Proc. of the International Conference on Electrical, Computer and Energy Technologies (ICECET)
9-10 December 2021, Cape Town-South Africa

Real-time Speech Recognition and Subtitles Display

Ameema Arif
Computer & Information Systems Engineering
N.E.D University of Engineering & Technology
Karachi, Pakistan
arif4000423@cloud.neduet.edu.pk

Ifrah Ishtiaq
Computer & Information Systems Engineering
N.E.D University of Engineering & Technology
Karachi, Pakistan
ishtiaq4002070@cloud.neduet.edu.pk

Mahrukh Khan
Computer & Information Systems Engineering
N.E.D University of Engineering & Technology
Karachi, Pakistan
khan4000619@cloud.neduet.edu.pk

Syeda Sara Akif
Computer & Information Systems Engineering
N.E.D University of Engineering & Technology
Karachi, Pakistan
akif4008544@cloud.neduet.edu.pk

Fakhra Aftab
Computer & Information Systems Engineering
N.E.D University of Engineering & Technology
Karachi, Pakistan
fakhraaftab@cloud.neduet.edu.pk

***Abstract*— This paper is based on our project which provides a simple and efficient solution for reducing complications caused in routine communication e.g. environmental noise or hearing disability, speaker or language variability, vocabulary size etc. "Real-time Speech Recognition & Subtitles Display" offers not only real- time speech to text conversion but also real-time translation, providing the means for the users to view and read the speaker's speech in the same or translated language during live communication. The project is packaged as an android app named as "Live Subtitles" that overlays subtitles on the bottom of the camera screen. This approach of speech recognition and translation in the real time eases people to understand each other and come together despite communication barriers.**

Keywords—Automatic Speech Recognition, Speech to Text, Real-Time, Translation, Subtitles, Android Application

I. INTRODUCTION

Communication plays an integral role for human existence without which life may become unmanageable. Speech communication makes the sharing and exchanging of mere thoughts, fascinating ideas, valuable information or messages practically easier and efficient but has its own barriers [1]. To overcome those barriers and allow people to understand each other better, this paper discusses the application of subtitling technique in live interactions.

These communication barriers may occur when people can't understand the speaker's speech due to unfamiliarity with the language or accent, inability to keep up with the speaking speed or due to partial or complete loss of hearing. The number of people having loss of hearing is increasing to a great number. According to [2], nearly 2.5 billion people worldwide ─ or 1 in 4 people ─ will be living with some degree of hearing loss by 2050. This situation calls for a solution that deals with not only the issue of language translation but also provides an effective way of communication for people with hearing deficiency.

"Live Subtitles" not only helps people with all kinds of disabilities and experiences such as auditory processing disorders, dyslexia and other intellectual disabilities but also supports people with different language preferences in real-time. The app "Live Subtitles" increases the retention and the accessibility of the presentation as it allows the users to soak up every word from the speaker, engage with him and his content with confidence. It provides reinforcement in real life so that listeners don't miss any words, enabling them to fully absorb the content and feel engaged in the subject matter. There is no need for people to try to perform lip reading, straining to follow and grasping the speech at the same time. Such a process becomes an exhaustive mental load, and presentation of live subtitles relieves that burden by allowing complete focus and participation.

"Live Subtitles" offers great support in making accessibility of speaker's content easier for people; without having them to take their eyes off of the speaker as the video plays a vital role to help people understand and comprehend the information being provided. Hence, by now the need to have subtitles for the speaker's content in real time videos become clearly understandable targeting people having auditory problems, people having the gaps between their preferred and spoken language as well as general public to prevent from having to miss a point while hearing. This is efficiently done by the use of subtitles in real-time scenario and can be delivered to anybody; people attending an in-person lecture or a formal meeting, people having casual chatting or asking for directions on the way.

Now the paper is divided into 4 sections. First section is about literature review done in regard of this project. Second section is about the methodology and working of the project.

In the third section testing and deployment of the project is discussed. In the last section some future enhancements that can be incorporated in this project are discussed.

