**QUEST INTERNATIONAL UNIVERSITY**

**PERAK**

**FACULTY OF SCIENCES AND TECHNOLOGY**

**Bachelor of Computer Science (Hons)**

**BCS3105 CS PROJECT PHASE 1**

**PROJECT PROPOSAL**

**PREPARED BY:**

|  |  |
| --- | --- |
| **Name (Matrics No.)** | **TAN ZHONG MING (QIUP-201804-002120)** |
| **Project Title** | **DETECT STUTTERED SPEECH BY USING MACHINE LEARNING TECHNOLOGY** |
| **Supervisor** | **DR LEE LAM HONG** |
| **Moderator** |  |
| **Submission Date** | **8 MAY 2020** |

Table of Contents

[Abstract 3](#_Toc39868457)

[Introduction 4](#_Toc39868458)

[Problem Statement 6](#_Toc39868459)

[Objectives 7](#_Toc39868460)

[Scope 8](#_Toc39868461)

[Journey map 15](#_Toc39868462)

[System architecture 17](#_Toc39868463)

[Modules and functionality 18](#_Toc39868464)

[High level system flow 19](#_Toc39868465)

[Milestones 20](#_Toc39868466)

[Conclusion 20](#_Toc39868467)

[References 21](#_Toc39868468)

# Abstract

The traditional method to remove a stuttered speech from video required a lot of human effort. The video editor must play the whole video and watch numerous time to remove the stuttered speech. Apart from that, to remove a stuttered speech from the footage required to learn editing tools such as Adobe Premiere, Sony Vegas, Hitfilm and more. The learning curve for editing tools is very high, and they are different from each other due to different concept and different workflow. To remove a stuttered speech, the user must learn the timeline, timecode, audio wave and more. Next, existing speech recognition is not able to detect the stuttered speech, so when the user requests the API, it will return a weird result.

In this project, a camera app will develop and implementing machine learning. The goal of the app is using machine learning to identify and remove the stuttered speech from the video. It can reduce the learning curve significantly and reduce the human effort required to edit a video. First, the audio will extract from the video and undergo pre-processing algorithm such as amplify and normalization. Pre-processing audio can provide a better result for the next step. Next is implement the MFCC algorithm to extract the audio features. After generating the features for each data in the dataset, the features will ingest to a neural network SVM for training purpose. The trained model will save in a specific format file protobuf. Then, the protobuf will include as an asset in the android app. By using the Android SDK and Android NDK, we can use the function provided by TensorFlow Lite to classify the input by using this special format file. The stuttered part in the video will be labelled by a remove algorithm to generate a timestamp. Then the video will undergo an L-cut algorithm will cut the video base on the generated timestamp. After that, the app will render the video base on the timestamp and store it in mp4 format with the subtitle.

# Introduction

Stutter

According to the Cambridge dictionary, stuttering is a speech disorder when someone tries to speak something with difficulty, especially the first part of a word. For instance, pause before the sentence ends or repeating the same name or phonemes. Speech is one of the communication methods used by human to express their feelings, idea and thoughts. According to (Vikhyath Narayan K N, S P Meharunnisa, 2016) and (Swapnil D. Waghmare, Ratnadeep R. Deshmukh, Pukhraj P. Shrishrimal, Vishal B. Waghmare, Ganesh B. Janvale and Babasaheb Sonawane, 2017), there is an approximate 1 % of the population in the world faces the stuttering problem. Stuttering is a common speech disorder when patients in a tense environment. There are several types for the dysfluencies. Repetition is one type of dysfluencies which include syllable repetition, whole world repetition and phrase or sentence repetition. The following type of dysfluencies is prolongation, interjection and pauses.

Repetition occurs when a syllable or sound is repeated at the beginning of the word. For example, "the baby-baby ate the soup" and "W-W-W- Where are you going?". Prolongation occurs when the speaker prolongs the sounds or syllable such as "The baaaaaaby ate the soup". Interjection commonly uses "um" and uh" to fill up the pause sector also known as filled pause or fillers. For example, "the baby um uh ate the um soup". Pauses occur when the speaker trying to pronouns syllable with difficult and causes broken words or sentence. For example, "IT was won[pause]derful".

Machine Learning

According to (Tom Mitchell, 1998), machine learning is a well-posed learning problem which means a computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E. Machine learning is a new capability for computers such as data mining, self-customizing program, the application can't program by hand. For example, Natural Language Processing. Machine learning divided into two main category which are supervised learning and unsupervised learning. There is more category from that, such as reinforcement learning, recommender systems and more. Supervised learning divided into two categories. First, regression problem is to predict a valued output based on the input such as house price. Supervised learning does provide a right answer at the end compare to unsupervised learning. Second, classification problem is to classify the object base on the parameter given. The common classification problem is to classify have or don't have breast cancer based on the tumour size. The classification problem can become more complex based on the situation. Support vector machine (SVM) is introduced to solve when there is a two-group of classification problem. (Stecanella, 2017). Unsupervised learning is one of the machine learning algorithms. The goal of unsupervised learning is to draw inferences base on the datasets given without any labelled responses. In other meaning is ask the machine to classify and provide an inference based on the dataset given. The standard algorithm included hierarchical clustering, k-Means clustering, Gaussian mixture models, Self-organizing maps, Hidden Markov models. (MathWorks, n.d.)

