Secured Audio Encryption using AES Algorithm

Aishwarya Agarwal
Department of Information Technology
Galgotias College of Engineering & Technology
Greater Noida

Pratibha Raj Singh
Department of Information Technology
Galgotias College of Engineering & Technology
Greater Noida

Sandhya Katiyar, PhD
Department of Information Technology
Galgotias College of Engineering & Technology
Greater Noida

ABSTRACT
Speech to text conversion is the process of converting spoken words into written texts. In this paper, a voice encryption system is developed as a real-time software application. Basically, the speech is taken as an input and is encoded to be decoded by authenticated users only. The algorithm used to perform this cryptography is Advanced Encryption Standards (AES). This algorithm has its own particular structure to encrypt and decrypt sensitive data and is applied in hardware and software all over the world. We take the information in the form of audio as an input using a microphone which is to be transmitted over the channel to the intended receiver. The audio input is converted into text format which ensures speech-to-text conversion. Then, the text is encrypted using the AES algorithm to form cipher text. This cipher text is sent over a channel to the receiver. The receiver requests to perform decryption of the information only if he has the correct secret key otherwise the request is declined. If the key matches, the decryption is successful and the receiver get the message as text. For 128 bit, about 2128 attempts are needed to break. This makes it very difficult to hack it as a result it is very safe protocol.

General Terms
Cryptography, AES algorithm, Encryption/Decryption Speech System.

Keywords
Speech-to-text (STT)

1. INTRODUCTION
Speech is one of the most important and basic tool for the communication between humans and his environment. The environment may include computers, mobile phones, etc. The human computer interaction is termed as human computer interface. In order to establish such an interface, keyboard and mouse forms the most common method for interaction. When the amount of data to be entered is large, then these devices become time consuming. For an efficient communication to take place, we tend to change the method of interaction. According to human beings, the best way of communication between them is speech. If a system can understand what a human speaks, then it is the best method of interaction between a human and a computer. Speech control or widely known as Speech Recognition is the method to control something by human voices/speech. Speech recognition technology is one of the fastest growing engineering technologies. It has a number of applications in different areas and provides potential benefits. Nearly 20% people of the world are suffering from various disabilities; many of them are blind or unable to use their hands effectively. The speech recognition systems in those particular cases provide a significant help to them, so that they can share information with people by operating computer through voice input.In order to share some confidential data between people, secure communication must be ensured. This could be brought into picture by introducing encryption and decryption of the message to be delivered which in audio format i.e. speech. Encryption is considered one component of a successful security strategy. Successful encryption completely depends on robust passwords and pass phrases called "keys". The basic flow diagram for encryption and decryption system is as shown in fig 1.

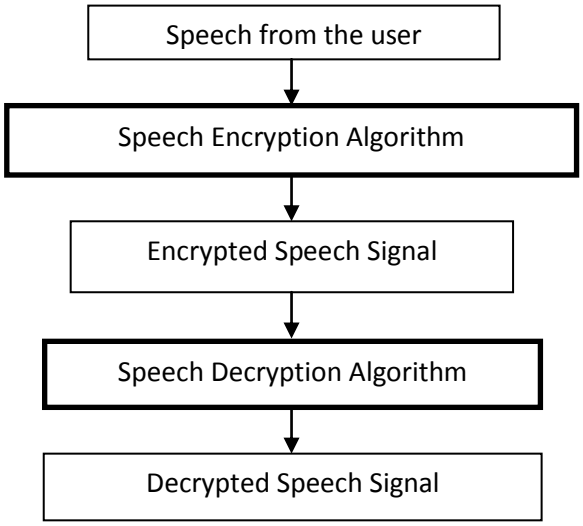

Fig 1: Implementation of Encryption/Decryption Speech System

2. TYPES OF SPEECH
Speech recognition system can be categorized into different classes based on the type of utterances they can identify:

2.1 Isolated Word: Isolated word recognizers usually require each utterance to have quiet on both side of sample windows. It accepts individual words or single utterances at a time. It has a Listen and Non Listen state.

2.2 Connected Word: Connected word system are similar to isolated words but allow to isolate sound to be run together with minimum pause between them.

2.3 Continuous speech: Continuous speech recognizers permits user to talk almost naturally, while the computer regulates the content. Recognizer with continues speech capabilities are one of the most difficult to create since they utilize distinctive sound and precise method to form utterance boundaries.

2.4 Spontaneous speech: At an initial level, it can be thought of as speech that is natural sounding and not rehearsed. An

Automatic Speech Recognition System with spontaneous speech ability should be able to manage different words and variety of natural speech feature such as words being run together.

