Subtitle Generating Media Player using Mozilla DeepSpeech Model

Waat Perera
Intelligence Research Laboratory,
Faculty of Computing,
General Sir John Kotelawala Defence University

B. Hettige
Intelligence Research Laboratory,
Faculty of Computing,
General Sir John Kotelawala Defence University

ABSTRACT
Subtitles plays a major role when comes to consuming media. Most of the time either media content comes without any subtitles or comes with basic subtitles in the native language. So, finding subtitles from another language than the native or creating subtitles for a new media content wasn't an easy task. For famous films, tv shows or sometimes songs could find subtitles in more than one language but there are majority of content that isn't exposed to internet. To address this issue this paper proposes a method to generate real-time subtitles for selected languages using English language media files through the existing Mozilla DeepSpeech and Google Cloud Platform Translation API. This proposed system takes any English media content from .mp4 file format as the input and generate subtitle according to the users desired language preference as a .srt output. Further, this paper also describes an overview of existing methods for Speech to Text conversion, advantages and disadvantages that are compared with Mozilla DeepSpeech model. The system has been tested with Human evaluation methods as well as automated evaluation method namely BLEU.

General Terms
Deep Learning, Language Translation, Media Player, Speech Recognition, Speech to Text, Subtitle Generation.

Keywords
Deep Learning, DeepSpeech, Language Translation, Media Player, Mozilla, Speech Recognition, Speech to Text, Subtitle Generation.

1. INTRODUCTION
Traditional media players have been in use in digital space for a long time but all of them are identical and share verry common feature set and small differences in shortcuts and looks and feels. So that the functionality or the purpose of a media player hasn't been changed for a while. To change this and make ordinary media player something better and more useful authors suggest an efficient method to build a subtitle generator to a media player[1].

Speech recognition and Speech to Text translation has been a tough area to compute for a while because of the nature of those sources, which is audio or sound. Naturally analog sound, or audio, is transmitted as waves of pressure through a medium such as air. From a computing point of view sound is time sequence data which is a measurement of an entity changing over time in this case its sound. So why is it so hard to process audio data? It's not the format that makes it hard it's the variation of the data. There can be countless variations of the same content or the word regardless of the similarity of the meaning. Human speech audio can be differ based on the environment, background noise, speaker, audio recording device (mic), physical state of the speaker (ill, tired, or well), emotions (angry, happy, sad, neural) and many more facts. This is what makes audio recognition so hard and why it's been as it is for a long time[2].

Recently with the rapid development of computer hardware (specially GPUs), data science and processing techniques, machine learning and computer algorithms there is an improvement in the audio processing and speech recognition field. Because of this, Automatic Speech Recognition (ASR) has advanced significantly in recent years thanks to the use of larger and larger deep neural networks.

Focus here is to use those techniques to achieve the goal, to build a media player with automatic subtitle generation. There are so many pre trained STT (Speech-to-Text) translation acoustic models out there that performs efficient enough as both online APIs and offline models. Also, there are language models and rescoring algorithms that have been improved over time and contributed by the community as open-source projects that can be used to achieve this goal. Compare with others Mozilla DeepSpeech model shows more successful results for the speech to text conversion especially for the English language. Therefore, proposed medial player has been designed using DeepSpeech model.

The rest of the paper is structured as follows. Section 2 provides brief description on commonly used Speech to Text conversion APIs and models. Then section 3 contains the proposed design for the DeepSpeech model based media player. In the 4th section paper will discuss the implementation methods of the proposed solution and steps of developing full media player. The next section, Evaluation methods and results of the proposed system will be provided. Finally, section 6 will include the conclusion further works.

2. RELATED WORK
Since training neural networks takes lots of computation power doing it in large scales is really hard as independent developer. Even to train a model to differentiate two objects by images takes considerable amount of processing power. Roughly the more epochs a model is trained the accurate its results get. Training a model to understand human spoken language as a whole is really resource and time-consuming task also, gathering enough vocal samples with enough variation and quality is a challenging task. So, in order to train an efficient and accurate model developer should overcome all the above-mentioned challenges. However large companies like Google, Amazon and IBM have their own in house STT models that were specifically trained for their services in huge server farms which perform unbelievably accurate. Some of them are provided to end users as API services by those companies. Among those Mozilla DeepSpeech project takes a special place because they provide their trained model as an open-source

project that the developers can use in offline applications[3], [4].

