Speech Recorder and Translator using Google Cloud Speech-to-Text and Translation

[1]J.Y. Chan and [2]H.H. Wang
[1, 2] Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
email: [1]jietying@outlook.com, [2]hhwang@unimas.my

Date received: 4 December 2020
Date accepted: 4 May 2021
Date published: 28 October 2021

Abstract - *The most popular video website YouTube has about 2 billion users worldwide who speak and understand different languages. Subtitles are essential for the users to get the message from the video. However, not all video owners provide subtitles for their videos. It causes the potential audiences to have difficulties in understanding the video content. Thus, this study proposed a speech recorder and translator to solve this problem. The general concept of this study was to combine Automatic Speech Recognition (ASR) and translation technologies to recognize the video content and translate it into other languages. This paper compared and discussed three different ASR technologies. They are Google Cloud Speech-to-Text, Limecraft Transcriber, and VoxSigma. Finally, the proposed system used Google Cloud Speech-to-Text because it supports more languages than Limecraft Transcriber and VoxSigma. Besides, it was more flexible to use with Google Cloud Translation. This paper also consisted of a questionnaire about the crucial features of the speech recorder and translator. There was a total of 19 university students participated in the questionnaire. Most of the respondents stated that high translation accuracy is vital for the proposed system. This paper also discussed a related work of speech recorder and translator. It was a study that compared speech recognition between ordinary voice and speech impaired voice. It used a mobile application to record acoustic voice input. Compared to the existing mobile App, this project proposed a web application. It was a different and new study, especially in terms of development and user experience. Finally, this project developed the proposed system successfully. The results showed that Google Cloud Speech-to-Text and Translation were reliable to use in video translation. However, it could not recognize the speech when the background music was too loud. Besides, it had a problem of direct translation, which was challenging. Thus, future research may need a custom trained model. In conclusion, the proposed system in this project was to contribute a new idea of a web application to solve the language barrier on the video watching platform.*

Keywords*:* Speech Recognition, Google Speech-to-Text; ASR; Google Cloud Translation;

1 Introduction

YouTube is the second most visited website in the year 2019 (Top Websites Ranking, 2019). The content on YouTube is localized in over 100 countries and can be accessed using 80 different languages (YouTube Offical Blog, 2019). There is a total of one billion hours of video viewed on YouTube every day worldwide. Hence, watching videos online has become an essential daily routine for many people who communicates in different languages. However, not all online videos provide multilingual subtitles. As a result, the audiences who speak a different language from the video could not understand the content of the video. Thus, this study implemented research and development on a web-based speech recorder and translator for the video watching platform.

To build a web-based speech recorder and translator, it uses Automatic Speech Recognition (ASR) technology. The process of Automatic Speech Recognition includes receiving audio input, analyzing the input patterns, and providing a text output (Lai & Yankelovich, 2007). There are diverse ASR applications nowadays. Based on an evaluation report that compares Google Cloud Speech-to-Text, Limecraft Transcriber, and VoxSigma, Google Cloud Speech-to-Text has the lowest Word Error Rate (WER) (Santo, 2017). There is also another research of android-based Google Cloud Speech-to-Text, specifically for the speech-impaired person (Anggraini et al., 2017). The result shows that the success rate of recognition for a normal voice is 100%. As for the speech-impaired person, the success rate is in between 83.3% to 90% which is considered pretty high. Based on these previous researches, Google Cloud Speech-to-Text is reliable to use in any application. Knowing the WER from existing studies, this paper discussed other aspects of the Speech-to-Text Application, such as client libraries and the number of languages supported. Instead of an android-based application in existing projects, this project proposed a new web-based system that will recognize acoustic input from a video. Hence, the main objective of this project is to design and develop a web-based speech recorder and translator for the video watching platform. The secondary goal of this project is to evaluate the usability of Google Cloud Speech-to-Text and Translation API when they face background music and accent challenges.

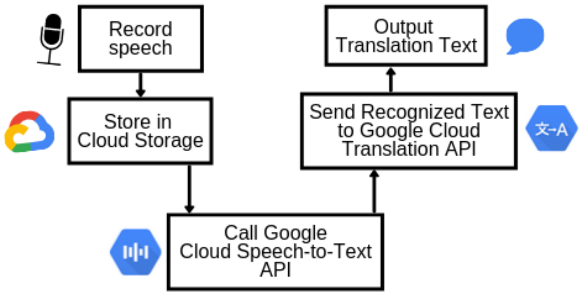

Record speech
Store in Cloud Storage
Call Google Cloud Speech-to-Text API
Send Recognized Text to Google Cloud Translation API
Output Translation Text

Figure 1. Overall Flow of Speech Recorder and Translator Application

Finally, the purposes achieved in this study. The web-based speech recorder and translator successfully developed by using Microsoft Visual Studio 2019. Google Cloud Speech-to-Text is applied to recognize the language in audio and convert the audio to text transcription. It uses Automatic Speech Recognition (ASR) to receive acoustic input through a microphone and produce a text output after some analysis by algorithms or models (Levis & Suvorov, 2013). Then, Google Translation API will translate the text output to the user's selected language. Figure 1 shows the overall flow of the speech recorder and translator application. The Translation API uses a pre-trained machine learning model to translate between languages. Besides, this paper discussed the evaluation results of the usability of Google Cloud Speech-to-Text and Translation API. The results showed that Google Cloud Speech-to-Text could not perform well when there is loud background music on a video. The word recognition is also challenging when the speaker in the video has a strong accent. Thus, future work may be needed, such as training a customized model for a specific accent. All in all, this study is a good start for future work in this area.

2. Related works

There was an evaluation report by (Santo, 2017) that compared the word error rate between Google Cloud Speech-to-Text, Limecraft Transcriber, and VoxSigma. It used a 60 minutes audio file from the European Organization for Nuclear Research (CERN) produced meetings as the audio input. It had a Python program to perform the actual measurement of time and accuracy. Table 1 shows the example of word error rate calculation whereby “S” represents substitution, “I” is an insertion, “D” is deletion and the symbol “=” means match. According to the research by (Santo, 2017), Google Cloud Speech-to-Text has the lowest word error rate. However, comparing word error rates is not enough for speech recognition on video watching platforms. Thus, this paper makes improvements by comparing other features such as client libraries, real-time streaming, and pricing. Table 2 below shows the additional features comparison between Google Cloud Speech-to-Text, Limecraft Transcriber, and VoxSigma.

TABLE 1. EXAMPLE OF CALCULATION FOR WORD ERROR RATE (WER)

Reference	this	is	A	paper	about	ASR	
	=	D	S	=	=	=	I
Hypothesis	this		Of	paper	about	ASR	system

TABLE 2. COMPARISON ON EXISTING SPEECH RECOGNITION APPLICATIONS

Features/ Application	**Google Cloud Speech-to-Text**	**Limecraft Transcriber**	**VoxSigma**
Request	• REST API • RPC API	REST API	REST API
Client libraries	C#, GO, Java, PHP, Python, Node.js, Ruby	No	C/C++, PHP, Java, JavaScript
Supported Languages	120	100+	15
Real-time streaming	Yes	No	Yes
Language identification	Yes	No	Yes
Support	Stack Overflow, Google Group, Support Package	Hotline Support	Hotline Support
Pricing	• Free for first 60 minutes, $0.024 per minute • One year free trial with $300 credit, 60 minutes per month	$109 per month	• $0.01 per minute • Offer free trial upon request
Word Error Rate from Previous Research	14%	45%	60%
Latency	3885 / 60 mins	N/A	4994 / 60 mins

Based on Table 1, Google Cloud Speech-to-Text has the best overall features. It enables developers to send a request through two different APIs (REST and RPC), while Limecraft Transcriber and VoxSigma only allow integration through REST API. Besides, Google Cloud Speech-to-Text offers the highest number of client libraries. It indicates that Google Cloud Speech-to-Text accommodates a wide range of development alternatives for the developers. Most importantly, it allows real-time streaming of audio input from the microphone, which suits this project requirement well. Furthermore, it offers a one-year free trial of its features. Thus, for speech recognition on video watching platforms, it is believed that Google Cloud Speech-to-Text is the most suitable application to be used.