II. Related Work

There are many mobile and web applications available for use which provide language translations and subtitles but only a very few products provide real-time translation and subtitling.

Be it *Google's Captioning on Glass*, *National Theatre's Smart Caption Glasses* or *Epson Moverio Smart Glasses*; they're all wearable technology and able to perform real-time speech to text conversion to display it on the smart glasses' heads-up display. Google's technology uses an Android app i.e. "Captioning on Glass" .The app is free but the glasses have a cost of $1,500 [2], [3]. On the other hand, smart glasses offered by National Theatre and Epson Moverio, allows the user to view captions within the glass lenses in real-time relying on a pre-loaded script from the theatre show [4], [3].

The former technology costs around $1,200 a pair and the later from $700 to $1,200 depending on the model [5], [6].

AR Captioning is an AR approach to real-time captioning which uses Head Mounted Display (HMD) to increase glance ability, improve visual contact with speakers, and support access to other visual information (e.g., slides). This approach allows users to manipulate the shape, number and placement of captions in 3D space [7].

The problem with above wearable technology is that it may not be feasible for routine usage or in everyday life. Users have even been appeared to struggle to keep the wide-framed glasses in place throughout the show and not all of these can be used by people who already wear contact glasses as described in [6].

Live Subtitles App from Shravan Apps by Oswald Labs is an android application that uses AR to display real-time subtitles on phone screen but this app is not free and requires pre-registering. It has not yet hit the market and the date for release is still unannounced [8].

Live Transcribe by Google is an android application which uses Google's automatic speech recognition and sound detection technology to provide users with free, real-time transcriptions of their conversations and sends notifications based on their surrounding sounds at home [9].

Live Caption App is a real time transcribing iOS app only with free trial and then requires subscription for $2.99/month. The languages available are Spanish, French, Japanese and Sanskrit [10]. Similarly, Live Transcribe is another such iOS app which supports 50+ languages. It is free to install but contains built-in App purchases [11].

The gotalk.to on Twitter offers live subtitles for video gatherings using Chrome's 'Live Captions' feature for both desktop and mobile which is accessible in English only. No apps and no sign up is required for video chatting in browser of desktop and mobile, both on Android and iOS [12]. SyncWords and VoCaption are live automated captioning and subtitling solutions which are built around Artificial Intelligence and deliver real time speech-to-text processing. However, these are not free and SyncWords is a pay-per-use service which starts for free and then charges $0.60 per minute [13], [14].

All of the above solutions are complex and distracting as user will have to choose to focus on either speaker or the content, only one at a time. Some of them also requires subscription fees which limits people reach as most of the people don't prefer paid apps.

III. Methodology

The project's implementation comprises of three main components:

1. *Automatic Speech Recognition*

Automatic Speech Recognition is the process of conversion of an acoustic speech signal into readable text by a machine. This capability to detect and recognize words from audio to generate a written format of speech is also simply called as Speech Recognition or speech-to-text [15].

2. *Real Time Translation*

Real time translation is basically a tech-driven method to perform conversion of content from the source language into the target language [16]. The real time translation is done with Firebase ML Kit, if the user requires translation.

3. *Text Display*

Lastly, the text is displayed on the screen for the user in the form of subtitles. For our project, we used Open Captions i.e. subtitles that are attached to the video and will be presented along with the video without being turned on and off [17]. Subtitles will be provided in either same or translated language, as specified by the user. To keep our project simple, we have carried out our project for two languages only; English and Urdu.