3. SPEECH TO TEXT SYSTEM

Speech is an extraordinary interesting mode to bring about human computer interface: it is "hands free"; it demands only modest hardware for accession i.e. a good quality microphone or microphones; and it arrives at a very moderate bit rate. It is hard to recognize human speech, especially connected speech for a vocabulary of sufficient complexity without undergoing any burdensome training. However, now the processing of speech signals has become much easier and with that we are also able to recognize the text which is talking by the talker due to the emergence of modern processes, algorithms, flow diagrams, and methods. In this system, we are going to develop an internet operated speech-to-text engine. The system accesses speech at run time through a microphone and works on the sampled speech to recognize the uttered word/words. The recognized set of words can be stored in a text file. It can add other larger systems, giving users a different choice for data entry. Visually impaired, hard of hearing, or physically debilitated clients tend to get most benefitted by this as a speech-to-text framework can likewise enhance framework accessibility by giving information passage alternatives to them. An application that grew in this work is voice SMS that permits a client to record and believer voice messages into SMS instant message. Client can send voice messages to the telephone number of the person he/she wishes to talk to. Speech identification is done via the Internet, connecting to Google's server. The language which is used to input messages in this application is English. The technique used for speech identification is based on hidden Markov models (HMM – Hidden Markov Model). It is currently the most successful and most flexible approach to speech recognition.

An Automatic Speech Recognition System with spontaneous speech ability should be able to manage different words and variety of natural speech feature such as words being run together.

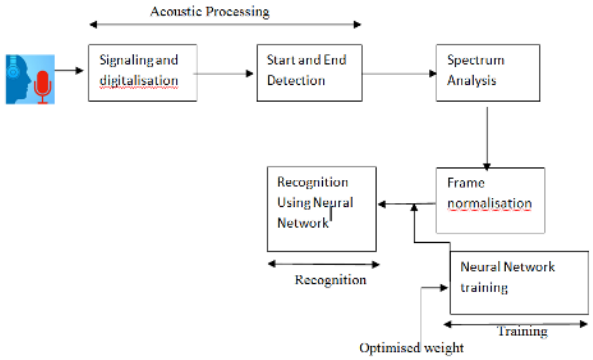

Fig 2: Speech-to-text Model

4. APPLICATIONS OF SPEECH TO TEXT SYSTEM

As our lives and jobs become more centralized around the use of a computer, typing becomes the everyday normal for basic communication, documentation and memorization. But as humans, we aren't meant to sit around typing out our every thought; it can be inefficient, sound commonplace, and be unhealthy for our bodies and minds at a certain point. The application field of speech-to-text system is spreading fast whilst the quality of the systems is also rising at a great speed. Dealing with speech system has become much affordable nowadays making it to be used every day by the customers at an effective cost. Some of the applications of speech-to-text system are:

1. Ease of communication- No more illegible handwriting

With advancements in technology, and the use of speech-to-text services and software, anyone can experience the benefits of cutting back on typing. Speaking directly into an application that can transcribe your words into text for you will drastically improve the accuracy of your documents.For a normal typist, about 8 out of 100 words will be spelled incorrectly and have to be fixed. By cutting out the time you'll need to spend making corrections, you'll be free to focus on the task at hand.

2. Aid for Visually Impaired people

Blind individuals are most benefitted by this system as they can utilize the system in the most effective way possible. They listen or hear to gain information quickly and efficiently from the books on tape, CD, etc. They assess what is happening around them by their sense of hearing.

3. Games and Education

A speech-to-text system can make tedious jobs streamlined and simplified. In the field of study and sports, synthesized speech can be used. The performance, accuracy and efficiency of the speech recognizer remains intact no matter what. It also provides an accurate and reliable method of documentation.

4. Telecommunication and Multimedia

Vocal information can be accessed over the telephone with the help of speech-to-text systems. User's voice or the telephone keyboard is used to process queries to such information retrieval systems with the help of speech recognizers. Short text messages could be spoken in a mobile phone with the help of synthesized speech.

5. Man-Machine Interaction

Human machine interaction or interface is made possible to a great extent by using speech recognition. For example, synthesized speech may be used to give close to accurate information of the ongoing situation in cautionary situations such as clocks, washing machines, alarm clocks. Speech signals are considered far better than buzzers and warning lights as it facilitates to react to the signal faster if the person is unable to get light due some obstacles.