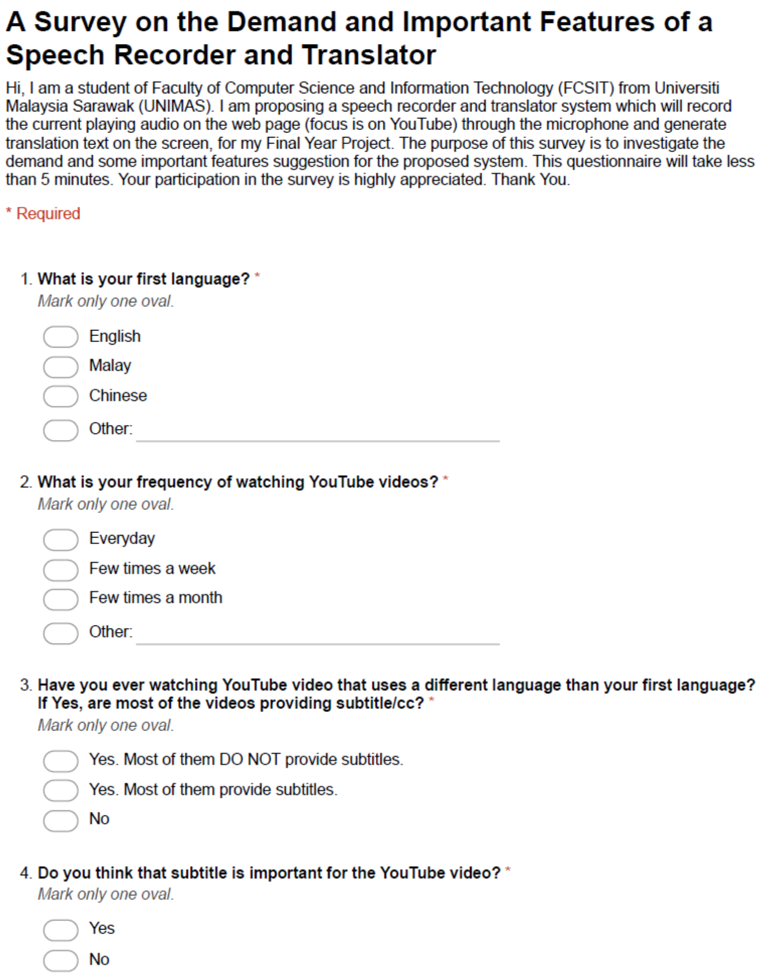
Besides evaluation research by (Santo, 2017), there is another related study accomplished by (Anggraini et al., 2018). It is a research of android-based Google Cloud Speech-to-Text, specifically for the speech-impaired person. The main topic of the study is to compare the recognition rate of Google Cloud Speech-to-Text when the audio input is normal voice or speech-impaired voice. The result shows that the success rate of recognition for a normal voice is 100%. As for the speech-impaired person, the success rate is in between 83.3% to 90% which is considered high success rate. However, there is a limitation that the android app could not automatically record acoustic input from a playing video. Thus, this project improves the function by using a device microphone to record the audio from the current playing video. Another limitation is that only an android mobile phone can use the app. Hence, this project also improves by creating a web-based application so that any device can use the application.

3. Methodology and Design

This project used agile system development methodology throughout the development process. By adopting an agile system development methodology, the developer continuously plans, learns, develops, and delivers the system to the targeted users (Tatvasoft, 2015). There are several modules for the targeted users to test from phase to phase. Section 4 discussed the detail of what happened during the testing phase in agile system development process.

3.1 Requirement Analysis

Before developing the proposed system, there is a questionnaire to obtain opinions and demands on a speech recorder and translator from the targeted individuals. There is a total of 10 multiple choices questions in the questionnaire. The project consists of 19 respondents who are students from different universities or colleges. The questionnaire uses Google Form to create and deliver to the respondents, as shown in Figure 2. Figures 3 to figure 12 are the results of each question.



A Survey on the Demand and Important Features of a Speech Recorder and Translator

Hi, I am a student of Faculty of Computer Science and Information Technology (FCSIT) from Universiti Malaysia Sarawak (UNIMAS). I am proposing a speech recorder and translator system which will record the current playing audio on the web page (focus is on YouTube) through the microphone and generate translation text on the screen, for my Final Year Project. The purpose of this survey is to investigate the demand and some important features suggestion for the proposed system. This questionnaire will take less than 5 minutes. Your participation in the survey is highly appreciated. Thank You.

* Required

1. **What is your first language?** *
   Mark only one oval.
   ○ English
   ○ Malay
   ○ Chinese
   ○ Other: ____________________

2. **What is your frequency of watching YouTube videos?** *
   Mark only one oval.
   ○ Everyday
   ○ Few times a week
   ○ Few times a month
   ○ Other: ____________________

3. **Have you ever watching YouTube video that uses a different language than your first language? If Yes, are most of the videos providing subtitle/cc?** *
   Mark only one oval.
   ○ Yes. Most of them DO NOT provide subtitles.
   ○ Yes. Most of them provide subtitles.
   ○ No

4. **Do you think that subtitle is important for the YouTube video?** *
   Mark only one oval.
   ○ Yes
   ○ No

Figure 2. Questionnaire Google Form

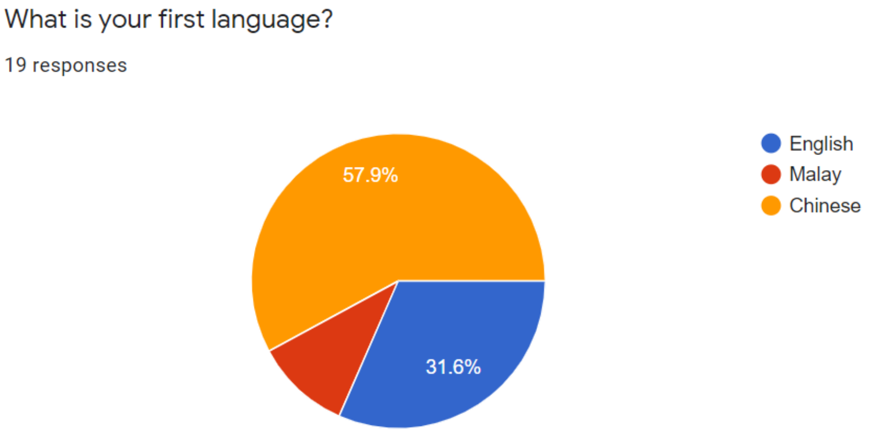


Figure 3. Question 1 of the Questionnaire: Respondents' first language

From Figure 3, all of the respondents used different first languages. Among 19 respondents, 57.9% (11) of them use Chinese as their first language, followed by 31.6% (6) use English and 10.5% (2) use Malay.

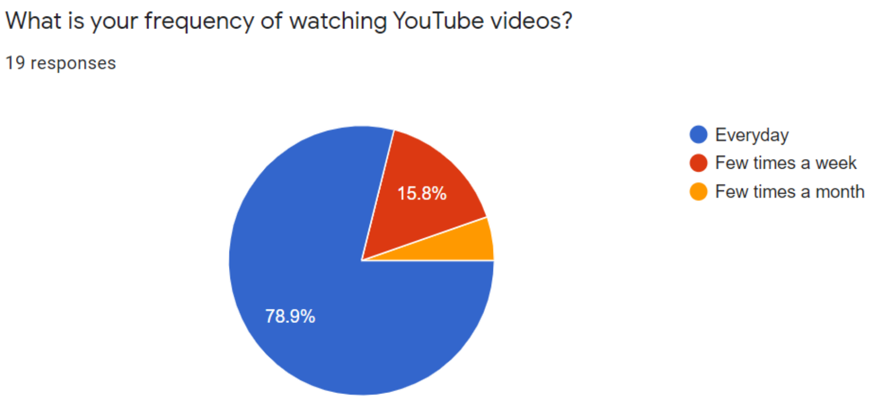


Figure 4. Question 2 of the Questionnaire: Frequency of watching YouTube videos

Figure 4 shows question 2 of the questionnaire, which aims to collect the frequency of the respondents watching YouTube videos. A high number of respondents (78.9% or 15 of them) watch YouTube videos every day. Only one respondent (5.3%) watches YouTube videos a few times a month. It means that the targeted users can test the proposed system on YouTube because most of them watch videos on the platform daily.

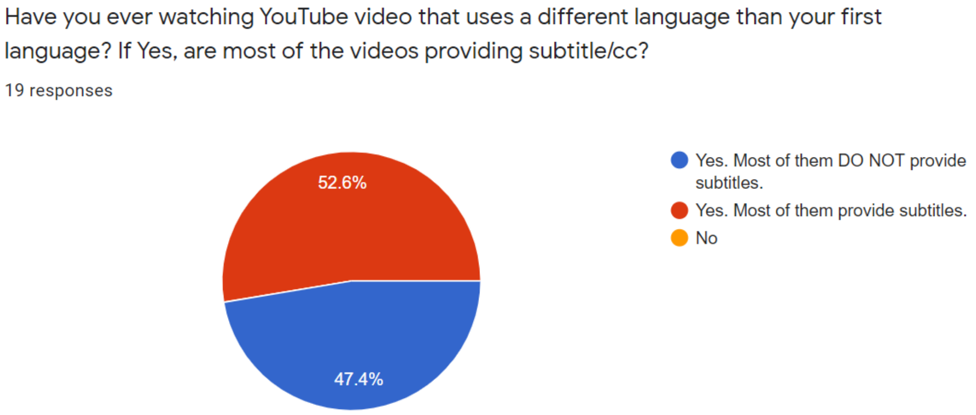


Figure 5. Question 3 of the Questionnaire: Video Subtitles

Figure 5 shows question 3 of the questionnaire, which collects data about the chances of the respondents watching a video that uses a different language than their first language and the probability of the videos providing subtitles. The results show that all of the respondents watch a video that speaks the non-mother language. Among the videos, 52.6% of them provide subtitles, while 47.4% do not provide subtitles. To conclude, the average probability of the video providing subtitles is almost half (50%).