Video

According to (Celie O’Neil-Hart, Howard Blumenstein, 2016), 6 out of 10 people is prefer to watch video online rather than television. According to (Youtube for Press, n.d.) the number of channels owns more than a million subscribers were grown by more than 65%. The number of video views grew 120,589,156 on 21 April 2020 for the highest number of subscriber channel, T-Series. T-Series channel grew around 100k video views daily. The growing of youtube significantly shows the importance of online video and shows the demand for a video editor. Most of the Youtuber does hire at least one or more edit, or to edit their video daily. However, to become an editor required a large amount of skill and years of training such as composition, colour science, graphic design skill, sound processing and various skill. According to Linus Tech Tips channel on YouTube, the average salary for his video editor is $29 per hour, and he has 7 video editors. It does cost $420,000 for a year.

# Problem Statement

1. **The editor must remove the stuttered speech manually by using existing tools**

Existing tools are not capable of removing shuttered speech from the video automatically. So, the editor must playback the video to identify the shuttered speech and remove it automatically. Most of the time editor will play the video numerous time to ensure there is not stuttered speech in the video. Remove manually will cause the editor to spend most of the time just for removing the stuttered part instead of the content. The cost to run a media company will also increase due to the long edit time. According to Linus Tech Tips channel, hire a video editor will cost $29 per hour, but most of the time editor just removing the stuttered speech and find out the successful part to make a video.

1. **Non smooth jump cut in video**

Although the removing the stuttered speech from the video is possible. But the cut perform by AI is not smooth and look well in video. In other meaning is it will look like a sudden jump in the video. This will provide a weird experience for the audience.

# Objectives

1. To identify and remove stuttered speech by using machine learning.

The traditional method is to identify the stuttered speech by listening and watching the video numerous time. A trained machine learning model will be capable of identifying the stuttered speech without any human involvement, including the threshold. As a comparison, a machine learning method can identify the stuttered speech in a short amount of time. Another side, human has to listen and watch the video over and over again. The trained machine learning model will identify the stuttered speech and provide a timestamp for remove stuttered speech algorithm. The second benefit after remove the algorithm is to provide a more reliable speech recognition by using API. Existing speech recognition API not able to detect the stuttered speech so this project also provides a tool for better speech recognition by removing the stuttered speech from video. Apart from that, after remove the stuttered speech from the video it can reduce the size of file that required send to API for speech recognition. In other meaning, this will save a lot bandwidth when calling the speech recognition API.

1. Perform L-cut algorithm

A L-cut algorithm can provide a better user experience, it will make the video look like more natural and smoother. L-cut required to overlap the audio with around 0.5s frame and bring forward the next scene around 0.5s. This is a standard cut technique that used in filming schools.

# Scope

1. Perform pre-processing techniques to audio data set for better sound quality
2. Build and train an Artificial Neural Network (ANN) for the stuttered speech classification model.
3. Develop an application and algorithm for Android to automatically remove the stuttered speech and transcribe the audio.
4. **Domain**

Extract features is a critical step to solve the classification problem and identification problem. Mel Frequency Cepstral Coefficient (MFCC) is widely used as a feature for voice recognition due to its adaptability to background noise, emotion and more. The goal of MFCC is to filter out the shape of the vocal tract such as the tongue and teeth. Amplitude also provides a significant result to detect the stuttered speech, and it does reduce the time required to train a classification model.

1. **Hardware**

Minimum requirement

|  |  |
| --- | --- |
| Hardware | Specification |
| Computer | RAM: 4GB minimum, 8GB recommended  Hard disk: 4GB of available disk space  Resolution: 1280 x 800 minimum  OS: Windows 7 (64bit) or later |
| Smartphone | ROM: 1GB  RAM: 2GB  OS: Android 6.0 (Marshmallow) |