6. Voice Enabled E-mail

Voice-enabled e-mail uses voice recognition and speech synthesis technologies to enable users to access their email from any telephone. The subscriber dials a phone number to access a voice portal, then, to collect their email messages, they press a couple of keys and, perhaps, say a phrase like "Get my e-mail." Speech synthesis software converts e-mail text to a voice message, which is played back over the phone. Voice-enabled e-mail is especially useful for mobile workers, because it makes it possible for them to access their messages easily from virtually anywhere (as long as they can get to a phone), without having to invest in expensive equipment such as laptop computers or personal digital assistants.

5. AES ALGORITHM

The Advanced Encryption Standard (AES), also known by its original name Rijndael, is a specification for the encryption of electronic data established by the U.S. National Institute of Standards and Technology (NIST) in 2001. It supersedes the Data Encyption Standard (DES), which was published in 1977.

The more popular and widely adopted symmetric encryption algorithm likely to be encountered nowadays is the Advanced Encryption Standard (AES). It is found at least six time faster than triple DES. It is considered more reliable than other encrypting algorithms when using a large length of secret key.

A replacement for DES was needed as its key size was too small. With increasing computing power, it was considered vulnerable against exhaustive key search attack. Triple DES was designed to overcome this drawback but it was found slow.

The AES algorithm consists of four stages that make up a round which is iterated 10 times for a 128-bit length key, 12 times for a 192-bit key, and 14 times for a 256-bit key. It is an efficient algorithm. It is usually unlikely to crack this algorithm.

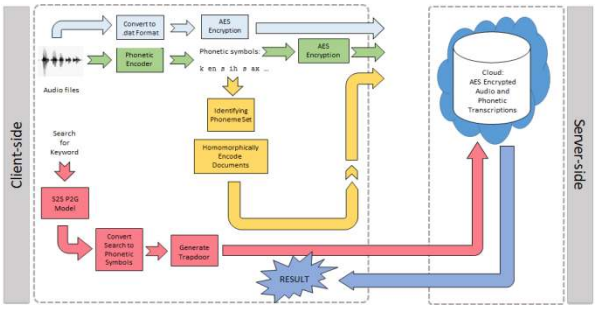

Fig 3: Basic idea of AES Algorithm

Stage 1: "SubBytes" transformation is a non-linear byte substitution for each byte of the block.

Stage 2: "ShiftRows" transformation cyclically shifts (permutes) the bytes within the block

Stage 3: "MixColumns" transformation groups 4-bytes together forming 4-term polynomials and multiplies the polynomials with a fixed polynomial mod (x^4+1).

Stage 4: "AddRoundKey" transformation adds the round key with the block of data.

In most ciphers, the iterated transform (or round) usually has a Feistel Structure. Typically in this structure, some of the bits of the intermediate state are transposed unchanged to another position (permutation). AES does not have a Feistel structure but is composed of three distinct invertible transforms based on the Wide Trial Strategy design method.

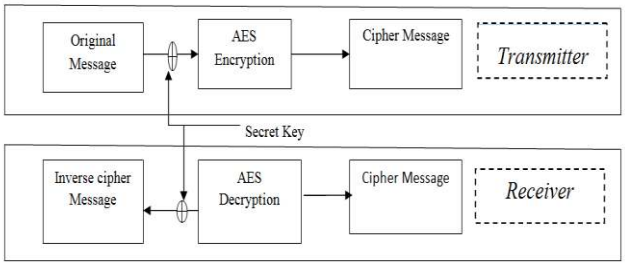

Fig 4: Work Model of AES Algorithm

5.1 Generating a Key

AES encryption needs a strong key. The stronger the key, the stronger your encryption. This is probably the weakest link in the chain. By strong, we mean not easily guessed and has sufficient entropy (or secure randomness).

That being said, for the sake of demonstration of AES encryption, we generate a random key using a rather simple scheme.

key = ''.join(chr(random.randint(0, 0xFF)) for i in range(16)) print 'key', [x for x in key] # prints key ['+', 'Y', '\xd1', '\x9d', '\xa0', '\xb5', '\x02', '\xbf', ';', '\x15', '\xef', '\xd5', '}', '\t', ']', '9']

5.2 Initializing Vector

In addition to the key, AES also needs an initialization vector. This initialization vector is generated with every encryption, and its purpose is to produce different encrypted data so that an attacker cannot use cryptanalysis to infer key data or message data.

A 16-byte initialization vector is required which is generated as follows.

iv = ''.join([chr(random.randint(0, 0xFF)) for i in range(16)])

The initialization vector must be transmitted to the receiver for proper decryption, but it need not be kept secret. It is packed into the output file at the beginning (after 8 bytes of the original file size), so the receiver can read it before decrypting the actual data.

5.3 Encrypted with AES

We now create the AES cipher and use it for encrypting a string (or a set of bytes; the data need not be text only).