Fig. 1 is a simple flow chart of the implementation mentioned above:

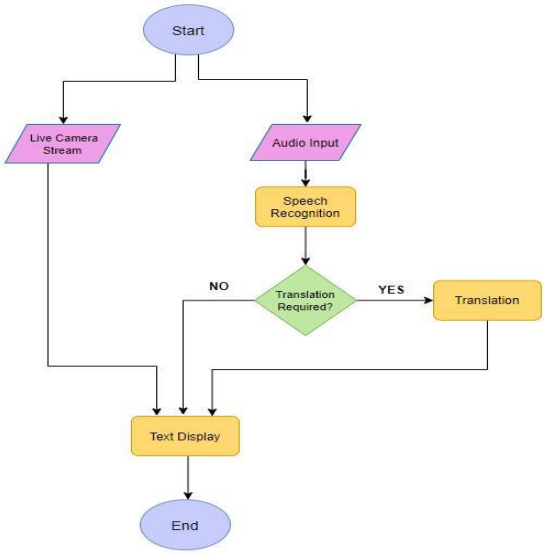

Fig. 1. Project Workflow

A. *Working of Application*

This app is built on Android Studio 4.1.2. Android Studio is one of the most popular platform used to build android applications, as it is open-source and free to use. We have used the Android version 4.1 Jelly Bean API 16 so that our app can run on approximately 99.8% of the devices.

1. *Front-end*

We kept the front-end of our application simple in order to make it user friendly. The main page of app shows three options:

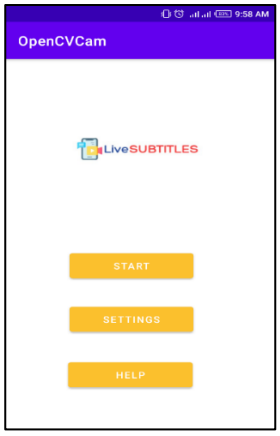

Fig. 2. App Main Screen

- **Start** – Selecting this option will enable user to view live camera streaming along with the subtitle display of audio.

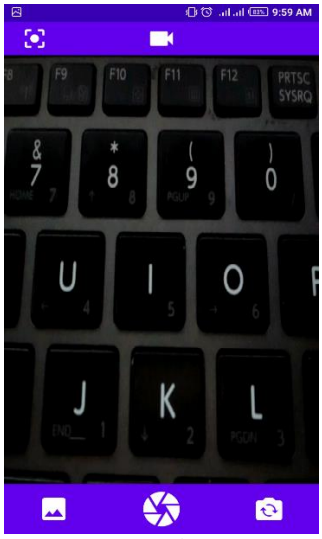

Fig. 3. Camera Interface

- **Settings** – Selecting this option will allow user to choose desired language of subtitle display.

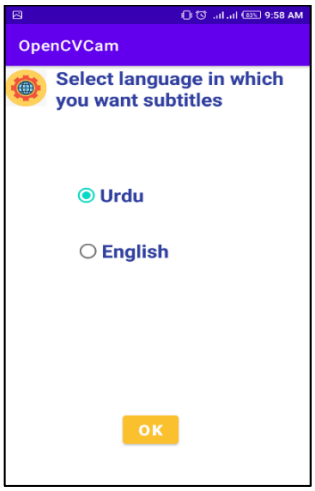

Fig. 4. Settings Interface

- **Help** – Selecting this option will show user basic guidelines on how to use the app.

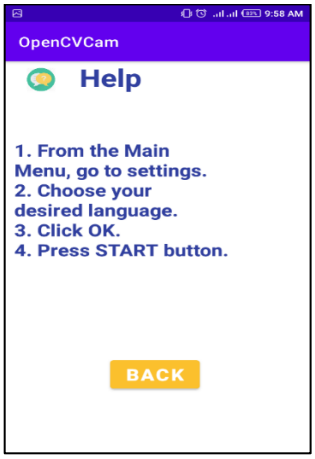

Fig. 5. Help Interface

2. *Back-end*

To develop the video functionality for this app, OpenCV Android-SDK version 3.4.13 have been used. When the video and audio recording starts first the language option that was chosen by the user in settings interface and stored in shared preferences, is fetched. Shared Preferences are used to store and retrieve small amount of primitive data in key/value pairs [24]. Then the recognized speech using Speech Recognizer class of Android Studio is converted into text and passed to the Firebase language identifier in order to identify source language of the video. With the help of source language and option it is decided whether translation is needed or not. If translation is needed it is done using Firebase translation API. Finally, the text is displayed on the screen by passing the text to text view. User's internal storage is accessed to store the video file of the recording when the user ends it.