2.1 Watson Speech to Text

Watson Speech to Text is a Speech to text translation service provided by IBM company. It's mainly based as an API service with inbuild AI features that makes this service more suitable for business uses. As for features of this model it can be customized with business specific vocabulary to understand brand names, products, and model numbers. Also, the AI is capable of learning itself and adjust to more often used vocabulary. A feature that stands from other STT models is that this service has the capability to identify different speakers in a conversation and transcribe the conversation as a dialogue. Because of all those business-oriented features, ease of use and with the premium price tag this model is mostly used in large scale businesses to improve their customer services[5]–[7].

2.2 Amazon Transcribe

Amazon Transcribe is the Speech to Text service provided by the world's leading web service provider AWS (Amazon Web Services). Same as the Watson STT service this web-based service also specifically business oriented with features capability to extract information and generate insights from conversations, video files, clinical calls and with safety features like masking sensitive information and privacy details. Because of these services build in security features and as its intended this is mostly used in call centers, online help centers and large business firms.

2.3 Google Cloud Speech to Text

This is the Googles own STT model that is used in almost every google service or product that involves STT translation. Because of the wide range of use cases and because this is a self-learning online model that is provided as an API this get exposed to massive number of samples daily so that this is one of the most accurate speech recognition models in existence and capable of almost handling any type of conversation. The most important aspect of this model and the advantage over others is that the language support, which can understand and translate the most number of languages. This model is trained to mostly work with day-to-day conversations which makes this suitable for any consumer application[3], [4], [7]–[10].

2.4 Mozilla DeepSpeech

Mozilla DeepSpeech project is taking a different approach in STT translation. This project is an open-source project maintained by Mozilla. Everything from source code, models to data set is accessible to public and because of it this project is community driven. Specially the data set which is the CommonVoice is a community driven data set where anybody can contribute their own voice samples for the data set. The dataset is robust enough because of the huge variation and also growing rapidly throughout the releases. The source code of the neural network and the architecture is provided with all the codes and instructions so that anybody can train their own model and even Mozilla is providing the pre trained model with a scorer for English language[11]–[13].

Table 1. Comparison of Modern Speech-to-Text Systems

Comparison of Modern Speech-to-Text Systems	Provider	License	Advantages	Disadvantages
Watson Speech to Text	IBM	Commercial	Inbuilt AI features Business oriented Customizable vocabulary Identify different speakers and generate dialogues	Only support English language Provided as an API service High price tag
Amazon Transcribe	AWS	Commercial	Business oriented Can extract information and insights Inbuilt security features to mask sensitive information	Only support English language Provided as an API service High price tag
Google Cloud Speech to Text	GCP	Free to use	Self-learning model High accuracy Can transcribe many languages	Provided as an API service No customization in vocabulary
Mozilla DeepSpeech	Mozilla	Mozilla public license (Open source)	Community driven data set Can train using use case	Accuracy depends on the data set

			specific data Model architecture can be modified Huge community support Can be trained to any language with a correct data set Better performance in noisy environments	

3. DESIGN

This section describes details on Mozilla DeepSpeech model and Design of the proposed system.

3.1 DeepSpeech Model

The latest DeepSpeech model which is version 0.9.3 is a speech recognition engine, released under the Mozilla Public License. The core of this model is a RNN (Recurrent Neural Network) specifically strained to analyze speech spectrograms and generate English text transcripts[14]–[17].

This model consists of 5 layers. The first layer takes the audio as MFCC (Mel Frequency Cepstral Coefficients) as shown in figure 1 and use ReLu (Rectified Linear Unit) as the activation function. the first 3 layers are non-recurrent layers for feature extraction. The fourth layer is a recurrent layer with forward recurrence. The fifth layer is also a non-recurrent layer which takes the forward units as the inputs from the previous layer. The output layer is standard logits that correspond to the predicted character probabilities for each time slice[18]–[21].

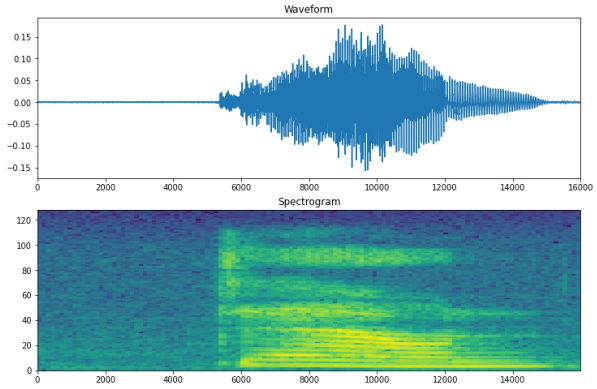

Fig 1: Mel Spectrogram Conversion

As a summery in figure 2 shows the entire architecture of the model as in the figure below.

