Improving Security of Voice Commerce in Smart Speakers

MSc Research Project
Cybersecurity

Devanshu Kaushik
Student ID: 20233132

School of Computing
National College of Ireland

Supervisor: Michael Prior

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Devanshu Kaushik

Student ID: 20233132

Programme: M.Sc. Cybersecurity **Year:** 2021-2022

Module: Research Project

Supervisor: Michael Prior

Submission Due Date: 15th August 2022

Project Title: Improving Security of Voice Commerce in Smart Speakers

Word Count: 5022………… **Page Count** 17

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Devanshu Kaushik

Date: 15 August 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Improving Security of Voice Commerce in Smart Speakers

Devanshu Kaushik
20233132

Abstract

Voice commerce is a growing technology where voice is used for purchasing something online. This technology works best with voice assistants and is therefore, prevalent in smart speakers.

Voice assistants store the crucial information like shipping address and payment details, in order to provide a convenient and fast method for shopping. The authentication is done by voice recognition to grant permission for carrying out the purchase. This voice recognition system has to be accurate to prevent any accidental or unauthorised purchase.

Speech and voice recognition systems are developed to understand the underlying technology of voice recognition. A structured test was conducted on Amazon's Alexa powered echo device to find any flaws in the authentication process.

This research will prove to be a foundation for a secure voice commerce via smart speakers by providing a better understanding of the authentication process as well as testing the voice recognition system implemented by Alexa.

1 Introduction

Smart homes, once considered as a luxury, have started to become a necessity in fast paced environment of a city life. Scheduling a timer, looking at a recipe, changing the light colour, as well as cleaning pet's hair from the floor can be all done using smart devices in a home. There are smart locks available which can be used by simply connecting them to your mobile device, which unlocks the door when a person moves closer to the door. Virtual assistant powered smart speakers work like a gateway to the other smart devices as they are capable to accept a voice command from a user and then send the signal to the smart device intended.

The smart speakers come with a virtual assistant which can be given a wake-up command in order to start an interaction with them. "Hi, Alexa" is used to wake up the Alexa devices by Amazon. "Hey, Siri" and "Ok, Google" are other popular wake-up words for the virtual assistants present in iOS and Google devices respectively. A virtual assistant is capable of handling a simple task, like answering "How is the weather today?", to performing a complicated task like "Book a ticket of Doctor Strange for 8 p.m. show on 15th of August at Cineworld Parnell, Dublin", provided that Cineworld has developed a skill for these virtual assistants.

One of the most important features of these virtual assistants is to introduce voice-based shopping to the customers of the devices. It is an ability to purchase something using just the voice. Online shopping experience is optimised by these smart speakers as they provide a way to buy something without going through the hassle of entering shipping address or payment details. The owner's account stores the shipping information and payment method is connected to the account as well.

This is the easiest and most convenient way to shop but it has brought much more inconveniences to many. It was found that there is a need for some kind of authentication when four pounds of cookies and a dollhouse worth $170 was accidentally ordered by a 6-year-old girl (K. Schooler, 2017). Christmas toys worth $700 were purchased by kids in a similar incident (Bild, 2019). The smart speakers only required a command from a speaker and no authentication measure was in place to authenticate the command. An African grey parrot was able to order his favourite snacks (NDTV, 2018) and a 3-year-old toddler successfully purchased a fart extension pack worth $2.6 (Moyes and Goss, 2021).

Amazon's Alexa is related to most of the incidents which involve such accidental purchases. Amazon's Alexa device is considered for this research as the market share of Amazon in US is 69% and globally it is around 26.4% in smart speakers (Bishop, 2021; Laricchia, 2022).

There are several measures taken by Amazon to ensure that such accidental purchases are not made. Despite these measures only 11.5% of the smart speaker buyers used their device for shopping (Perez, 2020).

The first security measure by Amazon is Passcode requirement while a purchase is being made. These passcodes are 4-digit codes which can be set and used for shopping. Alexa asks for the passcode and the user can speak these codes to authenticate the purchase. This is a security concern as anyone who is present at home can listen the passcode and use it whenever required.

The alternate to this is Voice ID feature. Voice ID can be used to build a personalised experience by making the device recognize individual's voice. These ID's can be used to create a kid's profile where the features can be restricted for them. An adult's permission will be required by a kid to purchase or access any particular restricted feature. In literature review it was found that most of the research done is on speech recognition instead of voice recognition. It was also found that hardware-based authentication, such as facial recognition, is suggested in research which increases the cost of the smart speakers (Sudharshan et al, 2019).

There is not much research done which discusses about the security concern of voice commerce or the voice recognition feature in smart speakers. Voice recognition is being widely used in banking sector to authenticate a customer and such technology sounds promising to authenticate unique users. Smart speakers, which can be easily accessed used by family members and unknowns should also be able to recognise the owner's voice to enhance the security of the device.

In literature review it can be seen that many research papers revolve around discussion of how to enable the voice recognition feature in the smart speakers instead of talking about the shortcomings or the way in which it can be improved. Therefore, there is a need to assess this feature so that the improvements can be done to ensure a security in voice commerce.

Research Question: How secure is Alexa's voice recognition for voice commerce? How is speech recognition and voice recognition implemented?

Potential Contribution: It is found that there is a need to discuss the difference between speech recognition and voice recognition to aware the consumers about the level of security being provided. The voice recognition feature has not been tested for the claims it has done and this research can be a steppingstone for enhanced security in voice commerce.

Structure of the Report: The remaining report is structured in a way that the Section 2 talks about the related work and how these have paved a way for the future work. The concepts of different research and how they can be implemented by merging them together is discussed in this section. Section 3 consists of research methodology whereas, Section 4 comprises the design specification. Section 5 is used to discuss implementation and section 6 discusses the evaluation of the same. Section 7 discusses the conclusion and future work. At the end of this research all references are provided in Section 8.