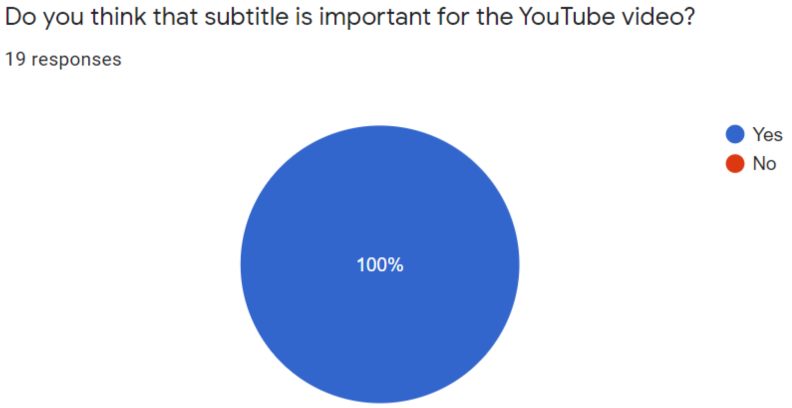


Figure 6. Question 4 of the Questionnaire: Importance of Subtitles

Figure 6 shows question 4 of the questionnaire. After obtaining the probability of video providing subtitles from the previous question, the questionnaire proceeds to gain opinions about the importance of subtitles for the videos.

All of the respondents think that it is vital for the YouTube video to provide subtitles. From the result, the speech recorder and translator gained high demand from the respondents.

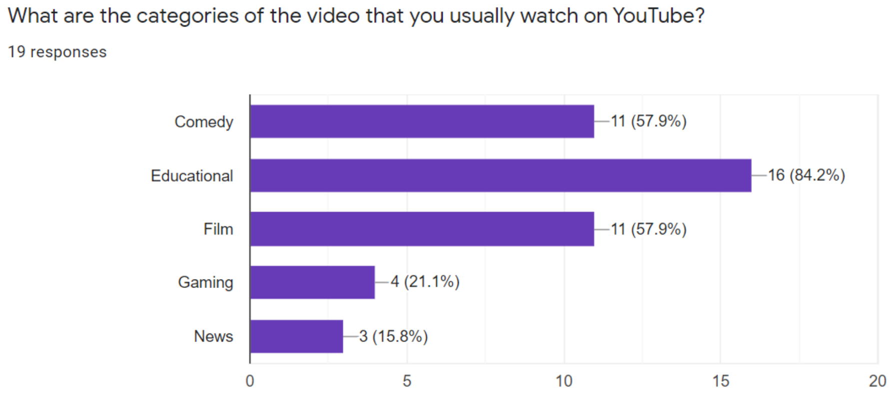


Figure 7. Question 5 of the Questionnaire: Categories of Watched Videos

Question 5 of the questionnaire collects the video categories that the respondents watch frequently. The result shows that most of the respondents watch educational videos. There is a total of 16 of them (84.2%) watch educational videos. The estimated reason for the result is that all of the respondents are students who often watch educational videos to learn for their studies. From this result, an educational video is a decent choice to test the proposed system.

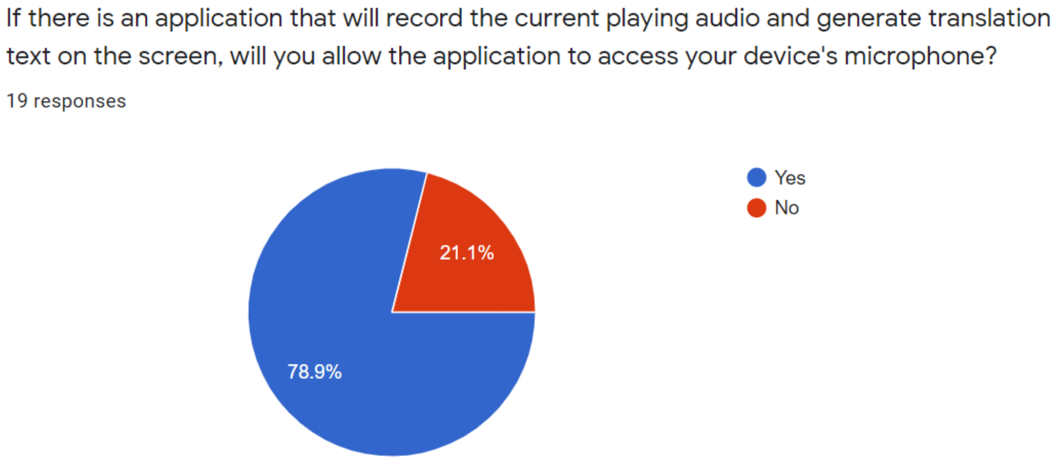


Figure 8. Question 6 of the Questionnaire: Willingness to Allow Access Device’s Microphone

Figure 8 illustrates the summary of responses for question 6 of the questionnaire, which is the willingness of the respondents to allow the system to access their device’s microphone. 78.9% of the respondents are willing to allow the system to access their microphone. 21.1% of them are reluctant to allow the system to access their device’s microphone. From the result, the application is acceptable and useable for most of the respondents.

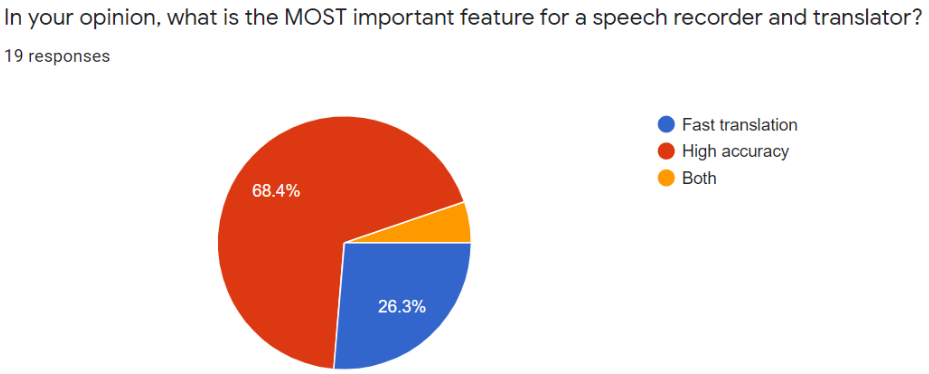


Figure 9. Question 7 of the Questionnaire: Important Feature of the Proposed System

Question 7 of the questionnaire collects data regarding the essential features of the proposed system, as shown in Figure 9. 68.4% of the respondents think that high accuracy is more significant than fast translation. 26.3% of the respondents argue that fast translation is more vital than high accuracy. Besides, one of the respondents has a different perspective whereby he or she feels that both fast translation and high accuracy are necessary. Thus, the proposed system could pay more attention to high accuracy than fast translation while selecting the translation service.

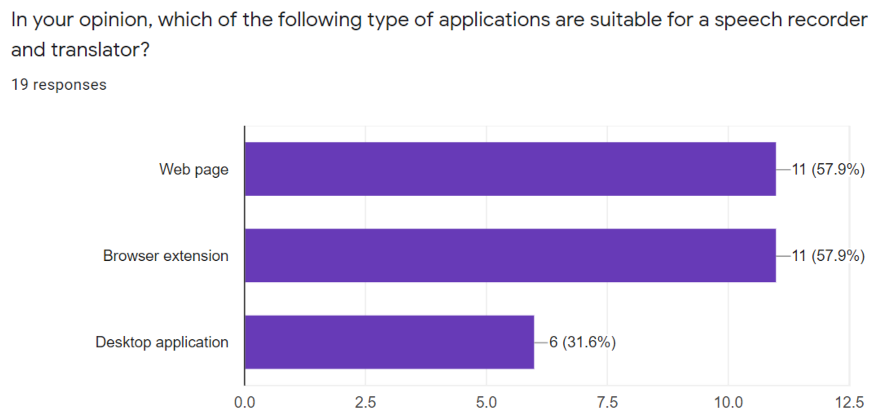


Figure 10. Question 8 of the Questionnaire: Type of application for speech recorder and translator

Figure 10 shows the summary of the responses for question 8 of the questionnaire. It aims to collect some opinions about the type of applications suitable for the speech recorder and translator. From the result, web page and browser extension receive the same amount of vote whereby 57.9% of the respondents choose web page and browser extension. In contrast, only 6 of the respondents think that desktop application is suitable for the proposed system. Hence, a web page or browser extension is acceptable as the proposed system.