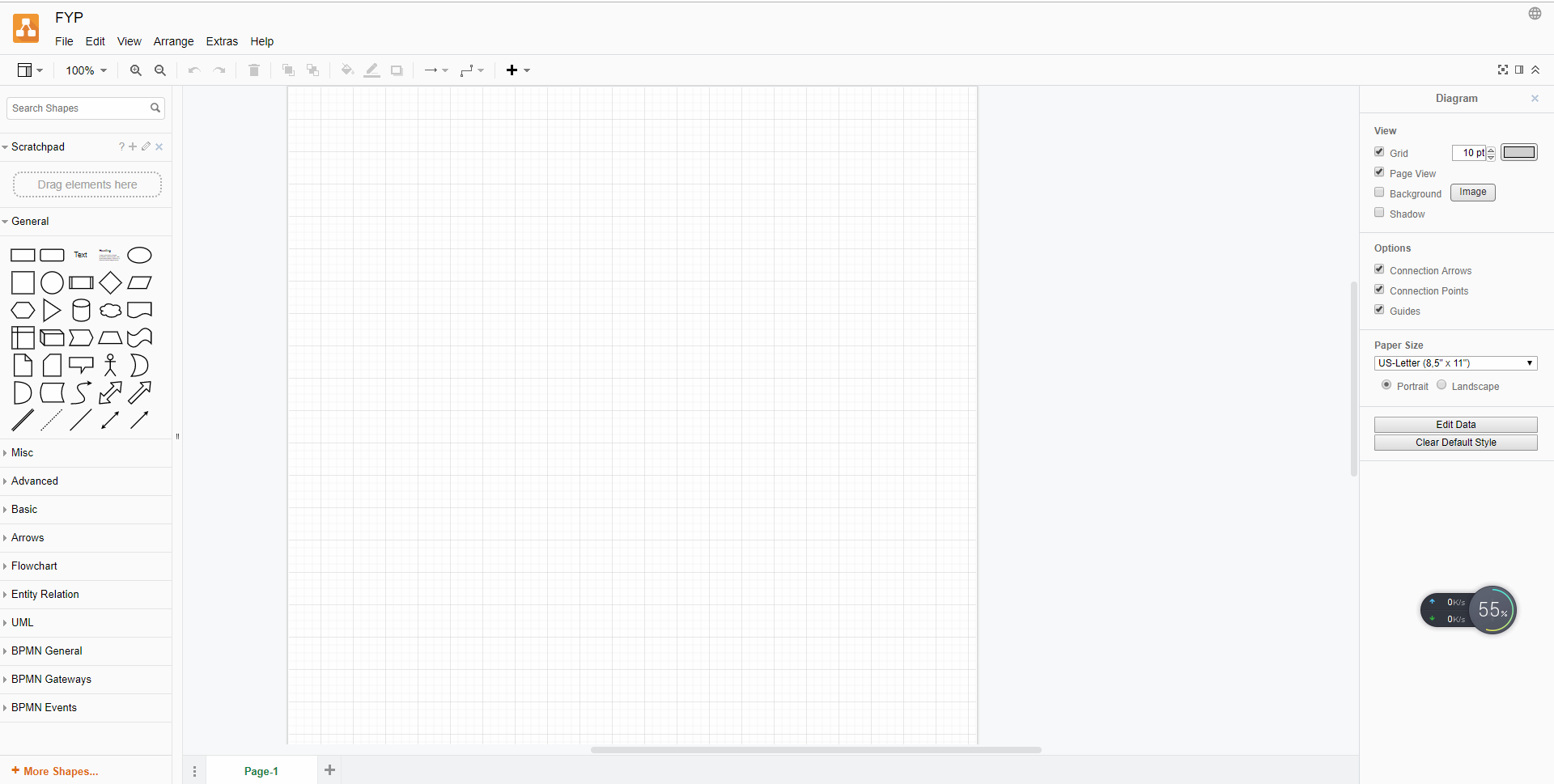
The minimum requirement for TensorFlow Lite is Android 6.0. According to (Mobile & Tablet Android Version Market Share Worldwide, 2020), there is only 8.72% people using android 6.0 and 37.4% of people using Android 9.0. So, the test platform will start with Android 6.0 since that is the minimum requirement.

Development and deployment

|  |  |
| --- | --- |
| Hardware | Specification |
| Laptop | Model: Acer E15 575G-55Z3  OS: Windows 10  RAM: 8GB  GPU: Nvidia GeForce 940MX  Processor: Intel i5-7200u  Addon:  SSD 480Gb  HDD 1TB |
| Smartphone | Model: Oppo A57 (CPH1701)  Display: 720 x 1280 pixels  Platform: Android 6 (Marshmallow)  OS: ColorOS3  Chipset: Qualcomm MSM8940 Snapdragon 435 (28 nm)  CPU: Octa-core 1.4 GHz Cortex-A53  GPU: Adreno 505  Main camera: 13 MP, f/2.2, PDAF, 1080p@30fps  Internal Storage: 32GB 3GB RAM |

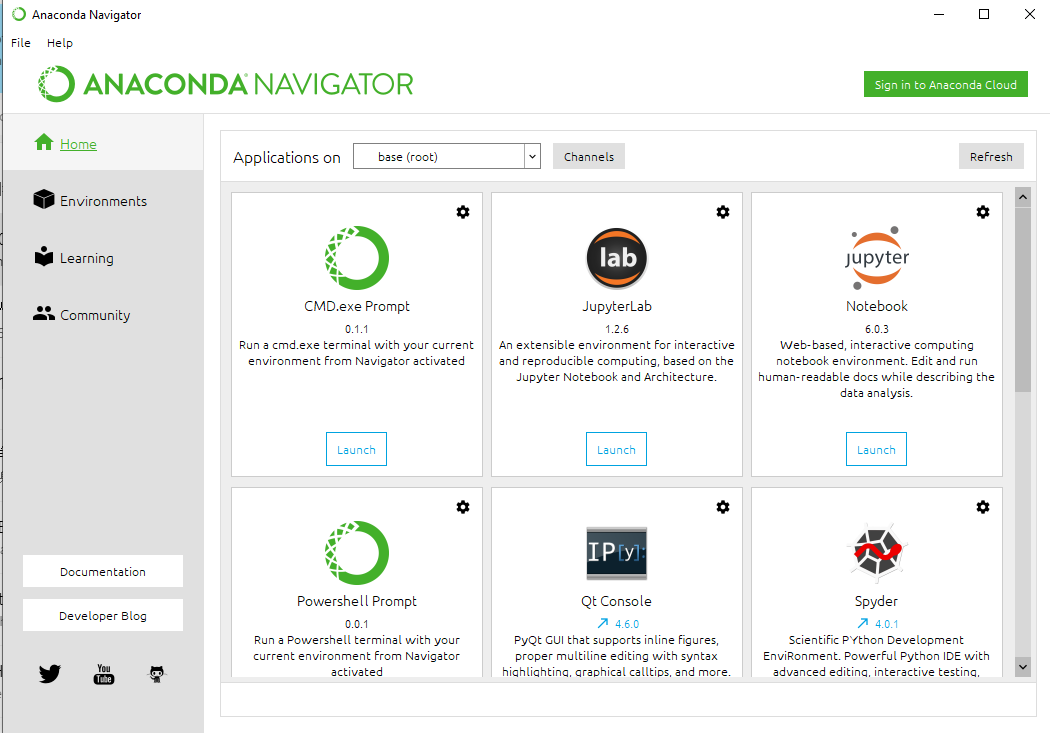
1. **Software**
   1. Minimum requirement
      1. Android SDK API equal or more than 23
      2. Python 3.5- 3.7
      3. Windows 7 or later
      4. Ubuntu 16.04 or later
      5. Android 6.0 or later
      6. NVIDIA GPU drivers CUDA 10.1 requires 418.x or higher
   2. Draw.io