2 Related Work

The easiest way to shop these days is by shopping from the comfort of your home. E-shopping is a part of almost every single household in today's world and to give more convenience to the customers, voice-based shopping is introduced. It is known as Voice commerce, where a person can order something by just using their voice. It is prevalent in voice assistant devices like Amazon's echo devices powered by Alexa.

These devices take voice input from a user and provide a relevant product or service which can be purchased by confirming the order. The different aspects of this voice-based shopping, from a user's request to the order confirmation, are discussed in the subsections given below.

2.1 Voice Commerce: The fear of unknown

According to Berko, M. (2022) voice commerce transactions in 2021 reached almost $5 billion dollars. This is comparatively very low number considering the number of households which own a smart speaker. It is estimated that about 333 million smart speakers were owned in 2021 across the globe.

Voice commerce was found to provide more customer satisfaction than e-commerce according to a study by Kraus et al (2019). The customer satisfaction was found to be influenced by the transaction process efficiency. Despite all these benefits of voice commerce, there are still hesitations related to the use of these features in smart speakers.

Zaharia, S. (2020) successfully carried out a study about the acceptance of voice commerce using smart speaker and found out that there is a perceived risk related to voice commerce, which lowers the intention of a user to use the feature. Prior use or experience with a smart speaker was unable to increase the confidence of the users to encourage the shopping via these devices.

Another study which involved the understanding of consumer behaviour was carried out by Rzepka, C. et al (2020). The team conducted about 30 interviews with owners of smart speakers. Efficiency, convenience, and enjoyment which are the three pillars of any shopping experience were found and ranked highest in voice commerce. Even after getting all the benefits above any other type of shopping experience, lack of trust, limited transparency, low technical maturity, and lack of control were the main reason why the subjects of this study stayed away from using the voice commerce.

Research was conducted by Lei et al (2017), when voice authentication was not present in Alexa devices, which comprised of 5 human participants who gave commands to the device. These commands were about various IoT devices present in a home. Alexa took commands from everyone and performed the tasks without any authentication. Different methods were used to exploit Alexa and an acoustic speaker present in a room was used to deliver a command which was successfully executed by the device. This research gave an insight on how a smart speaker can be maliciously used in various situations.

2.2 The difference between speech recognition and voice recognition

The terms speech and voice recognition are used by a lot of people interchangeably, but they are very different from each other.

Speech recognition works by converting the speech input by a user into a text. This is a very useful feature which can be used to easily communicate with computers. The speech recognition helps us in day-to-day tasks when we ask something from google assistant using our voice. The new feature of google where someone can simply “hmm” or hum the song to get information about a similar song is achieved using this. Gaikwad, S. et al (2010) reviewed different techniques which can be used to achieve speech recognition. They discussed in detail about various methods in which an audio can be processed for analysis, feature extraction, modelling, and testing. They used a local Indian language to show the work of speech recognition, but all the different suggested methods can be used for any other language.

Voice Recognition on the other hand, is identifying who is speaking instead of understanding what is being spoken. To understand the advantage of voice recognition with speech recognition we can take a scenario of a meeting. In this meeting, we will consider that x number of participants are present, and we have to transcribe the whole meeting process. Using a microphone and a computer with speech recognition is the easiest way to document everything without a need of a human. The speech recognition will simply convert all the voice input into the text without separating any speaker. If voice recognition is added to this setup, what is being said by whom can be captured as the data.

Voice recognition is also known as Speaker recognition and Bai et al (2021) used deep learning to produce highly abstract embedding features when compared to conventional methods. The team however concluded that speech can be easily contaminated by the noise, channel distortion, and reverberations and the speaker recognition can be fooled.

2.3 Voice Authentication in Smart Speakers

According to Meng et al (2020) voice Authentication is an authentication measure where voice of a speaker is compared with a voice sample which is already enrolled. If the comparison of these voice samples meets a set threshold, the user is authenticated.

Venayagamoorthy, G. et al (2002) used neural networks to show the voice authentication can be achieved by creating spectrograms of samples and then converting these spectrograms to Power Spectral Density graphs of these spectrograms. The research by the team was successful by correctly recognizing the speaker and they were certain about the technology to be used in place of conventional keys and other access purposes.

2.4 Voice Commerce and Voice Authentication

The research discussed above paved a way to introduce voice authentication in smart speakers for voice commerce. Zhang et al (2017) discussed how voice authentication is a promising type of authentication after achieving 99% detection accuracy during the research.

