Live Subtitles Using Augmented Reality

Subtitle as needed ***(paper subtitle)***

Ifrah Ishtiaq
Computer & Information Systems Engineering
NED University of Engineering & Technology
Karachi, Pakistan
ifrahishtiaq99@gmail.com

Ameema Arif
Computer & Information Systems Engineering
NED University of Engineering & Technology
Karachi, Pakistan
ameema.arif1@gmail.com

Mahrukh Khan
Computer & Information Systems Engineering
NED University of Engineering & Technology
Karachi, Pakistan
mahrukhkhan015@gmail.com

Syeda Sara Akif
Computer & Information Systems Engineering
NED University of Engineering & Technology
Karachi, Pakistan
saraakif123@gmail.com

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. "Live Subtitles Using Augmented Reality" 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 that uses augmented reality to overlay subtitles on the screen beneath the speaker's detected face. This approach of speech recognition and translation in the real time eases people to understand and come together despite communication barriers.

Keywords—Automatic Speech Recognition, Speech to Text, Face Detection, Face Tracking, …

I. INTRODUCTION (*HEADING 1*)

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. 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 speech 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 World Health Organization, Nearly 2.5 billion people worldwide ─ or 1 in 4 people ─ will be living with some degree of hearing loss by 2050 [1].This situation calls for solution that deals with not only the language translation issue but also provide 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 support people with different language preferences in real time. 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. Live captions provide reinforcement so they never miss a word, enabling them to fully absorb the content and feel engaged in the subject matter. With live captioning, people do not have to try to perform lip reading, straining to follow and trying to grasp the speech at the same time. Such a process is an exhaustive mental load, and live subtitles relieve that burden and allow complete focus and participations. It offers great support by making access 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, here the need becomes clearly visible to make real time videos available to the people having auditory problems and even more for the people to remove the gaps of their native language and also for having to miss a point while hearing. This can be best done by the use of subtitles in the real-time scenario. Live captions 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.

II. RELATED WORK

There are many mobile and web applications to provide language translations and subtitles which are available for use but very few products provide Real-Time translation and captioning.

Be it *Google's Captioning on Glass*, *National Theatre's Smart Caption Glasses* or *Epson Moverio Smart Glasses*; they're all wearable technology able to show speech-to-text on the glasses. Google's technology uses an Android app i.e. "Captioning on Glass" for real-time speech to text conversion and then displays it on the Glass heads-up display. The app is free but the glasses cost of $1,500 [2], [3]. On the other hand,

National Theatre's and Epson Moverio's technology 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 costs around $1,200 a pair and the later $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 everyday. Users even struggled to keep the wide-framed glasses in place throughout the show when Epson Moverio Smart Glasses were tested. Also not all of these can be used by people who wear contact glasses [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]. Similarly, 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 an iOS app only with free trial and then subscription for $2.99/month. It allows the user to caption while speaking. The languages available are Spanish, French, Japanese and Sanskrit [10]. Similarly, Live Transcribe is another iOS app which supports 50+ languages. It is free to install but contains built-in App purchases for real-time transcriptions [11].

•gotalk.to on Twitter offers Live subtitles for video gatherings using Chrome's 'Live Captions' feature for both desktop and mobile and as of now accessible in English, the feature will do its best to show what is being said in real time. Video chat can be done right in your browser on desktop and mobile, both on Android and iOS. No apps and no sign up is required. Share your screen, record gatherings, stream live, and more [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 $0.60 per minute [13], [14].

All of the above solutions are complex and distracting as user will face difficulty in being attentive and focused on speaker and the transcription of speech at the same time.

III. METHODOLOGY

The project's implementation comprises of four 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].

This is achieved by first recognizing speech from audio of real-time video through Speech Recognizer class of Android Studio, detecting source language using Firebase Language identifier and then finally converting the speech to text.

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.

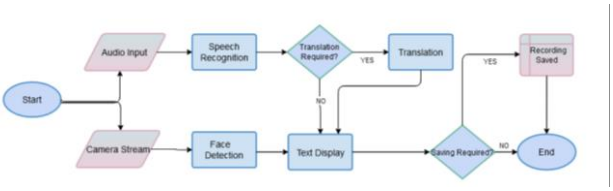
3. Face Tracking

Face Detection is the identification and extraction of features of human faces in images through Artificial Intelligence. This can also be done in digital video frames and is called as Face Tracking [17], [18]. We have used OpenCV module is to perform face tracking.

4. Text Display with Augmented Reality

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 [19]. 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.

The following figure is a simple flow chart of the implementation mentioned above:



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.