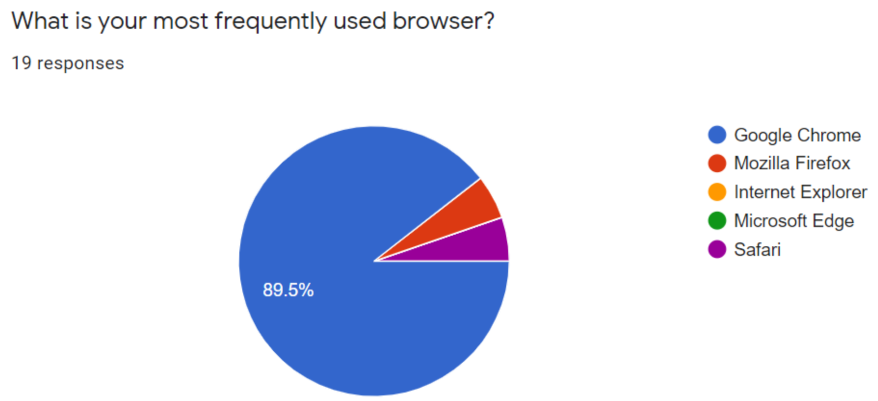


Figure 11. Question 9 of the Questionnaire: Respondents' Browser

Figure 11 illustrates the primary browser used by the respondents. 89.5% of the respondent use Google Chrome browser. 5.3% of them use Mozilla Firefox, and 5.3% of them use Safari. The data collected from this question is significant for the testing phase. From the result in Figure 3.10, the developer can focus on testing the proposed system by using Google Chrome.

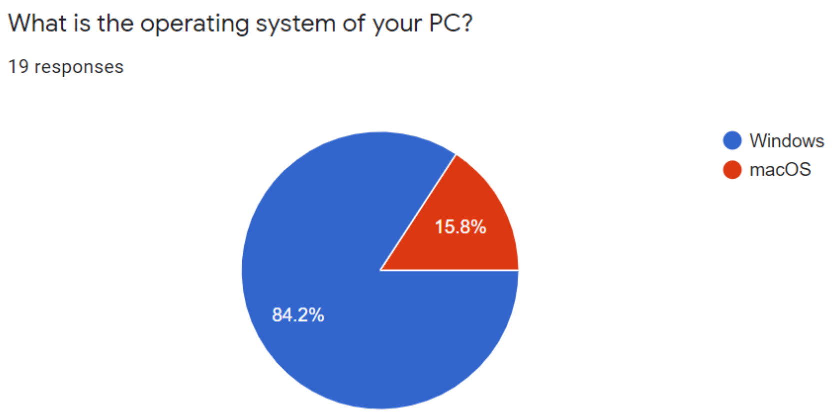


Figure 12. Question 10 of the Questionnaire: Respondents' PC Operating System

Figure 12 displays a summary of respondents' PC operating systems. 84.2% of the respondents use the Windows operating system, while 15.8% of them use macOS. This data is also vital for the testing phase because the developer could pay attention to one of the operating systems. In this case, the developer can focus on the Windows operating system.

3.1.1 Questionnaire Summary

To sum up the questionnaire results, there is a high number of respondents watch YouTube videos daily. All of them think that subtitles in a video are very significant. However, among 19 respondents, half of them respond that the videos on YouTube did not provide subtitles. Most of the respondents watch educational videos. Thus, the developer could test the proposed system on education videos from YouTube. There is a high number of respondents willing to allow the system to access their device's microphone. It means that the acceptance of the proposed application to use their device microphone is high. For the significant feature of the proposed system, 68.4% of the respondents think that high accuracy is more prominent than fast translation. Hence, the developer could pay attention to this feature while developing and testing the proposed system. The respondents also reflect that it is suitable to use a web page as a speech recorder and translator. They also accept a combination of a web page and browser extension. During the development and testing phase, the developer could focus on the compatibility of the proposed system in the Windows operating system and Google Chrome browser because most of the respondents use them.

3.2 System design

Figure 13 shows the cross-functional flowchart of the proposed system in which the processes are simple. Firstly, the user will select a recording language and translation text output language. There is no automatic language detection in the proposed system because the language identification feature is expensive. Thus, the user will select the language by themselves. Then, the browser will prompt whether to allow access to the microphone. If the user grants access, Google Cloud Speech-to-Text API will record the streaming audio and produce the text output. Next, Google Cloud Translation will receive the text and translate it into the translation language selected by the user. If the user does not grant access to the microphone, there is no translation output produced on the screen. The whole process of how the speech recorder and translator works is simple and easy to manage.

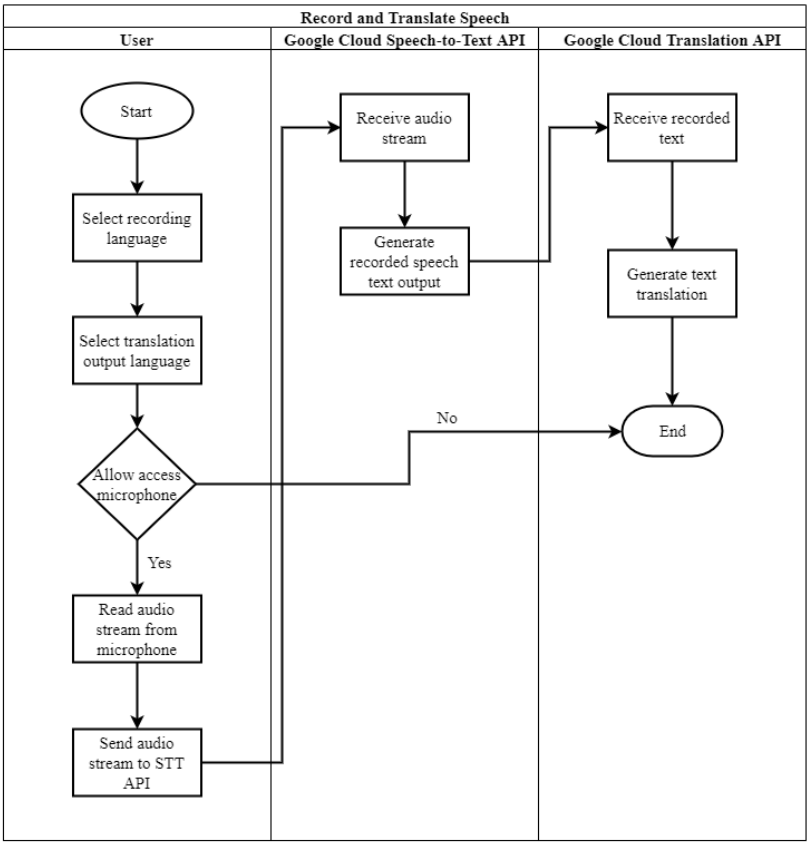
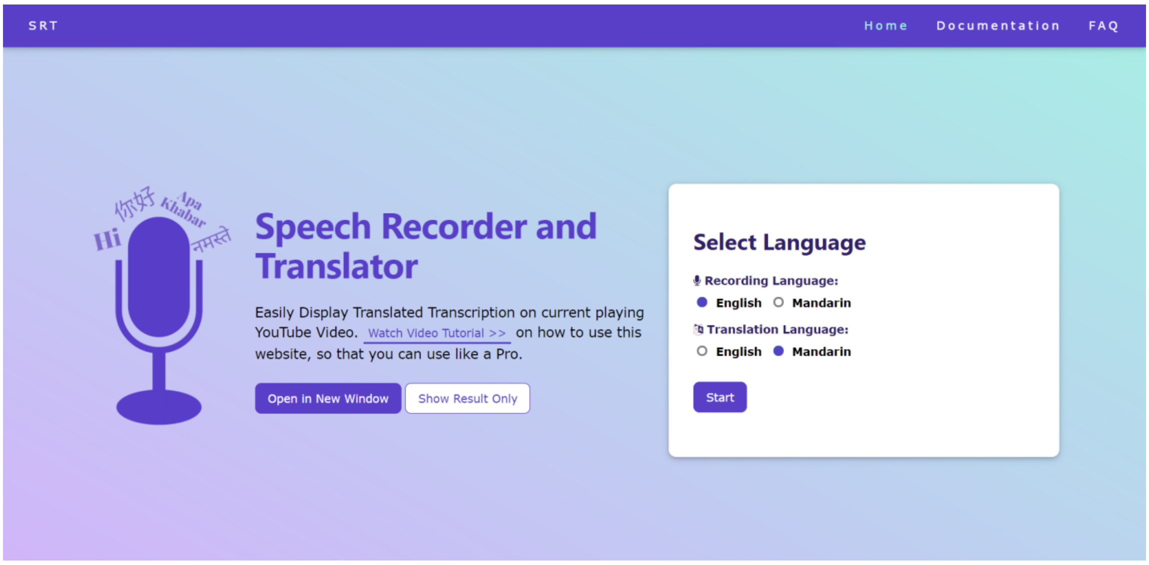

Record and Translate Speech
User
Google Cloud Speech-to-Text API
Google Cloud Translation API
Start
Select recording language
Select translation output language
Allow access microphone
Yes
No
Read audio stream from microphone
Send audio stream to STT API
Receive audio stream
Generate recorded speech text output
Receive recorded text
Generate text translation
End

Figure 13. Cross-Functional Flowchart of the Proposed System

3.3 User Interface

The following figures show the user interfaces of the proposed system. It consists of three parts which are the home page, documentation, and also FAQ.