To use Firebase Language Identifier and Firebase Translation API, Firebase Machine Learning Kit has been included in the project.

Fig. 6 and Fig. 7 shows how the subtitles of recognized and translated speech are displayed on the screen.

Fig. 6. English Subtitles

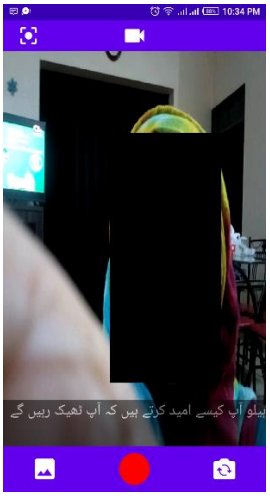

Fig. 7. Subtitles with Urdu Translation

IV. TESTING AND DEPLOYMENT

A. Application Testing

The correct working of the entire application has been tested by individually testing all the functionalities of the system. The test results are presented in Table 1.

TABLE I. APPLICATION TESTING

S.No	Test	Input	Result	Discussion
1	Passing the preferred language Option through radio buttons.	Check one radio button from two, either English or Urdu.	When we start taking video the subtitles shown is in the exactly same language.	We gave input through both radio buttons one by one to ensure they both are working fine.
2	Checking the flip camera option.	Click on the flip camera icon in the bottom right corner.	The flip camera icon is working fine.	We try to make videos using both front and rare camera to ensure the functioning of flip camera.
3	Checking picture taking feature.	Click on the shutter icon in the center bottom bar.	Picture taking feature is working fine.	When shutter is clicked it turns gray indicating that the picture has been taken
4	Checking video feature.	Click on the camcorder icon on the top bar to enter video mode then click on the circle icon to start recording.	The video feature is working properly.	First we enter the video mode and on clicking the circle icon it turns red indicating that the video is being recorded. When the circle button is clicked again the video stops recording and it turns white.
5	Checking the saved videos and photos.	Click the gallery icon in the bottom left corner.	All pictures and videos has been saved.	All the pictures and video can be view in the app gallery in the order of oldest to latest.
6	Checking English Subtitles.	First go to settings and set English as preferred language. Then start making video in English.	English subtitles are displayed properly.	Since the preferred language is English and video is also in English, so subtitles are displayed without translation.
7	Checking Urdu Subtitles.	First go to settings and set Urdu as preferred language. Then start making video in Urdu.	Urdu subtitles are displayed properly.	Since the preferred language is Urdu and video is also in Urdu, so subtitles are displayed

				without translation.
8	Checking Translated English Subtitles.	First go to settings and set English as preferred language. Then start making video in Urdu	English subtitles are displayed but not properly translated as they are displaying Urdu in Roman.	Since the preferred language is English and video is in Urdu, so subtitles are displayed after translation.
9	Checking Translated Urdu Subtitles.	First go to settings and set Urdu as preferred language. Then start making video in English.	Urdu subtitles are displayed properly.	Since the preferred language is Urdu and video is in English, so subtitles are displayed after translation.

B. Performance Testing

To measure the performance of the system, performance testing has been carried out. Performance metrics like App Crashes, Cost, Stability, Battery Consumption, Storage, CPU and Memory Usage has been evaluated and the results are described in Table 2.

TABLE II. PERFORMANCE TESTING

S.No	Performance Metric	Value	Comments
1	App Crashes	1-2%	App Crashes one time when it is used first time after installation if all the permission are not granted.
2	API Latency	1-2 seconds response time	One API is used in this app for translation. So the average response time between some speech and its translation is almost 1 to 2 seconds.
3	Session Length	Depends on user	Session length is the time between app open and close. The session length of this app depends solely on the user.
4	Cost	Minimum 15k to 20k	This is the cost of mobile that is used for running the app. The app itself is free of cost.
5	Stability	97 – 98 %	It means the percentage of sessions that are crash free.
6	App Launch Time	1 – 2 seconds	App usually takes 1 to 2 seconds to launch even when there are already 9 to 10 app processes running on the device.
7	User Interface Response Time	Less than 150 miliseconds	It usually takes less than 150ms to respond to user input.
8	Battery Consumption	10 to 15% per hour on average.	May vary from device to device.
9	CPU Usage	0 - 5mAh	The app is not using extensive CPU power.
10	Memory Usage	0.5 MB memory per hour on average	Memory used by the app per hour.
11	Storage	120 MB	It is the storage space used by the App. This may increase when data related to the app is increased.