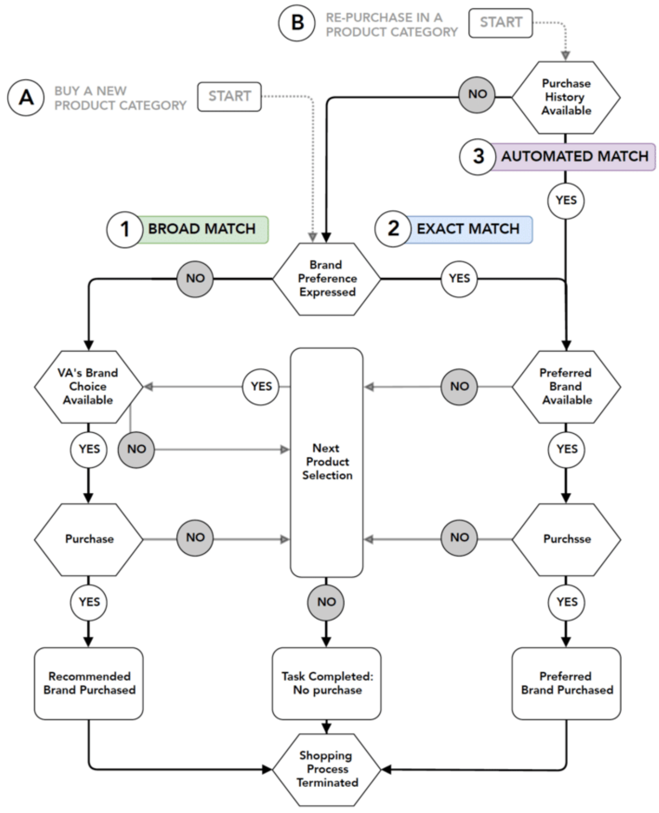

B
RE-PURCHASE IN A PRODUCT CATEGORY
START
A
BUY A NEW PRODUCT CATEGORY
START
Purchase History Available
NO
3 AUTOMATED MATCH
YES
1 BROAD MATCH
2 EXACT MATCH
Brand Preference Expressed
NO
YES
VA's Brand Choice Available
YES
Next Product Selection
NO
Preferred Brand Available
YES
NO
YES
Purchase
NO
NO
Purchsse
YES
NO
YES
Recommended Brand Purchased
Task Completed: No purchase
Preferred Brand Purchased
Shopping Process Terminated

Figure 1: Voice Commerce

Figure 1. depicts how voice can be easily used to purchase from a broad to an exact match. The user just has to ask for the required product, show interest by saying yes or no, and simply say yes to purchase an item. B part in the figure also shows how repurchasing a product is a very simple process.

The process seen above can be followed by anyone including a 3-year-old and an African Grey Parrot. Therefore, there was a need to introduce authentication measures. The three security measures by Amazon for Alexa devices are discussed in the table given below

Table 1: Security Measures by Amazon

S. No.	Security Measure	Definition	Drawback
1	Passcode Authentication	A user can choose a 4-digit passcode which is asked while making a purchase	Anyone who is present nearby can memorise the passcode and use it later
2	Voice ID	Voice or Speaker recognition which can automatically detect the owner and can authenticate a user based on profile	No tests done for voice recognition system as of now
3	Switch off the voice commerce feature	Switching off voice commerce in feature using Alexa App	Switching off a feature is taking away the convenience of shopping using voice

There are many research articles on speech recognition, voice recognition and voice commerce. There is however a requirement to do tests on voice recognition provided by smart speakers to improve the technology.

3 Research Methodology

As this research is intended to figure out the various aspects of voice used in smart speakers. There are 3 phases of this research:

1. Speech Recognition
2. Voice Recognition
3. Testing Alexa's Voice ID

3.1 Speech Recognition

Smart Speakers like Amazon's Alexa use Automatic Speech Recognition (ASR), which is a machine learning technology. Kalita, D. (2022) discussed how ASR works using two different types of models.

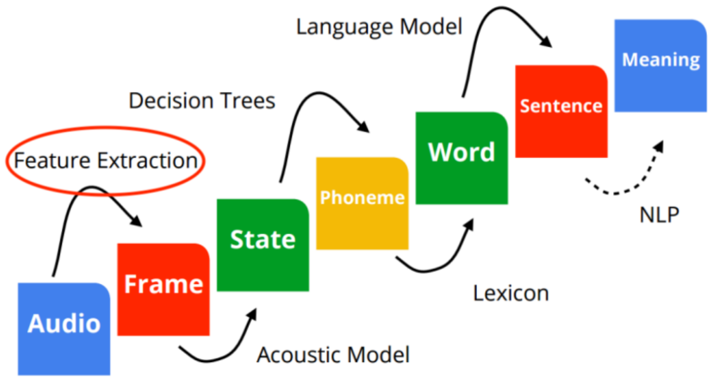


Figure 2 ASR Model

Figure 2. shows Acoustic and Language models which are used to convert speech to text. Acoustic model work by comparing sounds which make up a word. In English there are approximately 40 different sounds that can make up a word. Language model works by comparing these sounds with words which sound similar and then matching with word sequences.

There are many pretrained models which can be used to create a simple speech recognition system. These models are trained using huge chunks of data to achieve the accuracy required to work in a live environment. Considering this we used Wav2Vec 2.0 which is widely used by giants like Amazon and Facebook. Wave2Vec 2.0 has pretrained models and can be used to create a simple speech recognition system.

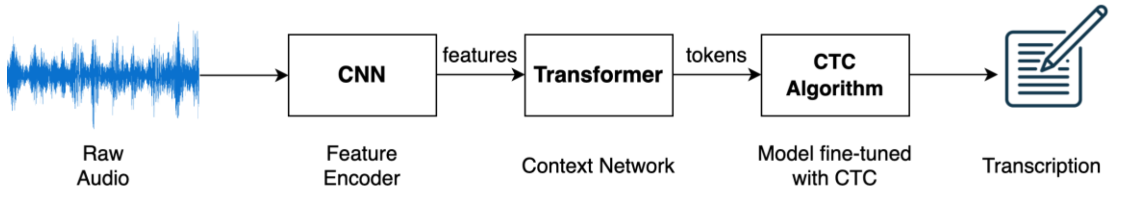


Figure 3: Wave2Vec 2.0

In Figure 3. we can see how raw audio goes through feature encoder to extract the features. These features are fed to context network which generates tokens for CTC Algorithm. The CTC algorithm is Connectionist Temporal Classification, an algorithm which is used to train neural network.

3.2 Voice Recognition

This is the second phase of our research. We will be using Gaussian Mixture Model to identify the speaker. In order to achieve this, the first step will be to extract the features of the audio clip. Mel frequency cepstral coefficient (MFCC) is used for this task as seen in Figure 4.