Speech Recorder and Translator

Easily Display Translated Transcription on current playing YouTube Video. Watch Video Tutorial >> on how to use this website, so that you can use like a Pro.

Select Language

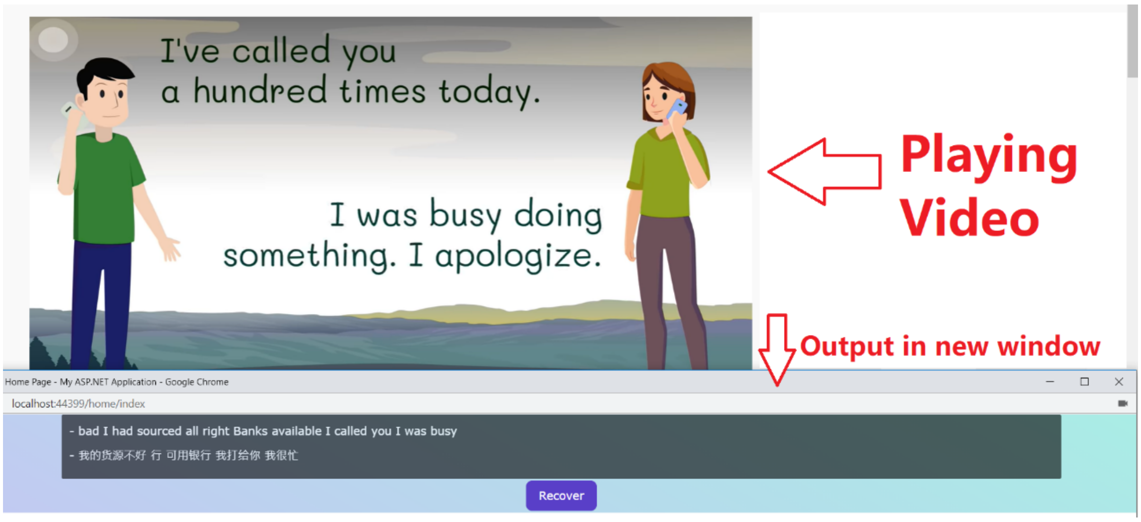
Recording Language:

● English ○ Mandarin

Translation Language:

○ English ● Mandarin

Figure 14. Home Page



I've called you a hundred times today.

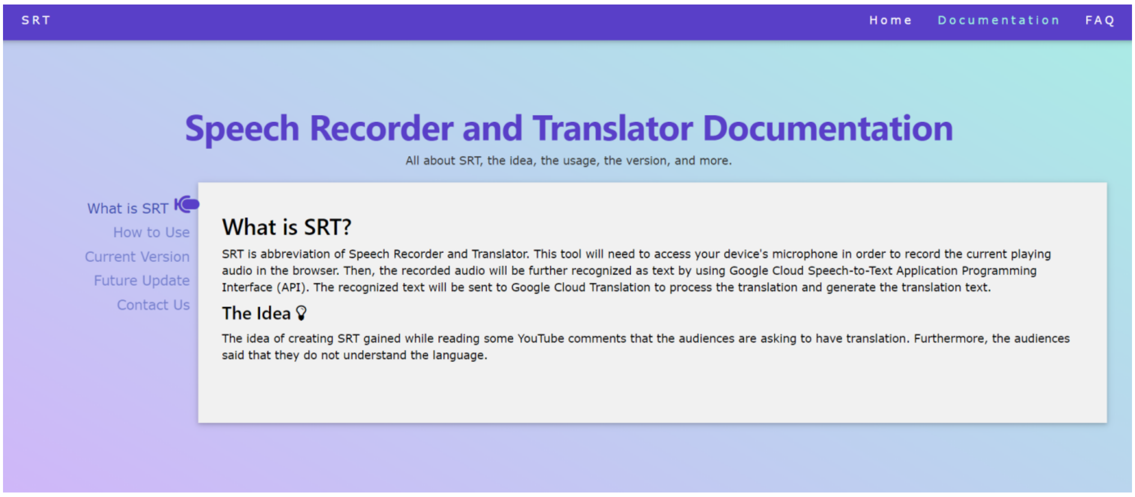
I was busy doing something. I apologize.

Playing Video
Output in new window
Home Page - My ASP.NET Application - Google Chrome
localhost:44399/home/index
Recover

- bad I had sourced all right Banks available I called you I was busy
- 我的货源不好 行 可用银行 我打给你 我很忙

Figure 15. Translation Text Output in New Window

Figure 14 shows the home page of the proposed system. There are only two selections of languages. It is because the scope of the project is focusing on translation between English and Mandarin. When the user clicks on the Start button, there will be translation output displayed on top of the page. There are also "Open in New Window" and "Show Result Only" buttons, whereby the users can open the application in a new window so that the users can see the translated results and watch the video together, as shown in Figure 15.



Speech Recorder and Translator Documentation

All about SRT, the idea, the usage, the version, and more.

What is SRT?

SRT is abbreviation of Speech Recorder and Translator. This tool will need to access your device's microphone in order to record the current playing audio in the browser. Then, the recorded audio will be further recognized as text by using Google Cloud Speech-to-Text Application Programming Interface (API). The recognized text will be sent to Google Cloud Translation to process the translation and generate the translation text.

The Idea

The idea of creating SRT gained while reading some YouTube comments that the audiences are asking to have translation. Furthermore, the audiences said that they do not understand the language.

Figure 16. Documentation Page

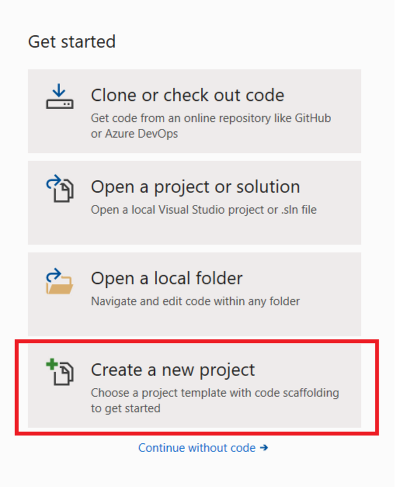
Figure 16 shows the user interface of the documentation of the system. Generally, every digital product needs to have documentation to guide the users on using the application. Thus, the proposed system included descriptive and informative documentation. The left vertical menu enables the user to have quick access and easy navigation to other sections. It allows the user to skip the current part and read other topics easily. The details on the documentation page include application concept explanation, application usage, versions, future updates, and contact us.

3.4 Development

The developer needs to install and configure the essential tools properly before starting to develop the application. The tools include Google Cloud Platform account, Microsoft Visual Studio Community 2019, and Microsoft Azure DevOps.

A Google Cloud Platform (GCP) account is needed to use Google Cloud Speech-to-Text and Google Cloud Translation services. Thus, the first step is to get a GCP account registered. After successfully signing up, the developer needs to create a console project and enable Cloud Speech-to-Text API and Cloud Translation API for that project. Then, from the left navigation menu, go to "IAM & Admin" and "Service Account" to create a service account and get a private key. Next, download the key as a JSON file. The location of the JSON file is needed to use in the coding part later.

Microsoft Visual Studio is used to develop the proposed web application. Firstly, the developer needs to download and install Microsoft Visual Studio Community 2019 from the official site. While installing the software, when it comes to the selecting workloads section, select only Asp.Net and web development and .NET desktop development. These two workloads are needed to use Microsoft Visual Studio to develop Asp.Net web application. After successfully installing the software, open Visual Studio and create a new project as shown in Figure 17.



