In addition to the 2048-bit RSA technique, while constructing a secure speech recognition application, it is also feasible to substitute the security solution with any other cryptosystem, depending on the needs when utilising it. contemporary reality. Here, the authors have replaced the RSA 2048 cipher generation with the 256-bit AES-GCM cipher to prove that several other ciphers may be used to safely handle the speech signal after it has been translated, digital speech signal. Experimental findings reveal that the time for voice processing is identical to the above since both parties use the same processing technique, but the time of key creation, encryption, and decryption is slower if using AES-GCM 256 and faster with RSA. This demonstrates that the selected solution employing the 2048-bit RSA cryptosystem as a security solution for voice signals in digital form is suitable for the authors' application.

4.2. Testing the source code of soft voice-RSA software

The authors utilised the Fortify Static Code Analyzer toolkit (Version 22.1.0.0166) to examine and assess the source code of the soft voice-RSA software. Table 2 presents thorough findings while doing testing, assessment, and analysis of the source code of the soft voice-RSA program. The findings reveal that soft voice-RSA software has no faults during development. Here, the authors primarily discuss functions used in programming. In the software application soft voice-RSA, the authors have not constructed a protective solution for the process of producing keys and storing public keys and private keys of the 2048-bit RSA cryptosystem in compliance with key management and storage requirements of NIST [16-18]. These are also the solutions the authors' team hopes to enhance in future investigations.

Table 2. Soft voice-RSA software source code testing results using Fortify Static Code Analyzer toolkit.

Category	Fortify priority (audited/total)				Total issues
	Critical	*High*	*Medium*	*Low*	
Buffer overflow	0	0	0	0	0
Poor style: variable never used	0	0	0	0	0
Type mismatch: integer to character	0	0	0	0	0
Type mismatch: signed to unsigned	0	0	0	0	0
Unchecked return value	0	0	0	0	0
Weak cryptographic hash	0	0	0	0	0

Thus, via analysing, assessing, and testing the flaws in soft voice-RSA software using the Fortify Static Code Analyzer toolkit (Version 22.1.0.0166), it indicates that soft voice-RSA software authors established development and construction have also assured the safety of the source code. This is enough to prove that soft voice-RSA software can assure security against some of today's assaults when applied to these real-world commercial applications used by people every day.

5. Conclusions

The findings gained in this research have created a voice recognition application according to the HMM of ANN. In particular, the program has incorporated an RSA public key cryptography solution (according to PKCS#1 Version 2.1 standard) to secure the secrecy of speech material in digital form after identification. The execution speed of soft voice-RSA programs accomplishing encryption and decryption times is always between 1-4 ms (for 2048-bit RSA encryption implementation) and 6-8 s (for decryption time). These findings have been improved and have been compared with the solution while using the AES-GCM 256 cipher. Experimental findings reveal that the speech processing time is identical to that above because both sides utilise the same processing, but the time of key creation, encryption, and decryption for AES-GCM is several times slower than RSA. The authors concluded that the solution in this research also has certain drawbacks, such as the restricted voice recognition capabilities. This is also the research direction that the authors will concentrate on and will report on in the upcoming investigations.

CRediT author statement

Nhu-Quynh Luc: Conceptualisation, Methodology, Software, Resources, Writing - Review and Editing; Duc-Huy Quach: Data curation, Software, Writing - Original draft preparation; Chi-Hung Vu: Resources, Visualisation, Investigation; Hong-Truong Nguyen: Supervision; Thanh-Long Vo-Khac: Supervision, Software, Validation.

ACKNOWLEDGEMENTS

The authors thank the Academy of Cryptography Techniques for supporting this work.

COMPETING INTERESTS

The authors declare that there is no conflict of interest regarding the publication of this article.

REFERENCES

[1] N. Das, S. Chakraborty, J. Chaki, et al. (2021), "Fundamentals, present and future perspectives of speech enhancement", *Int. J. Speech Technol.*, **24(4)**, pp.883-901, DOI: 10.1007/s10772-020-09674-2.

[2] X. Han, Z. Zhang, N. Ding, et al. (2021), "Pre-trained models: Past, present and future", *AI Open*, **2**, pp.225-250, DOI: 10.1016/j.aiopen.2021.08.002.

[3] G.A. Fink (2008), *Markov Models For Pattern Recognition: From Theory to Applications*, Springer Berlin Heidelberg, 275pp, DOI: 10.1007/978-3-540-71770-6.

[4] Z. Han, Q. He, M.V. Davier (2019), "Predictive feature generation and selection using process data from PISA interactive problem-solving items: An application of random forests", *Front. Psychol.*, **10**, DOI: 10.3389/fpsyg.2019.02461.

[5] I. Farkaš, P. Masulli, S. Wermter (2020), *Artificial Neural Processing of the Networks and Machine Learning - ICANN 2020*, Springer, 918pp.

[6] G.R. Yang, X.J. Wang (2021), "Artificial neural networks for neuroscientists: A primer", *Neuron*, **107(6)**, pp.1048-1070, DOI: 10.1016/j.neuron.2020.09.005.

[7] R. Dastres, M. Soori (2021), "Artificial neural network systems", *Int. J. Imaging Robot.*, **21(2)**, pp.13-25.

[8] N. Bansal, S. Singh (2020), "RSA encryption and decryption system", *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, **6(5)**, pp.109-113, DOI: 10.32628/CSEIT206520.

[9] E. Barker (2020), *Guideline for Using Cryptographic Standards in The Federal Government*, National Institute of Standards and Technology Special Publication 800-175B Revision 1, 91pp.

[10] F. Ernawan, N.A. Abu, N. Suryana (2011), "Spectrum analysis of speech recognition via discrete Tchebichef transform", *Proceedings of SPIE - The International Society for Optical Engineering*, **8285**, DOI: 10.1117/12.913491.

[11] S. Sadhu, H. Hermansky (2021), "Radically old way of computing spectra: Applications in end-to-end ASR", *Proc. Interspeech 2021*, pp.1424-1428, DOI: 10.21437/Interspeech.2021-643.

[12] A. Abel, A. Hussain (2009), "Multi-modal speech processing methods: An overview and future research directions using a MATLAB based audio-visual toolbox", *Multimodal Signals: Cognitive and Algorithmic Issues*, **1177**, pp.121-129, DOI: 10.1007/978-3-642-00525-1_12.

[13] T. Drugman, T. Dubuisson, T. Dutoit (2011), "Phase-based information for voice pathology detection", *2011 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Prague, Czech Republic, pp.4612-4615, DOI: 10.1109/ICASSP.2011.5947382.

[14] RSA Laboratories (2002), *PKCS #1 v2.1: RSA Cryptography Standard*, 61pp.

[15] G. McGraw (2006), "Software security: Building security in", *Proceedings of The 17th International Symposium on Software Reliability Engineering*, DOI: 10.1109/ISSRE.2006.43.

[16] A. Apvrille, M. Pourzandi (2005), "Secure software development by example", *IEEE Secur. Priv. Mag.*, **3(4)**, pp.10-17, DOI: 10.1109/MSP.2005.103.

[17] J. Koziol, D. Litchfield, D. Aitel, et al. (2004), *The Shellcoder's Handbook: Discovering and Exploiting Security Holes*, 2nd Edition, Wiley Publishing, 744pp.

[18] M. Howard, D. Leblanc (2008), *Writing Secure Code*, 2nd Edition, Microsoft Press, 737pp.