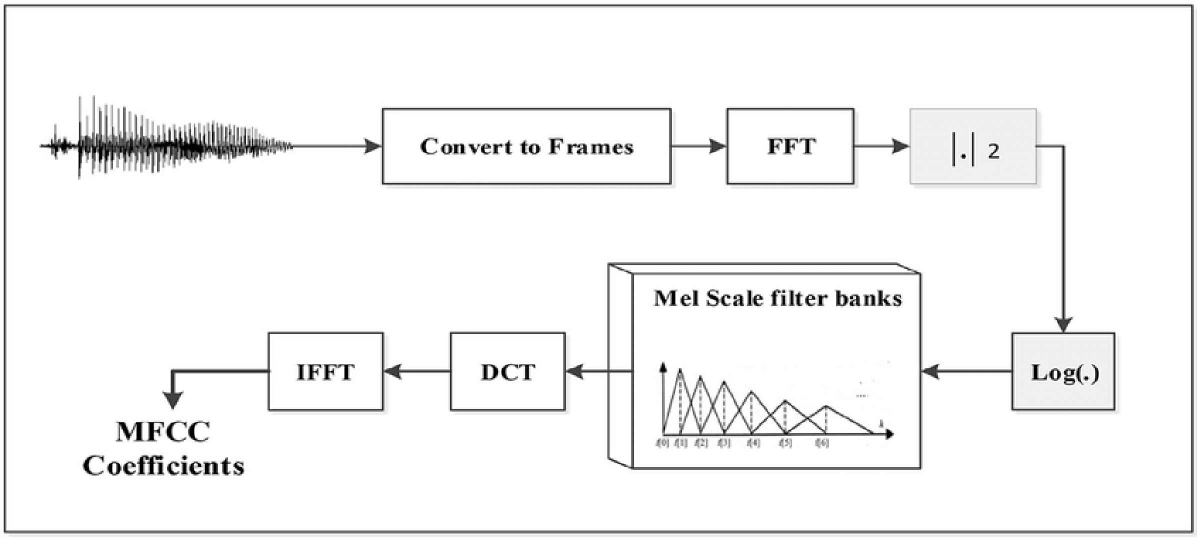


Figure 4: MFCC

After the extraction of the features from audio is done, Gaussian Mixture Model (GMM) is used to train the voice recognition system along with the MFCC. After the training is done, scores of the features are calculated for all the models using GMM model. The speaker model which has the highest score is then selected as the speaker.

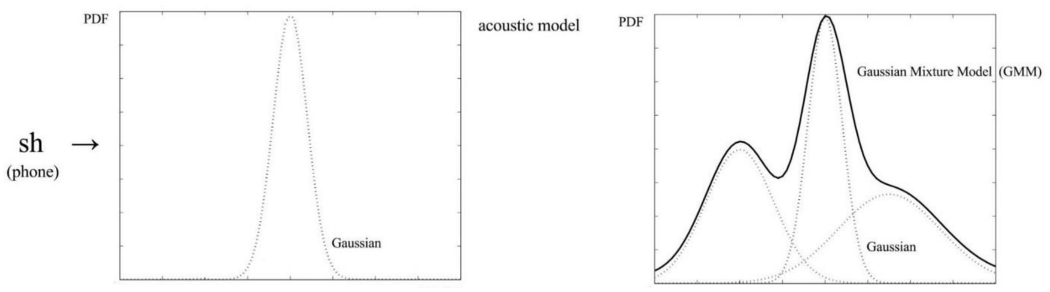


Figure 5: GMM

Figure 5. shows how a single Gaussian model works by plotting a single acoustic model vs how a GMM model plots 3 different acoustic model in a single graph.

3.3 Testing Alexa's Voice ID

Testing the Voice ID of Alexa for any shortcomings is the final as well as most important phase of this research. This test will help us in understanding how secure the voice commerce is by providing accuracy ratings of the voice recognition.

In order to test Alexa's Voice ID, various test cases are structured along with 4 different speakers with different account types in Alexa. The test cases involved questions being asked to Alexa as well as crucial tasks such as purchasing something using voice are performed.

4 Design Specification

4.1 Speech Recognition

Tokenization

One of the main functions of speech recognition is tokenization of the data. Tokenizer is responsible for breaking the raw speech into small chunks of data. This speech when broken into words or sentences is known as token.

These tokens are used to provide the understanding of the context and to interpret the meaning of the speech after the analysis is done on the sequence of the words. For example, the speech "How are you" can be tokenized into "How", "are", "you".

When the speech is split into words it is known as word tokenization, and when it is split into sentences it is known as sentence tokenization.

There are various types of tokenization techniques as shown in Figure 6. Word Tokenization can be space, punctuation, or rule-based tokenization. A space-based tokenization technique is the simplest tokenization technique where any space in speech or text input is split. This splitting results in words as tokens from a sentence or a paragraph.

The dictionary tokenization works by giving tokens to all the words present in the dictionary and then finding the same tokens in the input. The tokens which are not present in the dictionary are assigned using special rule. This is an advanced technique when compared to white space tokenization technique.