- 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;

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

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

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

- Back-end

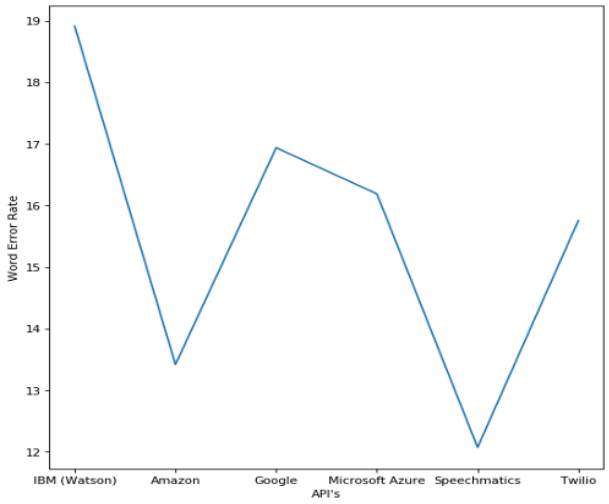
User's internal storage is accessed to store the video file of the recording when the user ends it.

, In this section we will discuss some approaches that are used for speech to text conversion and then putting that text as subtitles on the video.

B. Other Approaches

1. For Automatic Speech Recognition

   - (REFERENCE https://medium.com/analytics-vidhya/comparison-of-speech-to-text-apis-d8e0410ec924) The speakers in the Audio all speak with three different accents; American, Asian, British. I have calculated the mean of each of the API's, and drawn a graph based on that.These Cloud API's keep on working to improve its model using customer data. The data computed would be different from the time you are reading this



   - Microsoft Azure vs Google Cloud STT

     Google's software stands out for its multi-language support. Speech-to-Text is capable of transcribing audio in any of 120 languages to text. By comparison, Microsoft's speech to text software only supports 29 languages at this time. On its face, Microsoft Azure Speech Service is significantly cheaper than Google Cloud Speech-to-Text. Microsoft offers five hours of free transcription per month and then charges $1 per hour of audio after that. Google provides just one hour of free transcription, after which the service costs $1.44 per hour of audio. For straightforward audio transcription, Microsoft Azure Speech Service tends to perform better than Google Cloud Speech-to-Text. The difference is that Microsoft's software uses AI to make sure that what it's transcribing makes linguistic sense. Since this software can accept custom speech models, it also handles accents, lisps, and other speech impediments significantly better than Google's Speech-to-Text platform. (https://www.techradar.com/news/speech-apps-microsoft-vs-Google)

2. For Text Display with Augmented Reality

   - Using Android Studio Only

     Doing everything in android studio using Java alone and not using any external module or software.

     Result: Couldn't find any function in Java to extract audio from time to time to convert it to text and put it on the screen as subtitles.

   - Using Python with Java

     Making video using Java in android studio and passing that video to Python function as a parameter to do the further processing. This is done by using Chaquopy which is a platform to execute Python scripts in android studio.

     Result: Couldn't find any function in Java to pass frames one by one to Python function for processing.

   - Using Python Only

     When the button is clicked calling Python function to do everything from capturing video, extracting audio, converting speech to text to displaying that text in the form of subtitles.

     Result: This approach build successfully but App crashes on run time.

   - Using SRT files in Python

     Write the recognized audio in an SRT (SubRip Text) file time to time and read that file to put subtitles on the video, using Python. Note that SRT files are subtitles files that contain Block no. , starting and ending time for the subtitle to be displayed

on the screen and the subtitle to be displayed within that time.

Result: Reading and writing the SRT file simultaneously causes a lot of delay and runtime errors.

- Using OpenCV in Python and Java

  Make video in android studio using OpenCV module and pass that video as a parameter to a Python function to display subtitles using OpenCV. Note that we are distributing the tasks between Java and Python because there are some functions of OpenCV that are available in Python but not in Java.

  Result: App keeps crashing when we start recording video.

- Using Python and OpenCV in Java

  Making video in android studio using OpenCV get the recognized audio from python function using speech recognition module and put the text returned from python function on the screen using OpenCV putText function in android studio.

  Result: App keeps crashing when we start recording video.

- Using OpenCV in Java

  Making video in android studio using OpenCV module and putting a hardcoded text on it using OpenCV putText function.

  Result: App working fine and displaying the overlayed text on the screen when we start recording video.

- Using OpenCV in Python

  Making video using OpenCV in Python and recognizing audio using speech recognition module and putting that recognized text on screen using OpenCV.

  Result: Things working fine but slow.

IV. TESTING AND RESULTS

3. For Automatic Speech Recognition

S. No.	Test Condition	Status (Successful/ unsuccessful)	Remarks
1	Speaker is at a distance	Successful	Can recognize speech up to a distance of 17 to 18 cm
2	With or without earphones	Successful	Working with earphones as well as without them.
3	Noise	Successful	Can tolerate noise of up to 75db
4	Tested on all members in all languages	Successful	Tested with both English and Urdu
5	Strong and weak Wi-Fi connection	Successful	Working fine in both conditions.
6	Working on multiple devices at same time	Successful	Tested on multiple mobile devices at the same time.