Get started

Clone or check out code
Get code from an online repository like GitHub or Azure DevOps

Open a project or solution
Open a local Visual Studio project or .sln file

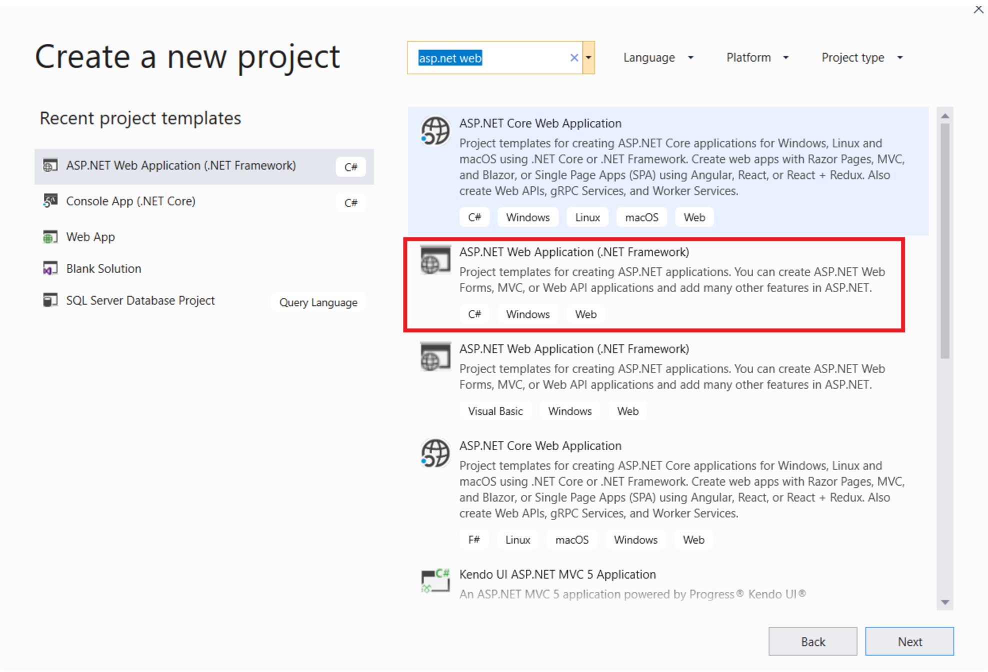
Open a local folder
Navigate and edit code within any folder

Create a new project
Choose a project template with code scaffolding to get started

Continue without code →

Figure 17. Create New Project in Visual Studio

Next, search for "Asp.Net Web Application" and select the one with the "C#" tag, as shown in Figure 18. Then, enter the project name, location to store the project and .NET framework. As displayed in Figure 19, for the .NET framework, this project uses .NET 4.6.1.



Create a new project

asp.net web

Recent project templates

ASP.NET Web Application (.NET Framework)	C#
Console App (.NET Core)	C#
Web App	
Blank Solution	
SQL Server Database Project	Query Language

ASP.NET Core Web Application
Project templates for creating ASP.NET Core applications for Windows, Linux and macOS using .NET Core or .NET Framework. Create web apps with Razor Pages, MVC, and Blazor, or Single Page Apps (SPA) using Angular, React, or React + Redux. Also create Web APIs, gRPC Services, and Worker Services.
C# Windows Linux macOS Web

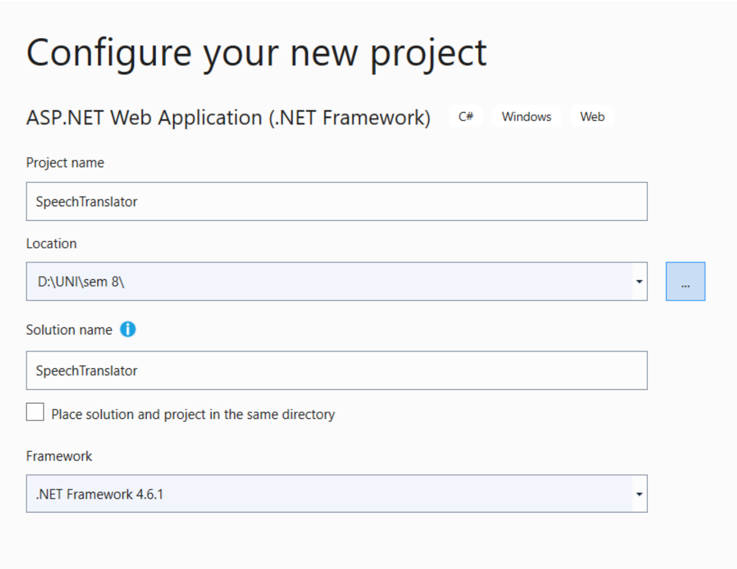
ASP.NET Web Application (.NET Framework)
Project templates for creating ASP.NET applications. You can create ASP.NET Web Forms, MVC, or Web API applications and add many other features in ASP.NET.
C# Windows Web

ASP.NET Web Application (.NET Framework)
Project templates for creating ASP.NET applications. You can create ASP.NET Web Forms, MVC, or Web API applications and add many other features in ASP.NET.
Visual Basic Windows Web

ASP.NET Core Web Application
Project templates for creating ASP.NET Core applications for Windows, Linux and macOS using .NET Core or .NET Framework. Create web apps with Razor Pages, MVC, and Blazor, or Single Page Apps (SPA) using Angular, React, or React + Redux. Also create Web APIs, gRPC Services, and Worker Services.
F# Linux macOS Windows Web

Kendo UI ASP.NET MVC 5 Application
An ASP.NET MVC 5 application powered by Progress® Kendo UI®

Figure 18. Select Type of Project



Configure your new project

ASP.NET Web Application (.NET Framework) C# Windows Web

Project name: SpeechTranslator

Location: D:\UNI\sem 8\

Solution name: SpeechTranslator

□ Place solution and project in the same directory

Framework: .NET Framework 4.6.1

Figure 19. Configure Project Environment

The next step is to install client libraries for Google Cloud Speech-to-Text and Translation in Microsoft Visual Studio 2019. Go to "Tool" tab, then Nuget Package Manager, then open Package Manager Console and type the following commands.

```
Install-Package Google.Cloud.Speech.V1 –Pre
Install-Package NAudio -Version 1.10.0
Install-Package Google.Cloud.Translation.V2
```

Google.Cloud.Speech.V1 is a client library for using Google Cloud Speech-to-Text service. NAudio is to receive the streaming audio from the microphone and send the audio to Google Cloud Speech-to-Text. Google.Cloud.Translation.V2 is a client library for using Google Cloud Translation in Asp.Net C# MVC.

After setting up the project in Visual Studio, the developer needs to register an account on Microsoft Azure DevOps official site. Then, log in to the account and create a new project in Azure DevOps. Set a project name, for example, Speech recorder and translator project. For the Process options, select the agile strategy. After that, go to Microsoft Visual Studio, open Team Explorer, and click on the green cable icon as shown in Figure 20 below.

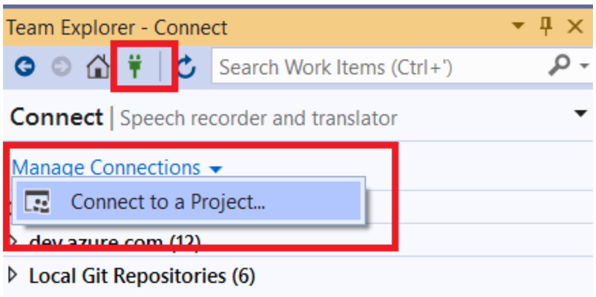

Team Explorer - Connect
Search Work Items (Ctrl+')
Connect | Speech recorder and translator
Manage Connections
Connect to a Project...
dev.azure.com (12)
Local Git Repositories (6)

Figure 20. Connect to Azure DevOps in Microsoft Visual Studio 2019

Then, click on Manage connections and Connect to a project. In the pop-up dialog, choose the project which the developer created in Azure DevOps just now. Next, go to Solution Explorer, right-click on the solution and click on Add Solution to Source Control. After completing the steps, there will be a green plus icon on all the files, then right-click on the solution and click on check-in. This step will upload all the files to Azure DevOps. It is a best practice in web development to back up all the project files. Finally, it's all set up and the developer can start to code. The pseudocode of the primary function is as shown in Table 3 below.

TABLE 3. PSEUDOCODE OF SPEECH-TO-TEXT AND TEXT TRANSLATION FUNCTION

```
Declare and initialize Google Credential to private key of Google Cloud API
Declare and initialize Channel to the Google Credential
Declare and initialize Speech Client to the Channel
Declare and initialize translated text to empty string
Declare and initialize transcript to empty string
Declare and initialize NAudio to new wave input
Declare and initialize Google Cloud Speech request with ByteString audio data
Assign received recording language parameter to Google Cloud Speech request's language code
Declare and initialize Result class to empty class
Assign empty string to recognized text of Result class
Assign empty string to translated text of Result class
While getting responses from Google Cloud Speech to Text API
    Declare and initialize transcript to the final response
    If transcript is not null
        Assign transcript to recognized text of Result class
        Declare and initialize Translation Client to the
        Google Credential and default translation model
        Get translation client result
        Assign translation client result to translated text of
        Result class
Return Result class to display
```

4 Testing and Results

In the testing phase, the developer starts to test the proposed system and record the testing results. There are several test cases in the testing phase. There are three types of testing carried out, including functional testing, non-functional testing, and usability testing. If there is any error found in the testing phase, the developer will solve it immediately. Thus, development and testing often implement together. In agile system development methodology, after the developer carrying out the functional testing, the module will continue to be tested with the targeted users to obtain feedback from the user and improve the module.