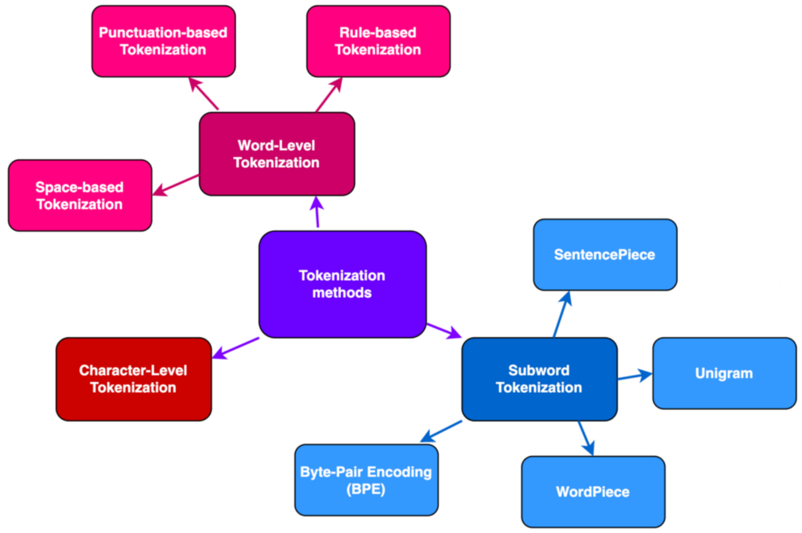


Figure 6: Tokenization techniques

Connectionist Temporal Classification

This a very old and reliable model of deep speech recognition. This model was designed to align the audio with the transcription.

If we consider the speech input, X = [x1,x2,…,xn] and the set of transcript Y= [y1,y2,…..,yn], X and Y are both variables. Sampling rate (X) and words in speech (Y) can be very random and the CTC provides a way to compute the conditional probability and the solution.

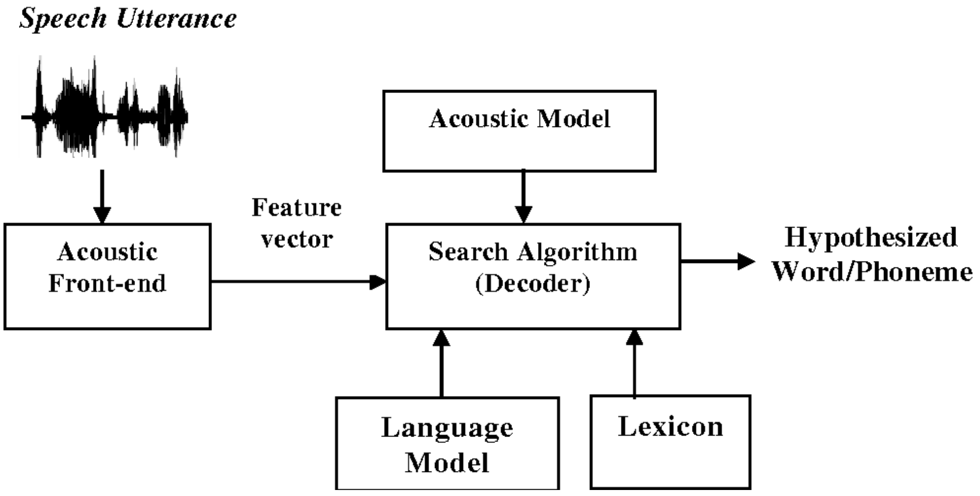


Figure 7: Speech Recognition Architecture

Figure 7. depicts the way in which the feature is extracted from the speech and is processed using Language and Acoustic model in order to give a hypothesized word.

4.2 Voice Recognition

Mel-Frequency Cepstral coefficients (MFCC)

One of the main components of voice and speech recognition is Mel-frequency cepstral coefficients (MFCC). It is one of the widely used technique to extract the features from the audio signal.

The steps involved in MFCC are shown in figure 8.

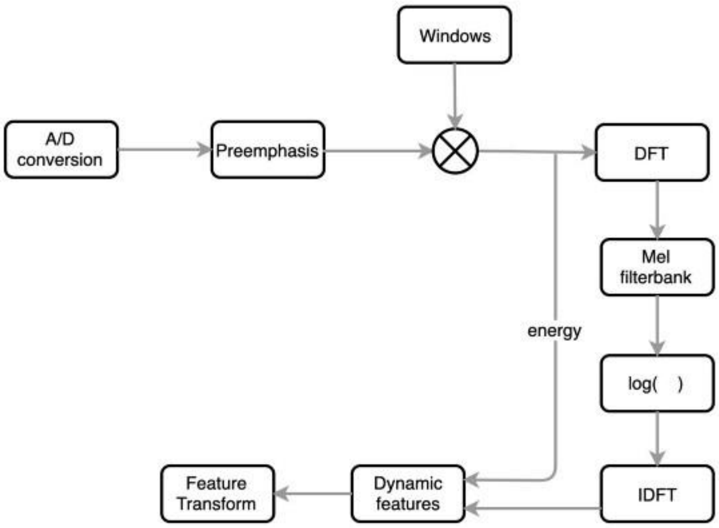


Figure 8. MFCC

In first step, audio is converted to a digital format. The second step of Preemphasis is used to increase the level of energy in the higher frequency. This is done as in the vowels it is found that energy is much less at higher frequencies when compared to the lower frequencies. This increases the phone detection accuracy.

The third step, windowing, is used to break the segment of audio into smaller segments of 25 ms with a 10ms signal space. For each of the segment around 39 features are extracted. The Discrete Fourier Transform (DFT), or the fourth step converts the signal into frequency domain from the time domain. The fifth step, Mel filterbank, is one of the most crucial steps of MFCC as it models the human ear to machine which improves the performance of the model as machine detects the same resolution at different frequency. The dynamic feature step is used to provide 39 features.

Therefore, MFCC provides 39 features from a given audio signal which can be used for comparison for speech and voice recognition

Gaussian Mixture Model (GMM)

Every observation in a signal can be modelled as a Gaussian process, with a (multivariate) Gaussian (normal) distribution. However, speech signals have a different structure which can not be modelled using simple Gaussian model. Mixture Model provides different classes which have unique statistical model. Unvoiced and Voiced classes are separated by modelling them in separated classes and the joint distribution of these classes is the weighted sum. The weight here corresponds to the frequency in which they appear in a signal. For example: if unvoiced signal constitutes 25% of the speech sound, then the weight will be 0.25 of the unvoiced class.