The AES cipher is created with CBC Mode wherein each block is "chained" to the previous block in the stream. (You do not need to know the exact details unless you are interested. All you need to know is – use CBC mode).

Also, for AES encryption using pycrypto, you need to ensure that the data is a multiple of 16-bytes in length. Pad the buffer if it is not and include the size of the data at the beginning of the output, so the receiver can decrypt properly.

```
aes = AES.new(key, AES.MODE_CBC, iv)

data = 'hello world 1234' # <- 16 bytes

encd = aes.encrypt(data)
```

5.4 Decryption with AES

Decryption requires the key that the data was encrypted with. You need to send the key to the receiver using a secure channel (not covered here).

In addition to the key, the receiver also needs the initialization vector. This can be communicated as plain text, no need for encryption here. One way to send this is to include it in the encrypted file, at the start, in plaintext form. We demonstrate this technique below (under *File Encryption with AES*). For now, we assume that the IV is available.And that is how simple it is. Now read on to know how to encrypt files properly.

```
1 aes = AES.new(key, AES.MODE_CBC, iv)

2 decd = adec.decrypt(encd)

3 print decd

4 # prints

5 hello world 1234
```

5.5 Save the initialization vector

As explained above, the receiver needs the initialization vector. Write the initialization vector to the output, again in clear text.

```
Fout.write(iv)
```

6. RESULT ANALYSIS

The actual output is the secured audio which is achieved by using the AES Algorithm. With the help of AES Algorithm, we were successful to encrypt and decrypt the audio signal which was converted to text format before cryptography implementation.

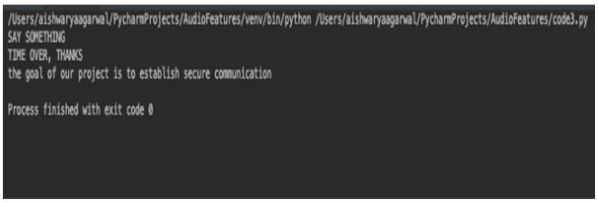

Fig 5: Input Audio using a Microphone

The spoken audio is converted into text file format from where it will be encrypted and decrypted.

Encryption requires a key with which a cipher text is created. Cipher text is the encoded version of the original text.

Decrytion requires the key that the data was encrypted with. You need to send the key to the receiver using a secure channel (not covered here).

Fig 6: Encrypted Audio Data

Hence the image shows the encrypted form of the speech, with the matched private key we found this encrypted form of the speech and hence successful form of encrypted speech is successfully done.

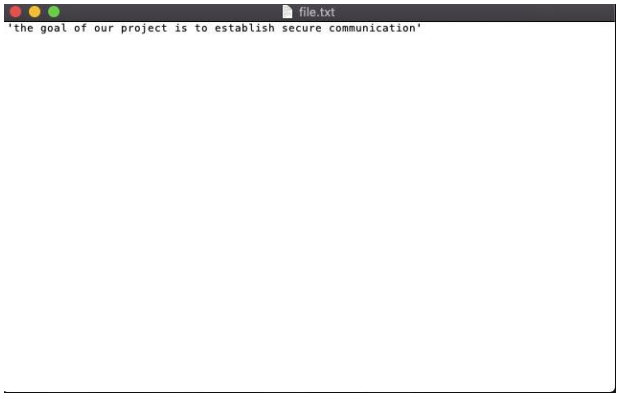

Fig 7: Decrypted Audio Data

Hence the image shows the decrypted form of the speech, with the matched private key we found this decrypted form of the speech and hence successful form of decrypted speech is successfully done.

7. CONCLUSION AND FUTURE SCOPE

In this paper, we discussed the topics relevant to the development of STT systems. The speech to text conversion may seem effective and efficient to its users if it produces natural speech and by making several modifications to it.. Speech to Text synthesis is a critical research and application area in the field of multimedia interfaces. In this paper gathers important references to literature related to the endogenous variations of the speech signal and their importance in automatic speech recognition. A database has been created from the various domain words and syllables. The desired speech is produced by the Concatenative speech synthesis approach. Speech synthesis is advantageous for people who are visually handicapped. This paper made a clear and simple overview of working of speech to text system (STT) in step by step process. The system gives the input data from mice in the form of voice, then preprocessed that data & converted into text format displayed on PC. The user types the input string. We have successfully converted the encrypted speech to decrypted speech with the help of the shared private key. And hence the converted text can be read.

Therefore, future scope of work is:

1. To test the proposed model for checking it against its fault tolerant power.
2. The QOS (quality of service) of proposed model may be determined in terms of some availability, throughput and delay.
3. Voice Mail Encryption.
4. To Secure Voice Chat.