4. For Face Tracking

S. No.	Test Condition	Status (Successful/ unsuccessful)	Remarks
1	Face is at a distance		
2	With or without speech		
3	Multiple Faces		
4	Tested on all members in		

	all languages		
5	Strong and weak wifi connection		
6	Working on multiple devices at same time		

5. For Real Time Translation

S. No.	Test Condition	Status (Successful/ unsuccessful)	Remarks
1.	Speaker is at a distance	Successful	Can recognize speech up to a distance of 17 to 18 cm
2.	With or without earphones	Successful	Working with earphones as well as without them.
3.	Noise	Successful	Can tolerate noise of up to 75db
4.	Tested on all members in all languages	Successful	Tested with both English and Urdu
5.	Strong and weak wifi connection	Successful	Working fine in both conditions.
6.	Working on multiple devices at same time	Successful	Tested on multiple mobile devices at the same time.

6. For Text Display with Augmented Reality

S. No.	Test Condition	Status	Remarks
1.	Accuracy	50%	Can't convert Urdu speech to text properly. Instead of translating Urdu into English, it is writing Urdu in roman.
2.	Delay	1 to 2 seconds	That's the average delay between speech and text display.
3.	Speaker is at a distance	Working	Working fine up to a distance of 20cm.
4.	Works without face detection	Working	-
5.	Strong and weak wifi connection	Working	Working in both conditions.
6.	Working on multiple devices at same time	Yes	Tested on two mobiles at the same time.

V. FUTURE ENHANCEMENT AND CONCLUSION

A. Future Enhancement

This app can be considered as a prototype for future developments. Since this app possesses some great newly introduced features, so the research conducted in this domain can be proved very helpful for other developers. Subtitles are the need of this era since everyone want their products to be globally recognized and in order to increase understanding of the product among people subtitles are used widely. In this chapter we will discuss some enhancements that can be made in this application to make the user experience better.

1. Reduce delays

Since we are doing subtitling on real-time videos, initially there's a delay of 2 to 3 seconds between the speech recognition and text display on the screen. This delay can be reduced by using some more efficient speech to text converter tools. Note that this project has been made by using tools and platforms that are open source and easily available so that this application can be used freely by people for learning purposes. There are some tools that are more efficient but they are not free.

2. Save Videos Along with Subtitles

Up till now we cannot save the videos along with subtitles so that the user can access them later too. So this feature can be introduced save the videos along with subtitles so that they can be reused anytime. When saving the videos with subtitles the user can be allowed to correct the mistakes in the recognized text and translations.

3. Adding more languages

Up till now the app only supports English and Urdu Language. More languages can be added so that the app can support subtitles and translated subtitles in more languages. This can increase the use of this app widely.

4. Face detection and Speech Detection

The current version of the app does not support face detection feature. Face detection feature can be added in to the app along with speech detection so that the user may know from whom the recognized speech is coming from. Additionally we can provide subtitles in different color for different speakers so that if there are more than 1 speaker in the video his face can be highlighted with a rectangular box in the same color in the recognized speech from this user is displayed as subtitles.

5. Using Machine Learning

As we know that machine learning plays a great role in developing artificial intelligent applications. So Machine learning can be accommodated in this app to make it more efficient.

The machine learning model can be trained with videos and subtitles such that it can learn that how a word can be spoken in many different ways, how speakers generally says a word, the pronunciation of different names and so on. This can help them to recognize, convert and translate speech more efficiently. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

B. Conclusion

The first part of the project is some basic research on speech recognition, speech to text conversion, translation and displaying of subtitles. The research is based on some products and techniques related to these domains. This part also includes the need and usefulness of this project. The next part is software development of this project. This includes how this project is made and which tools and platforms are used for this purpose. In this part each and every practical aspect of this app is discussed in detail. This part also includes some practical research or we can say different hit and trial approaches that are used to achieve the final outcome. This research can play the role of do's and don'ts for further development in this domain. Later this project is tested and the results have been noted in accordance with these tests. The last part of the project is based on the discussion of some future enhancements that can made in this app to make it more useful and efficient.

ACKNOWLEDGMENT *(Heading 5)*

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 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.

REFERENCES

[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] C. Bernstein, "face detection," TechTarget, [Online]. Available: https://searchenterpriseai.techtarget.com/definition/face-detection.

[18] "Everything About Face Tracking," sightcorp, [Online]. Available: https://sightcorp.com/knowledge-base/face-tracking/.

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