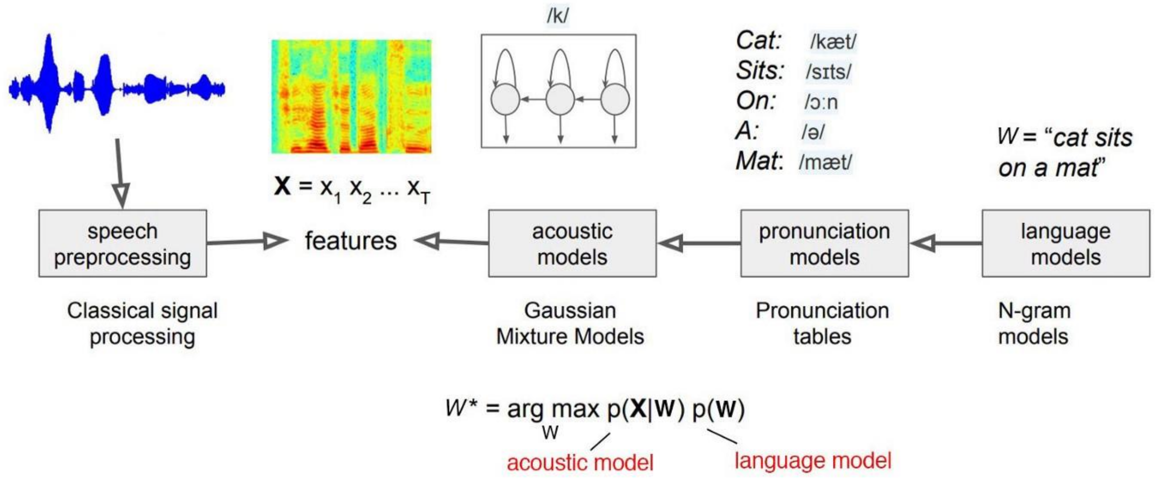


Figure 9: Voice Recognition Architecture

Figure 7 depicts the way in which the speech is pre-processed, and features are extracted for training, testing and verification.

4.3 Testing Alexa's Voice ID

In order to test the Voice Recognition feature by Alexa, an Alexa device is set up with a shipping information and payment details. Alexa's Voice ID feature is used, and the owner of the device is given a Voice ID Profile. 2 more Voice ID profiles are created to test the voice recognition feature, while one unrecognized voice is used to test the restrictions implemented in the system.

5 Implementation

5.1 Speech Recognition

Speech Recognition system is developed using Jupyter Notebook in PyCharm, a tokenizer and a pre-trained model. The model used in this research is pre-trained by Facebook. The sampling rate from the audio files was changed to 16000Hz with the help of Librosa, as it is the accepted range for Facebook's model.

This acceptable input audio is then transferred to the tokenizer which is processed by the model. The result is finally stored in a transcription.

The speech recognition system is implemented in Jupyter Notebook using the following libraries:

1. Torch
2. Librosa
3. Numpy
4. Soundfile
5. Scipy
6. IPython
7. Transformers

After the system is loaded with tokenizer and model, the audio is provided. It is a recorded .wav file which is then transcribed using the speech recognition system.

The output produced by this speech recognition system is transcription of the audio file.

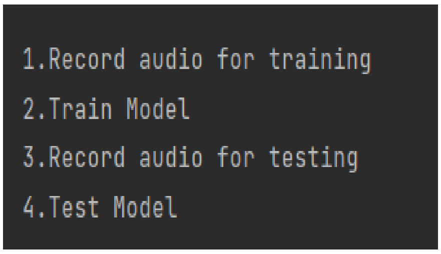
5.2 Voice Recognition

Voice Recognition system is developed using Python in PyCharm. The audio is recorded 5 times from the user for the purpose of training. After this audio is recorded, features are extracted from the audio samples using MFCC and then training of the model is done using GMM and MFCC.

For the testing purpose, the scores for all the models are calculated using GMM models. The model with the highest score is then selected as the speaker.

The python libraries which are used are:

1. PyAudio
2. Wave
3. Pickle
4. Speech_recognition
5. Numpy
6. Sklearn
7. Scipy
8. Python_speech_features



```
1.Record audio for training
2.Train Model
3.Record audio for testing
4.Test Model
```

Figure 10: Voice Recognition system options

Fig 10. Shows the four different options provided by the voice recognition system.
A single user was enrolled and then another unrecognised user was used to test if the voice recognition system is able to recognize the unknown speaker.

5.3 Testing Alexa's voice ID

4 different types of users are used for the testing purposes.

Table 2: Users for Alexa Voice ID

Sr No	User	Type	Access to Voice Commerce
1	Dev	Owner, Recognised	Yes
2	Ishaan	Recognised	No
3	Cheeku	Recognised	No
4	Abhishek	Unrecognised, unknown	No

A question, “Who am I?” or “Do you recognise me?” is asked by these users, in random order. Also, the Voice commerce restriction based on voice recognition is tested by these users.

6 Evaluation

6.1 Experiment 1: Speech recognition

Speech Recognition system is tested for 3 audio files. The first one is a short voice recording of about 4 seconds and it gives us the output:

“THIS IS A VOICE RECARDING FOR MY RESEARCH.”