Draw.io is free online diagram tools which allow the user to make flowcharts, process diagram, journey map, UML, network diagram and more.



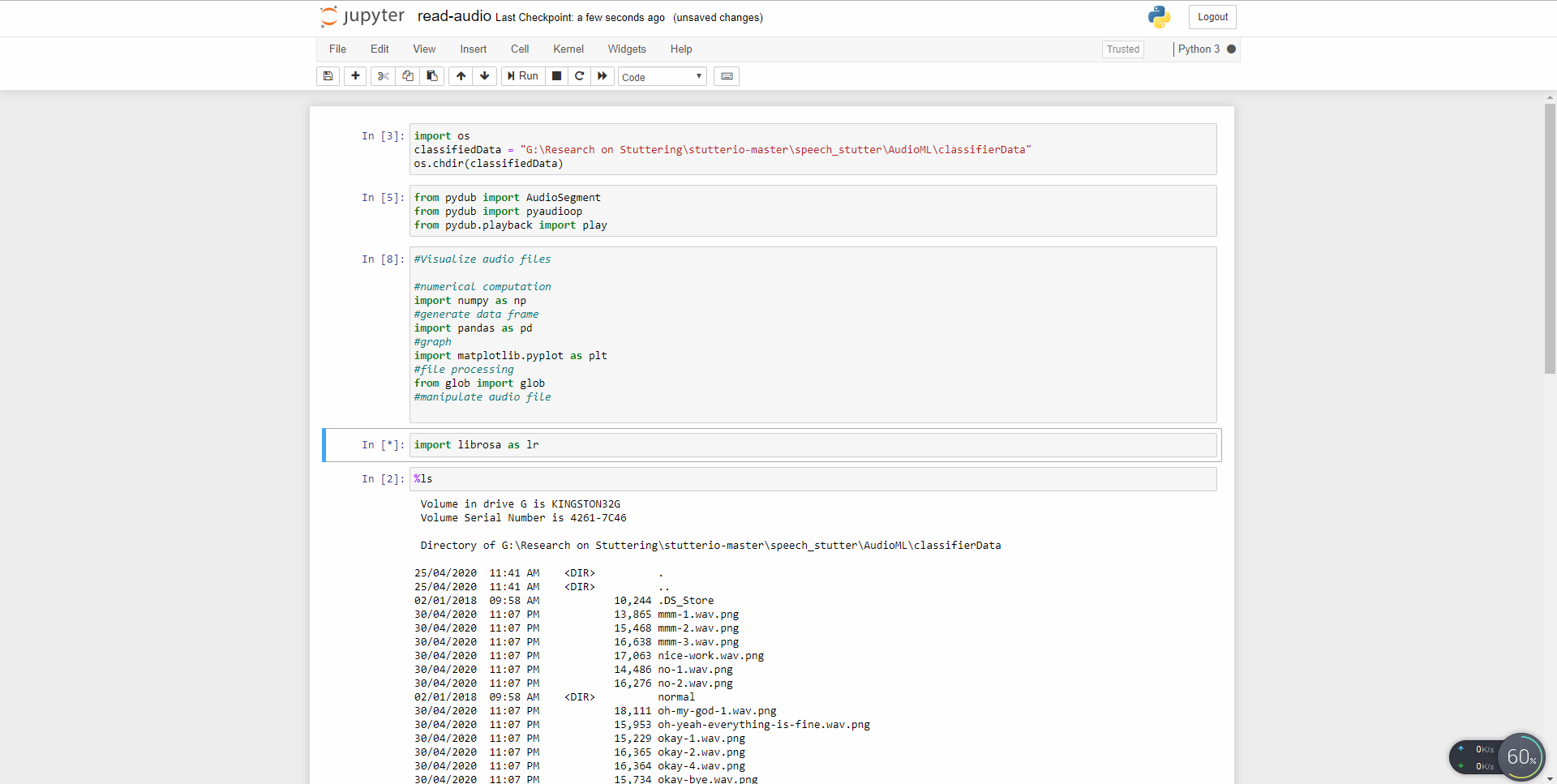
* 1. Anaconda3

Anaconda is a widespread free and open-source distribution of Python and R programming for data science, machine learning, and more. It supports Windows, Linux and macOS. It is easy to use and more reliable and more comfortable to use compared to other distribution. It does support VS Code, JupyterLab, Jupyter Notebook, Powershell Prompt, Qt Console, Spyder, RStudio and a bunch of application and environment.