The targeted users who will test the modules include two UNIMAS students who are Mandarin speakers from the Faculty of Economics and Business (FEB) and Faculty of Computer Science and Information Technology (FCSIT). They watch educational English YouTube videos frequently to help in understanding some theory better. The student from FEB is a Google Chrome and macOS user. As for the student from FCSIT, she used the Firefox browser and Windows 10. The developer gathers feedback from both users and then improves the system immediately. There are 3 phases for the users to give opinions about the developed system, as shown in table 4 below. The developer completed some improvements according to the feedback from both users. Table 5 to table 8 show the final functional testing result by the developer after the system improvement.

TABLE 4. THREE TESTING PHASES FOR TARGETED USERS IN AGILE METHODOLOGY

Testing Phases in Agile Methodology	**FEB Student's Comment**	**FCSIT Student's Comment**
Phase 1	Can I see the translation in a new window? In this way, I could watch the video and see the translation text at the same time.	I think you can add more sections on the home page to introduce the application. It let the user grasp the features of the application.
Phase 2	Open in a new window function is an excellent update. I think you could add a video tutorial on the home page to let users know exactly how to step-by-step use the web application.	When there is background music on the video, the translation output is empty. I think you could let the system output a dash symbol instead of showing nothing.
Phase 3	The video tutorial you added is so much beneficial to the user. In my opinion, you could also add a documentation page because every digital product has documentation to let users know the detail of the application.	Instead of just showing the translation text, I think the system could also display the recognized text in the video. For example, recognized text on top and translation text in the second row. It will be a noble improvement.

TABLE 5. TEST CASE FOR MICROPHONE PERMISSION

Test Case 1	Microphone Permission				
Step	**Step Detail**	**Input Data**	**Expected Result**	**Actual Result**	**Pass/Fail/Not executed**
1	First load home page	First load home page	Popup microphone permission prompt	Popup microphone permission prompt	Pass
2	Click on Allow button	Allow access microphone	Red recording icon displayed on browser tab	Red recording icon displayed on browser tab	Pass
3	Click on Block button	Block access microphone	Display microphone access required to use the application message	Display microphone access required to use the application message	Pass

TABLE 6. TEST CASE FOR OPEN IN NEW WINDOW

Test Case 2	Open in New Window				
Step	**Step Detail**	**Input Data**	**Expected Result**	**Actual Result**	**Pass/Fail/Not executed**
1	Click on "Open in New Window" button	Open in new window triggered	The page will be loaded in a new window, and old window will be dispose or closed.	The page will be loaded in a new window, and old window will be dispose or closed.	Pass

TABLE 7. TEST CASE FOR OPEN IN NEW WINDOW

Test Case 1	Language Selection				
Step	**Step Detail**	**Input Data**	**Expected Result**	**Actual Result**	**Pass/Fail/Not executed**
1	Select the same recording language and translation language	Recording Language = English, Translation Language = English	Display both language cannot be same message	Display both language cannot be same message	Pass
2	Select different recording language and translation language	Recording Language = English, Translation Language = Mandarin	Display translation results on top	Display translation results on top	Pass

TABLE 8. TEST CASE FOR OPEN IN NEW WINDOW

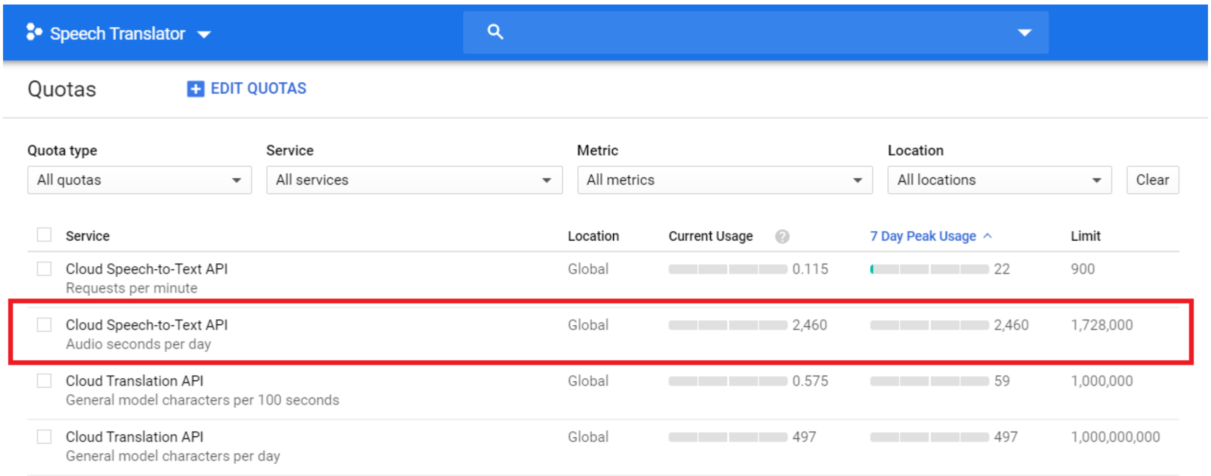
Test Case 1	Translation Results				
Step	**Step Detail**	**Input Data**	**Expected Result**	**Actual Result**	**Pass/Fail/Not executed**
1	Play a YouTube video and click on Start button	Recording Language = English, Translation Language = English	Display recognized English text on the first line and Mandarin translation text on second line	Display recognized English text on the first line and Mandarin translation text on second line	Pass
2	Evaluate the translation results	Results that displayed	Start display word from the start when the output word reached the end of line	Start display word from the start when the output word reached the end of line	Pass

For English to Mandarin transcribing functions, the developer tests three videos on YouTube and records the results in a table form. For Mandarin to English transcribing functions, there are also three videos being tested. However, this journal will only present the first English to Mandarin video transcribing result, due to the concern that all six tables are too long. The first video is a 5 minutes 5 seconds video. It is an English conversation educational video that teaches about the conversation in lending and borrowing things. If there are no results returned from the API, "-" will be displayed. Table 9 below shows the transcribing result.

TABLE 9. FIRST VIDEO TRANSCRIBING RESULT

Video Link	https://youtu.be/iYxs28X3U0I		
Time Span	**Actual Script**	**Recognized Text**	**Translated Text**
0.00	Background music	----	----
0.16	Hi Jennifer, can I borrow the car tomorrow?	can I borrow	我可以借吗
0.22	Why do you want to borrow the car?	want to borrow the car	想借车
0.26	I'm going to the beach with John.	I'm going to the beach	我要去海滩了
0.30	Last time you borrowed it you had an accident and dented the door.	Laugh if you had an accident and	笑 如果你出了意外
0.37	I promise I will drive carefully this time.	----	----
0.41	And the gas tank was almost empty.	And the gas	还有煤气
0.45	I will fill it up.	All fill it up	全部填满
0.47	Well, OK.	Well okay	哦，那好吧
0.52	Great. Thank you.	Great	大
0.54	Background music	----	----
1.04	Can I borrow 10 dollar?	I borrow $10	我借了十块
1.06	Sure. Why do you need it?	Sure	当然
1.11	I want to buy lunch.	Buy lunch	买午餐
1.15	Where is your money?	Where's your money?	你的钱在哪里
1.17	It's not in my wallet.	my wallet	我的钱包
1.21	You wallet is empty?	----	----
1.24	I don't have even one dollar in it.	don't have even $1	连一美元都没有
1.28	Being broke is no fun.	Being broke	被打破
1.32	Even if it's only for a short while.	it's only for a short while	只有一小会儿
1.36	It's always good to have friends.	It's always good	总是很好
1.39	Friends will lend you money when you are broke.	will lend you money when your	当您的
1.44	As long as you pay them back.	As long as you paid	只要您付款
1.47	Background music	----	----
1.51	Do you have some extra money?	Do you have	你有
1.53	How much do you want?	Which do you want	您想要哪一个
1.57	11 dollars.	----	----
1.59	Here you are.	you are	你是
2.01	Thanks a lot.	----	----
2.06	Can I borrow some money? I am so broke.	Can I borrow some So bro	我可以借一些吗索博
2.09	How much do you need?	do you need	你需要
2.12	60 bucks.	----	----
2.15	Here you go.	go	走
2.17	Thank you. I will pay you back soon.	I will pay you back soon	我会尽快还你的
2.20	Background music	----	----
2.24-4.39	"Other conversation still continue but no result returned."	----	----
4.43-5.05	Background music.	----	----

All of the testing of the general functions are passed. It means that all the discovered bugs are solved during the development. However, after 2 minutes 24 seconds of video playing, there are totally no results displayed. It is due to the free version of Google Cloud Speech-to-Text API has a quota or limit of recognizing 2460 audio seconds per day. After reaching peak usage, Google Cloud Speech-to-Text does not manipulate the request anymore. To monitor the API usage, the developer can do so in Google Cloud Console. For example, the developer can check the current quota usage in Google Cloud Console immediately when identifying there is no result returned. Figure 21 shows the quota usage in Google Cloud Console. The second row shows that current usage is 2460 audio seconds which is the same as the 7-day peak usage. It causes the dash symbol which indicates "no results" to be displayed as the translated text.