This output is almost correct with the only issue being the spelling of RECORDING. The second audio provided is a longer audio file, it is about 18 seconds and produces the output as:

“IN THIS ORDER ACCORDING IAM TRAIN TO SHOW THAT THE WORDS CAN BE CHANGED TO THE TEXT EASILY AND THIS PREDICTIS LITERATL EXCELEDAM TRANG TO SHOW THAT IT IS EASU TO FOOL ELEXISE O VOICE RECOGONITION SYSTEM.”

The output produced is jumbled, although any human can unjumble the given output and can conclude what is being said, the result is not accurate.

The third audio file used is slightly shorter than the second one, 15 seconds and was made keeping the results of the second test in mind. The voice is used in a clear manner by providing spaces between the words, so that the algorithm can catch the words right. The output produced is:

“IF THEIS ORDER REGARDING I AM TRYING TO SHOW THAT WORDS CAN BE CHANGED TROG TEXT EASILY AND THIS PROJECT IS A BOLD ENEXA AND HOWITS VOICE RECOGNITION SYSTEM CAN BE EASILY FOOLED”

The results of this test are similar to the second one.

Table 3: Speech Recognition Test

Test Number	Audio Length	Accurate Output
1	4 seconds	Yes
2	18 seconds	No
3	15 seconds	No

6.2 Experiment 2: Voice recognition

The voice recognition system was able to recognize the speakers correctly.

Table 4: Voice Recognition Test

Speaker	Type	Result
Dev	Recognized	Was recognized by the system correctly

Ishaan	Unrecognized, Unknown	Was declared unrecognized by the system

6.3 *Experiment 3: Alexa's Voice ID*

Four types of users were used for this experiment and the results are:

Table 5: Alexa's Voice ID Test

User	Type	Voice Commerce Permission	Voice Recognition Result	Voice Commerce Test
Dev	Owner, Recognized	Allowed to purchase	Recognized as Dev	Was able to Order
Ishaan	Recognized	No permission granted	Recognized as Dev	Was able to Order
Cheeku	Recognized	No permission granted	Recognized as Cheeku	Was not able to Order
Abhishek	Unrecognized	No permission granted	Recognized as Dev	Was able to Order

6.4 *Discussion*

Experiment 1 is done on a speech recognition system that employs the model trained by Facebook, Wav2Vec2.0 which is also used by Amazon. The expectation from the results of such system is high but the result proved to be average where only small speech input was converted to text successfully. Similarly, smart speakers often take input for short period of time before requiring the wake word to be spoken again. There are times when a user is speaking and Alexa device's ring led, which indicates the listening of the device is on, turns off in the middle of the speech. The experiment was successfully able to show that small speech data is processed in a better manner on a device, however, a device which performs this data over the cloud, like voice assistant, will be able to perform a better computation in less time because of better processing power.

Experiment 2 is done to understand how the GMM and MFCC work to differentiate one speaker from another. This is useful for the times when the voice-based authentication is required. The recognized user is successfully recognized by the developed system and the unrecognized speaker is not recognized as the recognized speaker. This is the best example of how voice-based authentication should be able to differentiate between different speakers.

Experiment 3 is the test of Voice ID by Amazon which is used to recognize the voices on the smart speakers. Voice ID does not only provide a personalised experience, where the voice assistant is able to give relevant information, but also a security feature where the purchase made through voice can be authenticated automatically. In the test, it was found that recognized user who does not sound similar to another recognized user is misidentified. Another unrecognized user was also identified as the recognized user with privileges. This is a big security concern as two different users were able to purchase something from voice commerce, without the requirement of going through any other type of authentication.

During the research it was found that "Hi Santa" which is a skill for the Alexa device sent a push notification to the smartphone to grant permission but not a single push notification was received regarding the order.

7 Conclusion and Future Work

In this research, a serious security concern of Alexa's voice recognition for voice commerce is found. The ability for an unauthorised and unrecognized person to be recognized as the authorized person because of the incompetence of voice recognition system can be a hazardous threat. Anyone who can come in close proximity of the smart speaker even by using another speaker in the room can give the commands to other connected IoT devices.

An unrecognized person can simply ask the Alexa to open the smart lock on the main gate, open a window, increase the temperature on the smart thermostat, purchase something over voice commerce, and even turn off the smart lights.

The use of Passcode also has a negligible effect on security as it can be heard from far. There is an immediate need to enhance the security of these smart speakers to ensure the safety of the user as well as to gain the trust of new users.

The limitation for the research conducted is that there is not enough user experiment data to come to the accuracy percentages, but the small number of users were able to produce the same result for higher number of times.

For the future work in this field, the implementation and improvement of CNN based voice recognition system over the cloud will be beneficial. A push notification system for voice transactions is suggested and the implementation of the same can be done using API by Amazon. A two-factor authentication for voice transaction can also be implemented in hardware as well as software.

A hardware based two-factor authentication measure will be more secure but will increase the cost of the product. A software based two-factor authentication can be One Time Passwords (OTP) or push notifications, like Google and Facebook, where the user has to only accept or decline the activity performed.

References

1. Schlosser, K., 2022. Girl, 6, ‘accidentally’ orders $170 dollhouse and 4 pounds of cookies using Amazon's Alexa. [online] GeekWire. Available at: https://www.geekwire.com/2017/girl-6-accidentally-orders-170-dollhouse-4-pounds-cookies-usingamazons-alexa/ [Accessed 05 April 2022]
2. Insider. 2022. A Michigan mom says her kids used Alexa to buy $700 worth of toys on her credit card. [online] Available at: https://www.insider.com/kids-alexa-buy-700-worth-of-toys-moms-credit-card-2019-12 [Accessed 05 April 2022].
3. NDTV.com. 2022. Naughty Parrot Uses Amazon Alexa To Shop While Owner Is Away. [online] Available at: https://www.ndtv.com/offbeat/naughty-parrot-uses-amazon-alexa-to-shop-while-owner-is-away-1963811 [Accessed 05 April 2022].
4. The Sun. 2022. Cheeky toddler, 3, shocks mum by buying FARTS through Amazon Alexa. [online] Available at: https://www.thesun.co.uk/news/17063683/toddler-buys-farts-amazon-alexa/ [Accessed 05 April 2022].
5. Sudharsan, B., Corcoran, P. and Ali, M. (n.d.). Smart speaker design and implementation with biometric authentication and advanced voice interaction capability. [online] Available at: http://ceur-ws.org/Vol-2563/aics_29.pdf.
6. Statista. 2022. Topic: Smart speakers. [online] Available at: https://www.statista.com/topics/4748/smart-speakers/ [Accessed 05 April 2022].