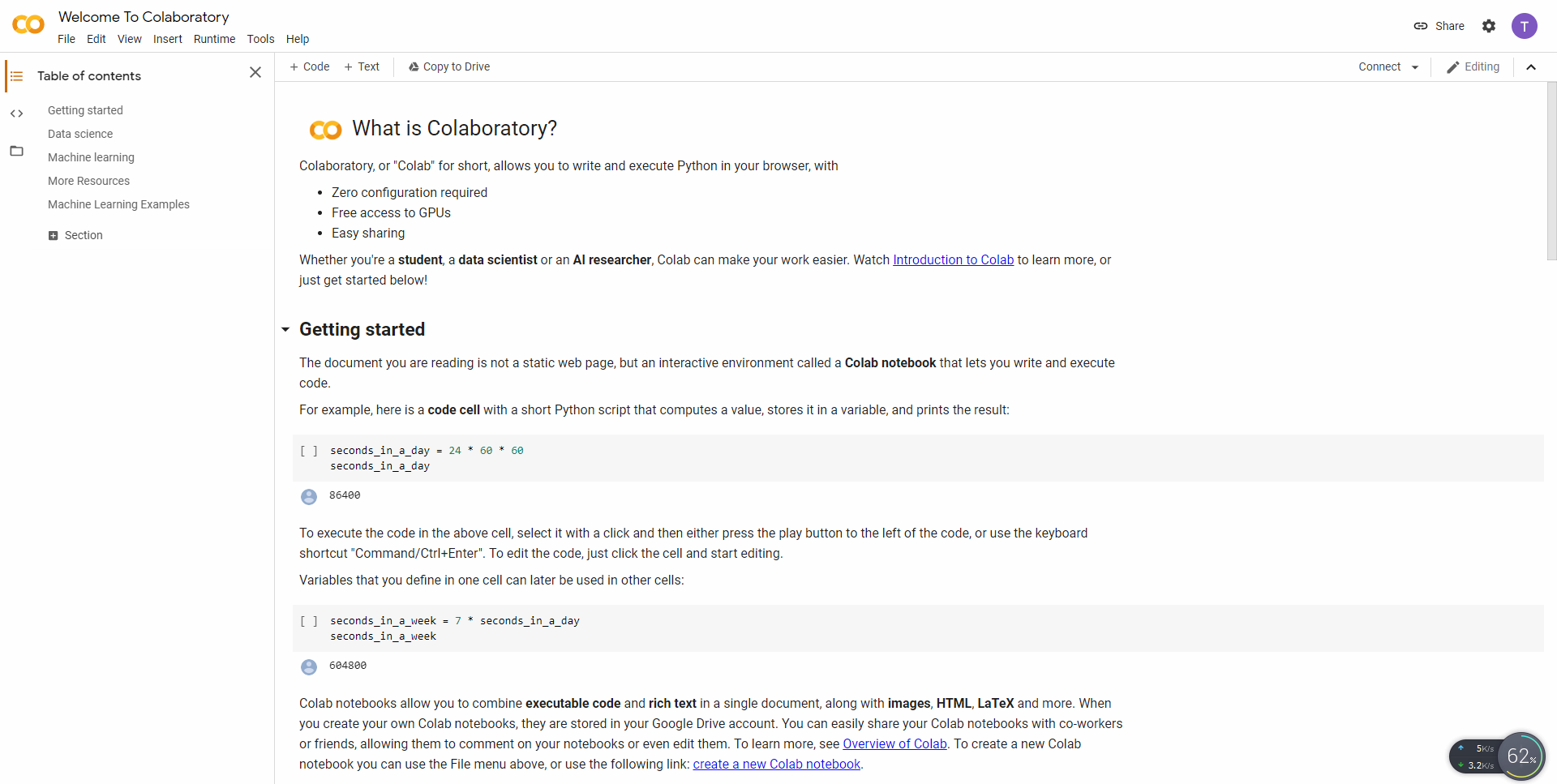


* 1. Jupyter Notebook / Colab

Jupyter Notebook is an opensource web application which allows user to run live code, equations, visualizations and more. It supports a lot of programming language such as Python, Java, R, Octave and more.



Colab is a cloud environment provided by Google which allows user to write and execute their python code in the browser. It provided free access to GPUs and allowed collaborative with other teammates. Colab implements Jupyter Notebook as its user interface.