C. Application Deployment

Currently this App is in the process of publishing on Android app stores like Google Play Store, SlideMe, Amazon App Store, GetJar, Aptoide, Opera Mobile Store. For testing purposes the app is locally deployed on some devices using Android Application Package APK. Table 3 mentions the set of android devices on which the application has been deployed to ensure correct functionality of the application on different device models and android versions.

TABLE III. APPLICATION DEPLOYMENT

S.No	Device Name	Device Model	Device Version	Display Size	Status
1	Nokia 3	TA-1032	Android 9.0 (Pie)	5 inches	Successful
2	Infinix HOT 4	Infinix X557	Android 7.0 (Nougat)	5.5 inches	Successful
3	Huawei Y5 Prime	DRA-LX2	Android 8.1.0 (Oreo)	5.45 inches	Successful
4	Lenovo K5 Play	-	Android 8.0 (Oreo)	5.7 inches	Successful

V. FUTURE ENHANCEMENTS AND CONCLUSION

From the outcomes of design, implementation, testing and elaboration, it is concluded that our user-friendly application “Live Subtitles” can be considered as a prototype for future developments. Since this app possesses some great newly introduced features, the research conducted in this domain can be proved very helpful for other developers. Subtitles are the need of this era since everyone wants their products to be globally recognized and in order to increase understanding of the product among people, subtitles are used widely. In this section, we will discuss some enhancements that can be made in this application to make the user experience even better.

1. Reduce delays

Because we are performing subtitling on real-time videos, initially there’s a delay of 2 to 3 seconds between speech

recognition and text display on the screen. This delay can be reduced by using some more efficient speech to text converter tools.

2. Save Videos Along with Subtitles

Until now we cannot save the videos along with subtitles so that the user can access them later too. This feature can be introduced so the subtitled videos can be reused anytime.

3. Adding more languages

So far the app only supports English and Urdu Languages. Modifications can be added so that the app can support subtitles and translated subtitles in more languages.

4. Face detection and Speech Detection

Face detection feature can be added in to the app along with speech detection so that the user may know the source of the recognized speech. Additionally we can provide subtitles in different color for different speakers so that if there are more than one speaker in the video, the faces can be highlighted with rectangular boxes of the same color as that of the subtitles generated from the users' recognized speech.

5. Use Machine Learning

The machine learning model can be trained with videos and subtitles for it to learn how a word is spoken in many different ways; how speakers generally say a word, the pronunciation of different names and so on.

The project has been developed after performing research on speech recognition, speech to text conversion, translation and display of subtitles based on some products and techniques related to these domains. This paper explains the need and usefulness of our application along with its software development i.e. how this project has been developed and which tools and platforms have been used for this purpose. Each and every practical aspect has been discussed in detail. Project testing and the results obtained from these tests have been noted to help evaluate the working of the application. This product can act as a source of guidance for further development in this domain. Furthermore, some future enhancements have also been mentioned which we think can make our product more useful and efficient.

ACKNOWLEDGMENT

First of all, we are grateful to Almighty Allah for giving us the strength, knowledge and understanding to complete and deliver this project. Secondly, we are thankful to our internal supervisor Ms. Fakhra Aftab for her immense support, time and guidance during the whole course of completion of this project. We would also acknowledge all our teachers who taught us and gave us the required skills which we were able to implement in our project. We are also thankful to our departmental staff and university staff, who assisted us during our stay at the university.