8. ACKNOWLEDGEMENTS

We would like to express our gratitude to Dr. S. K. Singh (Head of Department), Dr. Sandhya Katiyar (Deputy Head of Department) and our coordinator Mr. Sanjay Khakhil who have supported us in every phase of the research and have guided us to yield better results.

9. REFERENCES

[1] Sanjib Das, "Speech Recognition Technique: A Review", International Journal of Engineering Research and Applications" (IJERA) ISSN: 2248-9622 Vol. 2, Issue 3, May-Jun 2012.

[2] Ms. Sneha K. Upadhyay,Mr. Vijay N. Chavda, "Intelligent system based on speech recognition with capability of self-learning", International Journal For Technological Research In Engineering ISSN (Online): 2347 - 4718 Volume 1, Issue 9, May-2014.

[3] Deepa V.Jose, Alfateh Mustafa, Sharan R, "A Novel Model for Speech to Text Conversion" International Refereed Journal of Engineering and Science (IRJES)ISSN (Online) 2319-183X, Volume 3, Issue 1 (January 2014).

[4] B. Raghavendhar Reddy, E. Mahender, "Speech to Text Conversion using Android Platform", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622,Vol. 3, Issue 1, January - February 2013.

[5] Kaveri Kamble, Ramesh Kagalkar, "A Review: Translation of Text to Speech Conversion for Hindi Language", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.Volume 3 Issue 11, November 2014.

[6] Santosh K.Gaikwad,Bharti W.Gawali,Pravin Yannawar, "A Review on Speech Recognition Technique",International Journal of Computer Applications (0975 – 8887)Volume 10– No.3, November 2010.

[7] Penagarikano, M. Bordel, G., "Speech-to-text translation by a non-word lexical unit based system,"Signal Processing and Its Applications, 1999. ISSPA '99. Proceedings of the Fifth International Symposium on, vol.1, no., pp.111,114 vol.1, 1999

[8] Olabe, J. C.; Santos, A.; Martinez, R.; Munoz, E.; Martinez, M.; Quilis, A.; Bernstein, J., "Real time text-to-speech conversion system for spanish," Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84. vol.9, no., pp.85,87, Mar 1984.

[9] Kavaler, R. et al., "A Dynamic Time Warp Integrated Circuit for a 1000-Word Recognition System", IEEE Journal of Solid-State Circuits, vol SC-22, NO 1, February 1987, pp 3-14.

[10] Aggarwal, R. K. and Dave, M., "Acoustic modelling problem for automatic speech recognition system: advances and refinements (Part II)", International Journal of Speech Technology (2011) 14:309–320.

[11] Ostendorf, M., Digalakis, V., & Kimball, O. A. (1996). "From HMM's to segment models: a unified view of stochastic modeling for speech recognition". IEEE Transactions on Speech and Audio Processing, 4(5), 360– 378.

[12] Yasuhisa Fujii, Y., Yamamoto, K., Nakagawa, S., "AUTOMATIC SPEECH RECOGNITION USING HIDDEN CONDITIONAL NEURAL FIELDS", ICASSP 2011: P-5036-5039.

[13] Mohamed, A. R., Dahl, G. E., and Hinton, G., "Acoustic Modelling using Deep Belief Networks", submitted to IEEE TRANS. On audio, speech, and language processing, 2010. [14] Sorensen, J., and Allauzen, C., "Unary data structures for Language Models", INTERSPEECH 2011.

[14] Sura F. Yousif. Encryption and Decryption of Audio Signal based on RSA Algorithm. Vol 5, July 2018.

[15] Abhishek Kumar Sinha, IIJayaraj N "Encryption and Decryption Using AES in the Field of Network Communication based on Confidentiality", Jan-March, 2015.

[16] Edward C. Lin, Kai Yu, Rob A. Rutenbar and Tsuhan Chen "A 1000-Word Vocabulary, Speaker-Independent, Continuous Live-Mode Speech Recognizer Implemented in a Single FPGA", 2007.

[17] Santosh K. Gaikwad, Bharti W. Gawali and Pravin Yannawar "A Review on Speech Recognition Technique", 2010 submitted to International Journal of Computer Application.

[18] Muhammad Faheem Mushtaq, Sapiee Jamel, Abdulkadir Hassan Disina, Zahraddeen A. Pindar, Nur Shafinaz Ahmad Shakir, Mustafa Mat Deris "A Survey on the Cryptographic Encryption Algorithms" published in IJACSA (International Journal of Advanced Computer Science and Applications) in 2017.