1. **Framework**
   1. **TensorFlow / TensorFlow lite**

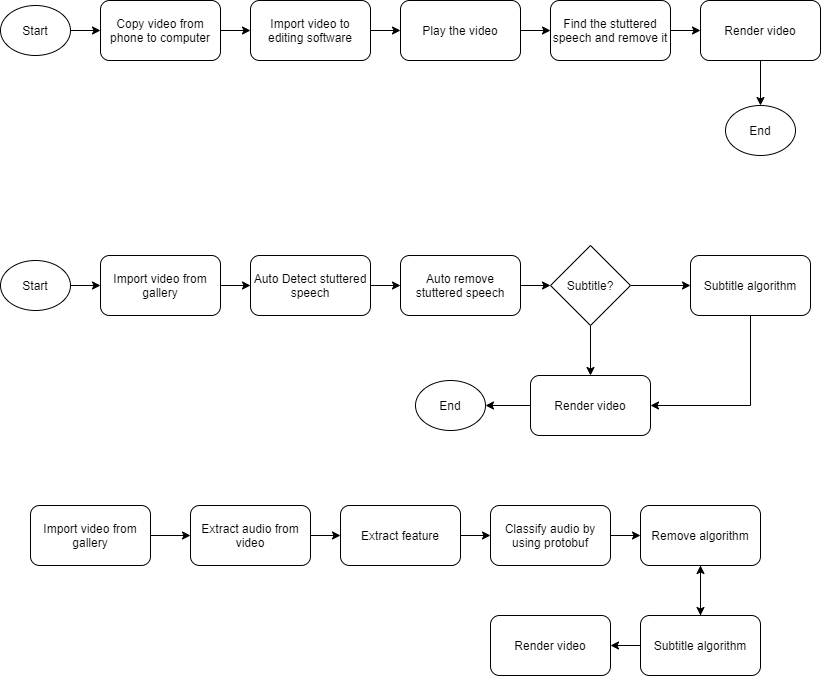
TensorFlow is an open-source platform commonly used in machine learning and deep learning tasks. TensorFlow is developed by Google Brain team to perform the heavy numerical task. TensorFlow provides Python and C++ API interface for the user for execution. TensorFlow uses C and C++ programming language for backend processing to provide faster processing time. It is based on data flow graphs.

* 1. **Flutter**

Flutter is a Google’s UI kit which will compile the application for mobile, web and desktop natively. Flutter is a learn once use anywhere concept. Flutter solved the issues when the developer has to launch an application in two different platforms such as Android and iOS. The traditional method to launch an application in two different platform developer must learn the two programming languages which supported by two different platforms. Apart from that, Flutter also solved the issues caused by different screen size for different mobile devices such as iPad, Android smartphone, iPhone, a screen on refrigerator, touch wall, TV Box and more.