REFRENCES

[1] World Health Organization, "WHO: 1 in 4 people projected to have hearing problems by 2050," WHO, Geneva, 2 March 2021.

[2] S. Gaudin, "Computer World," 8 October 2014. [Online]. Available: https://www.computerworld.com/article/2822820/i-understand-you-now-theres-a-google-glass-app-for-hard-of-hearing-users.html.

[3] N. Vega, "NewYork Post," 24 January 2020. [Online]. Available: https://nypost.com/2020/01/24/new-moverio-smart-glasses-could-help-deaf-theatergoers/.

[4] "National Theatre," [Online]. Available: https://www.nationaltheatre.org.uk/your-visit/access/caption-glasses.

[5] C. Kelsall, "American Theatre," 23 January 2020. [Online]. Available: https://www.americantheatre.org/2020/01/23/with-smart-caption-glasses-the-eyes-have-it/.

[6] C. Huston, "Broadway News," 28 January 2020. [Online]. Available: https://broadwaynews.com/2020/01/28/galapro-tests-out-smart-glasses-with-live-captions-on-broadway/.

[7] "MakeAbility Lab," 1 January 2016. [Online]. Available: https://makeabilitylab.cs.washington.edu/project/arcaptions/.

[8] "Oswald Labs," [Online]. Available: https://oswaldlabs.com/platform/shravan/apps/live-subtitles/.

[9] D. Copithorne, "Hearing Tracker," 7 February 2019. [Online]. Available: https://www.hearingtracker.com/news/google-live-transcribe-app.

[10] "Live Caption," [Online]. Available: http://www.livecaptionapp.com/.

[11] "App Store," [Online]. Available: https://apps.apple.com/us/app/live-transcribe/id1471473738.

[12] "gotalk.to," [Online]. Available: https://gotalk.to/.

[13] "SyncWords," [Online]. Available: https://www.syncwords.com/company/about.

[14] "BroadStream Solutions," [Online]. Available: https://broadstream.com/vocaption-live/#languages.

[15] IBM, "IBM Cloud Learn Hub," 2 September 2020. [Online]. Available: https://www.ibm.com/cloud/learn/speech-recognition.

[16] A. Ava, "Mars Translation," 12 August 2020. [Online]. Available: https://www.marstranslation.com/blog/real-time-translation.

[17] REV, "REV," 6 September 2019. [Online]. Available: https://www.rev.com/blog/resources/open-caption-vs-closed-caption-whats-the-difference.

[18] A. Saha, "Read, Write and Display a video using OpenCV |", LearnOpenCV – OpenCV, PyTorch, Keras, Tensorflow examples and tutorials, 2021. [Online]. Available: https://learnopencv.com/read-write-and-display-a-video-using-opencv-cpp-python/.

[19] S. Max, "Video Captioning with Keras", Medium, 2021. [Online]. Available: https://medium.com/analytics-vidhya/video-captioning-with-keras-511984a2cfff.

[20] E. Chidera, "A Beginner's Guide to Setting up OpenCV Android Library on Android Studio", Medium, 2021. [Online]. Available: https://medium.com/android-news/a-beginners-guide-to-setting-up-opencv-android-library-on-android-studio-19794e220f3c.

[21] "TextView , Android Developers", Android Developers, 2021. [Online]. Available: https://developer.android.com/reference/android/widget/TextView.

[22] H. Nguyen, "Use Android Speech Recognition so that it stops only at the press of a button", Stack Overflow, 2021. [Online]. Available: https://stackoverflow.com/questions/17144361/use-android-speech-recognition-so-that-it-stops-only-at-the-press-of-a-button.

[23] "How to pass data from one activity to another in Android using shared preferences?" Tutorialspoint.com, 2021. [Online]. Available: https://www.tutorialspoint.com/how-to-pass-data-from-one-activity-to-another-in-android-using-shared-preferences.

[24] "Shared Preferences in Android with Example - GeeksforGeeks", GeeksforGeeks, 2021. [Online]. Available: https://www.geeksforgeeks.org/shared-preferences-in-android-with-examples/.