7. Kraus, D., Reibenspiess, V. and Eckhardt, A. (2019). How Voice Can Change Customer Satisfaction: A Comparative Analysis between E-Commerce and Voice Commerce. Wirtschaftsinformatik 2019 Proceedings. [online] Available at: https://aisel.aisnet.org/wi2019/specialtrack01/papers/7/.
8. Anon, (n.d.). Is Voice Commerce the Future of Ecommerce? | Whiplash. [online] Available at: https://whiplash.com/blog/voice-commerce-future-ecommerce/ [Accessed 12 Aug. 2022].
9. Venayagamoorthy, G.K., Moonasar, V. and Sandrasegaran, K. (n.d.). Voice recognition using neural networks. Proceedings of the 1998 South African Symposium on Communications and Signal Processing-COMSIG '98 (Cat. No. 98EX214). doi:10.1109/comsig.1998.736916.
10. Tahseen Ali, A., Abdullah, H.S. and Fadhil, M.N. (2021). Voice recognition system using machine learning techniques. Materials Today: Proceedings. doi:10.1016/j.matpr.2021.04.075.
11. Abilitynet.org.uk. (2019). Voice Recognition - An Overview | AbilityNet. [online] Available at: https://abilitynet.org.uk/factsheets/voice-recognition-overview.
12. aws.amazon.com. (2022). Fine-tune and deploy a Wav2Vec2 model for speech recognition with Hugging Face and Amazon SageMaker | AWS Machine Learning Blog. [online] Available at: https://aws.amazon.com/blogs/machine-learning/fine-tune-and-deploy-a-wav2vec2-model-for-speech-recognition-with-hugging-face-and-amazon-sagemaker/ [Accessed 12 Aug. 2022].
13. Bhapkar, V. (2020). Speaker Identification Using Machine Learning. [online] Medium. Available at: https://medium.com/analytics-vidhya/speaker-identification-using-machine-learning-3080ee202920.
14. Zaharia, S. and Würfel, M. (2020). Voice Commerce - Studying the Acceptance of Smart Speakers. Human Interaction, Emerging Technologies and Future Applications III, pp.449–454. doi:10.1007/978-3-030-55307-4_68.
15. Rzepka, C., Berger, B. and Hess, T. (2020). Why Another Customer Channel? Consumers' Perceived Benefits and Costs of Voice Commerce. [online] scholarspace.manoa.hawaii.edu. Available at: https://scholarspace.manoa.hawaii.edu/items/754c1523-be45-44f2-a565-0d0eaf0d1045 [Accessed 25 Jul. 2022].
16. Mohd Hanifa, R., Isa, K. and Mohamad, S. (2021). A review on speaker recognition: Technology and challenges. Computers & Electrical Engineering, 90, p.107005. doi:10.1016/j.compeleceng.2021.107005.
17. Gaikwad, S.K., Gawali, B.W. and Yannawar, P., 2010. A review on speech recognition technique. International Journal of Computer Applications, 10(3), pp.16-24.
18. Bai, Z. and Zhang, X.-L. (2021). Speaker recognition based on deep learning: An overview. Neural Networks, 140, pp.65–99. doi:10.1016/j.neunet.2021.03.004.
19. Boles, A. and Rad, P., 2017, June. Voice biometrics: Deep learning-based voiceprint authentication system. In 2017 12th System of Systems Engineering Conference (SoSE) (pp. 1-6). IEEE.
20. Analytics Vidhya. (2022). A Comprehensive Overview on Automatic Speech Recognition (ASR). [online] Available at: https://www.analyticsvidhya.com/blog/2022/03/a-comprehensive-overview-on-automatic-speech-recognition-asr/.
21. Apple Machine Learning Research. (n.d.). Hey Siri: An On-device DNN-powered Voice Trigger for Apple's Personal Assistant. [online] Available at: https://machinelearning.apple.com/research/hey-siri.

22. Arias, J. (2020). How to build a Neural Network for Voice Classification. [online] Medium. Available at: https://towardsdatascience.com/how-to-build-a-neural-network-for-voice-classification-5e2810fe1efa.
23. Science ABC. (2020). How Does Alexa Identify Who Is Speaking? [online] Available at: https://www.scienceabc.com/innovation/how-does-alexa-identify-who-is-speaking.html.
24. Chakravarthy, S. (2020). Tokenization for Natural Language Processing. [online] Medium. Available at: https://towardsdatascience.com/tokenization-for-natural-language-processing-a179a891bad4.
25. Analytics Vidhya. (2021). MFCC Technique for Speech Recognition. [online] Available at: https://www.analyticsvidhya.com/blog/2021/06/mfcc-technique-for-speech-recognition/.
26. wiki.aalto.fi. (n.d.). Gaussian mixture model (GMM) - Introduction to Speech Processing - Aalto University Wiki. [online] Available at: https://wiki.aalto.fi/pages/viewpage.action?pageId=151492301 [Accessed 11 Aug. 2022].