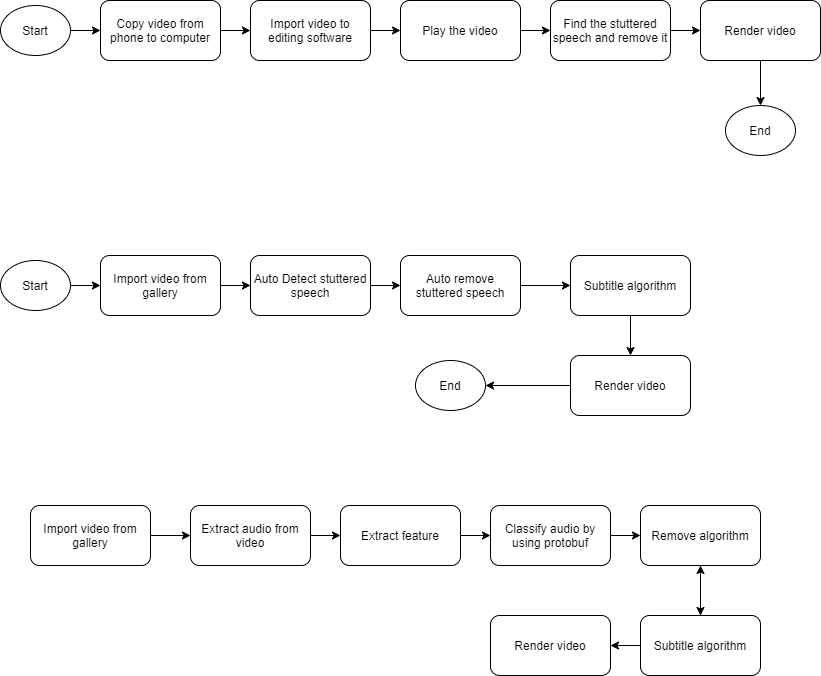
# Journey map

1. Traditional method



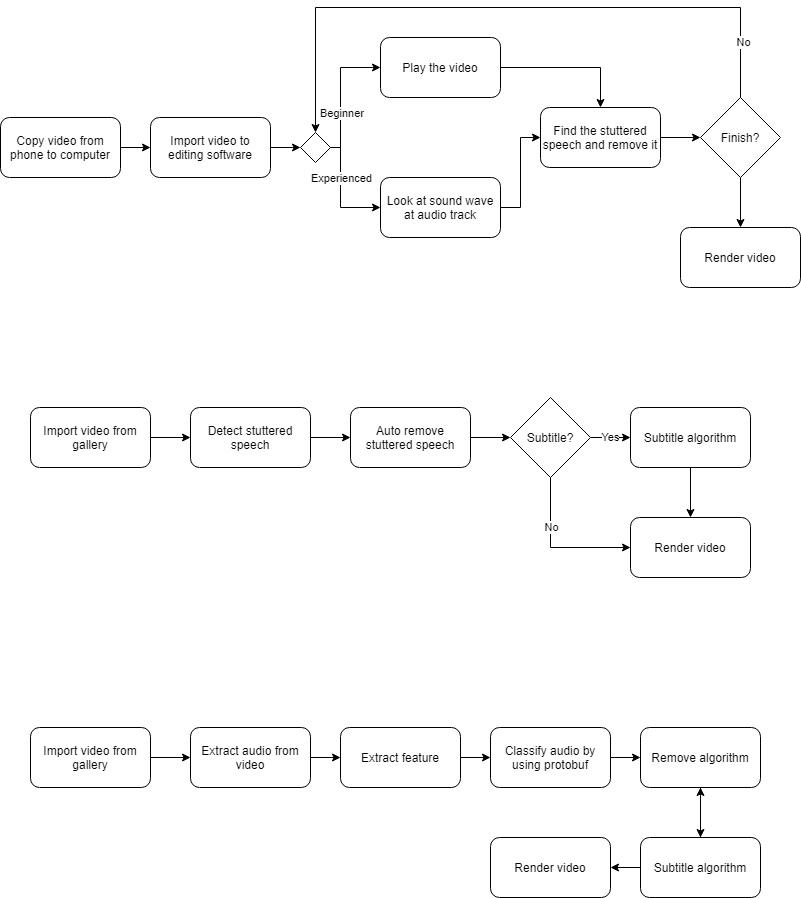
The user will connect their phone to the computer or launch an editing apps on their phone. Next, user will import the footage into the apps and insert into the timeline. User will play the video and try to identify the stuttered part and make a mark. Then, user will select cut tools or trim the video by using hotkey. This process will run over and over again until all the stuttered speech is removed. After all the stuttered speech is removed, user will render the video and save in a mp4 format.

1. New method
   1. External



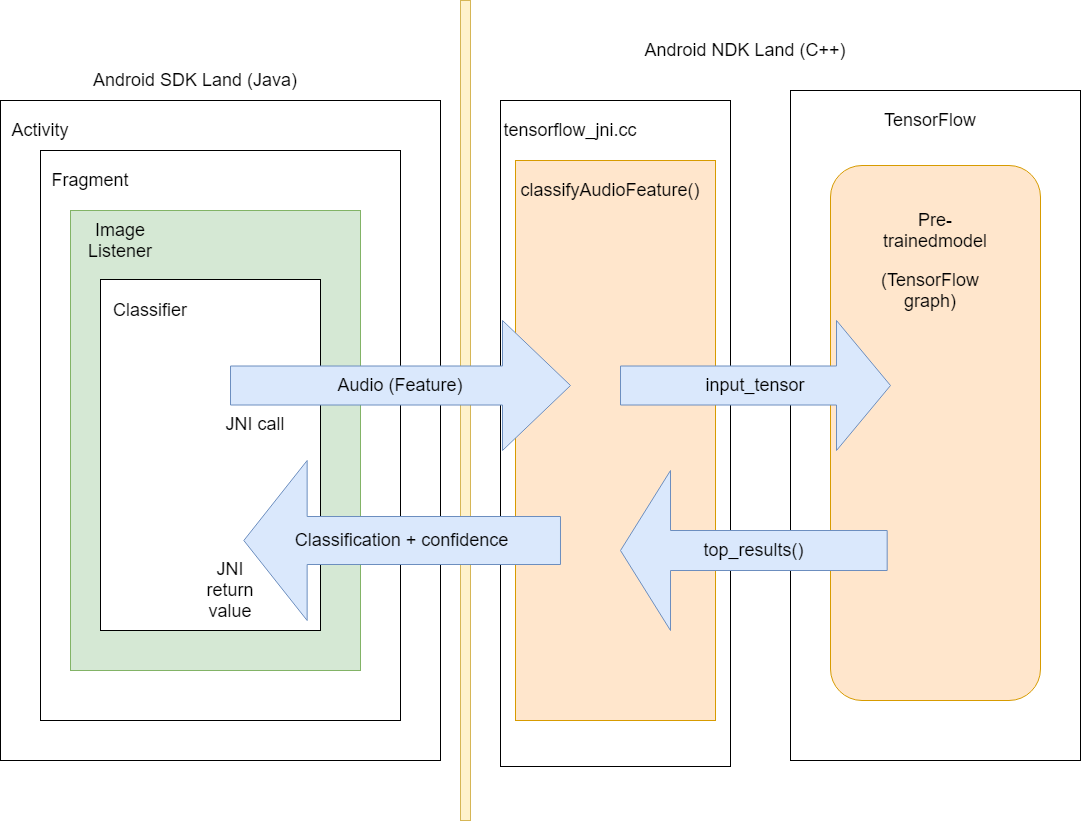
User will import the video from the gallery to our app. After that, the app will automatically identify the stuttered speech and make a mark. Next, the app will remove the stuttered speech automatically. The video will undergo subtitle algorithm and receive the subtitle with a timestamp from API. Then the video will render with the subtitle and save in an mp4 format with minimum human effort.

* 1. Internal



First, the user will import video from gallery. The apps will extract the audio from the video imported. Then the app will extract the feature automatically and classify by using protobuf. The remove algorithm will generate the timestamp need to remove automatically. Then the app will request an API for transcribing purpose. Finally, the video will render with the subtitle and save in an mp4 format.