Quota type: All quotas
Service: All services
Metric: All metrics
Location: All locations

Service	Location	Current Usage	7 Day Peak Usage	Limit
Cloud Speech-to-Text API Requests per minute	Global	0.115	22	900
Cloud Speech-to-Text API Audio seconds per day	Global	2,460	2,460	1,728,000
Cloud Translation API General model characters per 100 seconds	Global	0.575	59	1,000,000
Cloud Translation API General model characters per day	Global	497	497	1,000,000,000

Figure 21. Connect to Azure DevOps in Microsoft Visual Studio 2019

For the accuracy of the translation, it is directly translated from the word. For example, from table 9 at time 0.52, "great" in the sentence means "good" but it translated to a mandarin word means "big". Another example is the word "broke" in period 1.28 in Table 9. It means "no money", but it is directly translated to "break". While the background music is playing, the result shows "-" which indicates "no result". In other words, Google Cloud Speech-to-Text has an accurate recognition of no one is talking. However, it will sometimes miss out on some words. For example, in time 0.22 as shown in table 9, the actual sentence is "Why do you want to borrow the car?", but it is recognized as "want to borrow the car" whereby the words in front are missing. The reason is that Google Cloud Speech-to-Text is still processing the words in the previous second, while the words "Why do you" are speaking in the video. Thus, it bypasses the current playing speech. There is a total of 26 sentences of results recognized by Google Cloud Speech-to-Text. Besides, the results show that a word like "dollar" will result in a "$" symbol. It means Google Cloud Speech-to-Text will automatically recognize unit and output as the symbol of the unit. For the accuracy of speech recognition, it correctly recognizes all the text, except two wrongly recognized texts in 0.30 and 2.06. In the period of 0.30, "last time" is being recognized as "laugh". As for time 2.06, "so broke" is recognized as "so bro".

To sum up, Google Cloud Speech-to-Text recognition is accurate except it will miss out on some words sometimes and there is a 2/26 or 7.69% wrong recognition rate for one video, which is low. Google Cloud Translation will directly translate the words, however, it only happened two times in the whole testing ("great" and "broke") which is still can be accepted. For the recognition of Mandarin, Google Cloud Speech-to-Text returns results that are not accurate when the speaker has a strong accent. In terms of non-functional testing, the performance of the web-based speech recorder and translator reaches a satisfactory level. The loading speed of the landing page is less than 3 seconds. The web application is responsive on both laptops and mobile phones. When observing the two users mentioned in section 4, they do not zoom in on the page. It means that the users can read the translated text. They can also recognize all the buttons on the page. Hence, in terms of usability testing, the results show that the web-based application is considered easy to use.

5 Limitation and Future works

The web-based speech recorder and translator application have been successfully developed. It achieved the main objective of the study, which is to design and develop a web-based speech recorder and translator application. The study contributes detailed methods on how to develop the web application using Google Cloud API and Microsoft Visual Studio Community 2019. It also contributes a questionnaire result to the information technology field for knowing the demand on solving the video subtitle problem. In the testing phase, the study contributes some user feedback for the web application so that other researchers can pay attention to the features. The second objective is to evaluate the usability of Google Cloud Speech-to-Text and Google translation API, and it is also achieved. The usability of Google Cloud STT and Translation has been discussed in section 4 of this paper. It will let other researchers understand the usability of Google Cloud Speech-to-Text and Translation. For example, other researchers get to know Google Cloud STT can recognize background music accurately.

However, the speech recorder and translator application have few limitations. The services of Google Cloud Speech-to-Text and Translation are limited based on the selected plan of the Google Cloud Platform account. Due to the current application used the free version of Google Cloud STT and Translation, it has a quota or limit of recognizing 2460 audio seconds per day. Thus, to fully apply Google Cloud STT and Translation, the developer needs to upgrade the plan.

The second limitation is that the Google Cloud Translation service will have a direct translation. In other words, instead of understanding the meaning of the whole sentence, it manipulates the translation word by word. For example, “great” in the sentence means “good” but it translated to a mandarin word that means “big”. Another limitation is that the current project focused on the scope of translation between English and Mandarin.

To solve the limitations, the web-based speech recorder and translator application need to update and improve in the future. Firstly, the developer can upgrade the plan of the Google Cloud Platform account. Doing so will enable the project to use Google Cloud STT and Translation services more frequently. Furthermore, if there is enough budget, the developer can create a customized model for English-Mandarin translation to solve the direct translation of Google Cloud Translation. In the future, there can be an addition of other languages to make the application reach more users who understand different languages.

6 Conclusion

The web-based speech recorder and translator is successfully developed. It aims to increase the audience’s accessibility to the video content that is using a language that the audiences do not know. The application is functioning well except it is limited to the quota of the Google Cloud STT and Translation plan. In conclusion, the application is successfully developed with the functions of detecting audio, generating text, and translating text between English and Mandarin. The study’s objectives are achieved. However, the application can still be further improved to get it run permanently and support more languages. Nevertheless, this project is a good start for future research and development in the area of speech recognition and translation for video watching platform.

References

Anggraini. N., Kurniawan. A, Wardhani. L. K, & Hakiem. N (2018). Speech Recognition Application for the Speech Impaired using the Android-based Google Cloud Speech API, *Telkomnika* (pp. 2733-2739). Jakarta, Indonesia.

Arcadier (n.d). Retrieved from https://support.arcadier.com/hc/en-us/articles/115001371134-Translating-your-marketplace-using-Google-API-

Google. (n.d.). *Speech-to-text: Automatic speech recognition | google cloud*. Google. Retrieved from https://cloud.google.com/speech-to-text/?utm_source=google&utm_medium=cpc&utm_campaign=japac-MY-all-en-dr-skws-all-all-trial-b-dr-1003987&utm_content=text-ad-none-none-DEV_c-CRE_252507557187-ADGP_Hybrid%2B%7C%2BAW%2BSEM%2B%7C%2BSKWS%2B~%2BT1%2B%7C%2BBMM%2B%7C%2BML%2B%7C%2BM%3A1%2B%7C%2B.

Hees, M. V., Kozlowska, P & Tian, N. (2015). Web-based automatic translation: The Yandex Translate API.

Lai, J. & Yankelovich, N. (2007). Conversational Speech Interfaces and Technologies. Fundamentals, evolving technologies and emerging applications (pp. 381–91). New Jersey, US: Erlbaum Associates.

Levis, J. & Suvorov, R. (2013). Automatic Speech Recognition. The Encyclopedia of Applied of Applied Linguistics.

Santo, L. (2017, July 13). Speech to Text Software Evaluation Report. Switzerland: CERN.

Shetty, V. (2017, March 16). Translation APIs: Google, Microsoft, Yandex. Retrieved from https://medium.com/@vishweshshetty/translation-apis-google-microsoft-yandex-d1aa5ae90dd9

Tatvasoft (2015, April 15). Top 12 Software Development Methodologies & its Advantages & Disadvantages. Retrieved from https://www.tatvasoft.com/blog/top-12-software-development-methodologies-and-its-advantages-disadvantages/

Translator Text (2019). Retrieved from https://azure.microsoft.com/en-us/services/cognitive-services/translator-text-api/

Top Websites Ranking – Most Visited Websites in the world (2019). Top Website Ranking. Retrieved from https://www.similarweb.com/top-websites

Verwaest, M. (2016, September 5). Limecraft Transcriber gets a complete makeover. Retrieved from https://www.limecraft.com/2016/09/05/limecraft-launches-complete-makeover-transcriber-application/

VoxSigma Speech to Text Software Suite. (n.d.). Retrieved from https://www.vocapia.com/voxsigma-speech-to-text.html

Vocapia.com. (n.d.). *Voxsigma*. Vocapia. Retrieved from https://www.vocapia.com/voxsigma-speech-to-text.html.

YouTube Offical Blog (2019). YouTube for press. Retrieved from https://www.youtube.com/about/press/