# System architecture

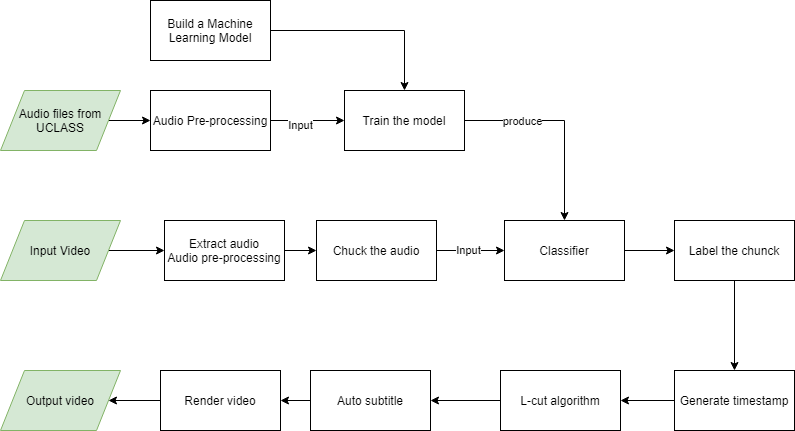


Android SDK stands for Android Software Development Kit. Android NDK stands for Android Native Development Kit, which written in C++. Step 1 is the classifier send the audio features to the tensorflow\_jni.cc class. tensorflow\_jni.cc is a wrapper written in C++ for Android and convert the input to a tensor and resized it. The converted tensor will send to TensorFlow pre-trained model, which call protocol buffer (.pf). Then TensorFlow will return a prediction which is a tensor to the tensorflow\_jni.cc file. Then tensorflow\_jni.cc will return a list of probability values in an array to Android SDK in Java.

# Modules and functionality

|  |  |  |  |
| --- | --- | --- | --- |
| **Module** | **Functionalities** | **Inputs** | **Expected output** |
| Audio pre-processing module | 1. Extract the audio from video 2. Amplify and normalize the audio | Audio files | A uniform formatted audio file |
| Machine learning classification model | 1. Construct an ANN classification model 2. Train the model with the UCLASS data set 3. Save the model | Processed audio files | The threshold for stuttered speech |
| Application | 1. Allow the camera to record video 2. Save the video 3. Deploy model 4. Detect stuttered speech 5. Remove the stuttered part from video 6. Auto subtitle algorithm 7. Render the video | Raw Video | Auto edited video |

# High level system flow



# Milestones

Refer to Appendix A

\*please ignore the working time

**Deliverables**

|  |  |  |
| --- | --- | --- |
| **Deliverables** | **Phase** | **Date** |
| Proposal | Analysis Phase (Planning) | 14/2/2020 |
| Documentation (Chapter 1 to 3) | Analysis Phase (Planning – Define Scope) | 22/5/2020 |
| Half-working prototype (Audio processing module) | Sprint 1 | 10/6/2020 |
| Refined documentation  ANN model | Sprint 2 | 25/6/2020 |
| Final documentation  Final prototype | Sprint 3 | 30/7/2020 |

# Conclusion

In this project, a UCLASS stuttered speech database will undergo pre-processing to improve the quality of the feature. Next, a classification neural network will develop to generate a model by using TensorFlow. The machine learning module will develop by using Python. The dataset will ingest to the neural network to train a model and save it in a special format file protobuf. Next is to develop an app which capable to shot a video and remove the stuttered speech from the video. The app will implement with TensorFlow as a classifier to classify input by using the trained model that saved in protobuf file. In conclusion, this project will significantly reduce human effort and learning curve to remove a stuttered speech from the video and provide a high quality of auto subtitle features.

# References

Celie O’Neil-Hart, Howard Blumenstein. (2016, April). *The latest video trends: Where your audience is watching*. Retrieved from Think With Google: https://www.thinkwithgoogle.com/consumer-insights/video-trends-where-audience-watching/

MathWorks. (n.d.). *Unsupervised Learning*. Retrieved from MathWorks: https://www.mathworks.com/discovery/unsupervised-learning.html

*Mobile & Tablet Android Version Market Share Worldwide*. (2020). Retrieved 7 5, 2020, from statcounter: https://gs.statcounter.com/android-version-market-share/mobile-tablet/worldwide

Stecanella, B. (2017, June 22). *An Introduction to Support Vector Machines (SVM)*. Retrieved from MonkeyLearn: https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/

*stutter*. (n.d.). Retrieved from Cambridge Dictionary: https://dictionary.cambridge.org/dictionary/english/stutter

Swapnil D. Waghmare, Ratnadeep R. Deshmukh, Pukhraj P. Shrishrimal, Vishal B. Waghmare, Ganesh B. Janvale and Babasaheb Sonawane. (2017, June ). A Comparative Study of Recognition Technique Used for Development of Automatic Stuttered Speech Dysfluency Recognition System. *Indian Journal of Science and Technology*, 2-13.

Vikhyath Narayan K N, S P Meharunnisa. (2016). Detection and Analysis of Stuttered Speech. *International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), 5*(4), 953-955.

*Youtube for Press*. (n.d.). Retrieved from Yotube: https://www.youtube.com/about/press/