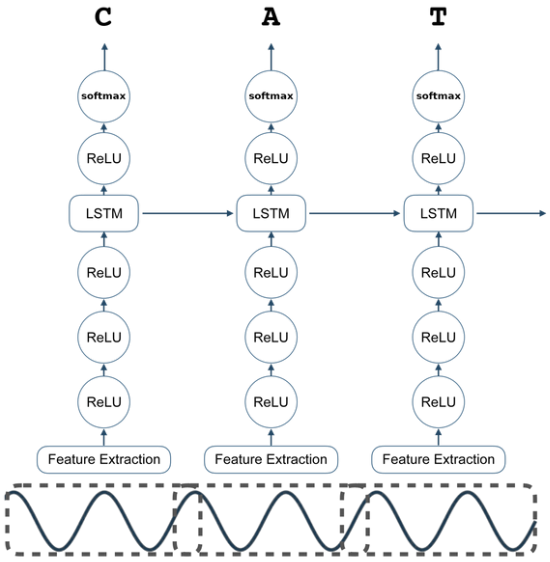

Fig 2: The DeepSpeech complete RNN model

3.2 DeepSpeech Scorer

For any Language when perform the STT conversation the process needs to tackle with two properties, the acoustic features, and linguistic features. The model will handle the acoustic features while a scorer is needed to handle the linguistic features.

DeepSpeech model comes with its own predesigned scorer which is based on Connectionist Temporal Classification (CTC) loss function which is shown in figure 3. Basically, what it does is that it calculates the likelihood of a word to be in a sentence in a given context. First it takes every word that a particular sound might be and then calculate the probability of that word will be in a sentence with the other words. Then it will output the most probable sentence as the correct one[22]–[25].

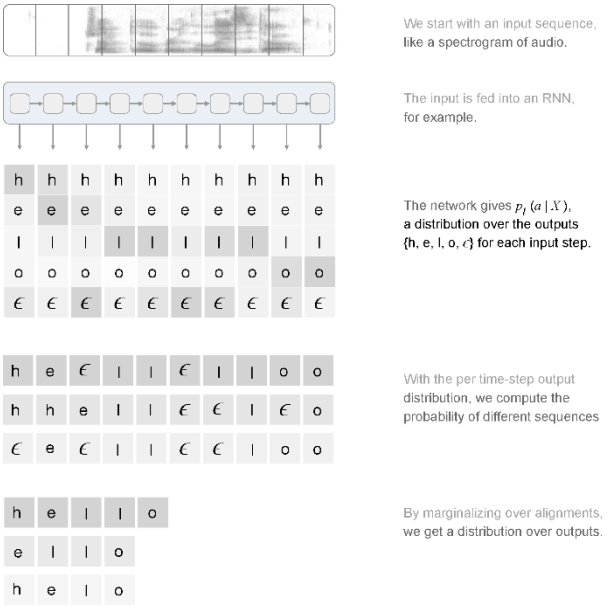

Fig 3: CTC Loss Function

Default scorer is accurate enough for most occasions and it works efficiently with the model, but DeepSpeech provides the relevant documentations to design own scorer if need. As for now it is enough to use the default scorer for the current purpose.

3.3 STT Conversion

As for the purpose of building the speech tot text translation engine it's better to take the pre trained model as it can be applied to any conversation. Latest version will be the most relevant and because all the model versions are architecturally same and only the trained data set and the performance and the accuracy is different it won't be an issue unless the version is really old with a different architecture.

3.4 Inputs

Best audio format for this model is .wav because it is the recommended input format by the creators. However, in this use case the data is in video format. To perform the STT translation using the model firstly it needs to separate the audio from the video. The method used to perform this task should be relatively easy to execute and should take less computing power and less time.

Proposed solution is to perform this task by using ffmpeg, which is a complete, cross-platform solution to record, convert and stream audio and video. Also, for better utilization this can be run as a sub task using subprocess library.

Then this wav data should be converted into 16Khz numpy array. To handle this, using SoX (Sound eXchange) is recommended because of its light weight and capabilities and it's better to run this also as a subprocess.

3.5 Audio Segmenting

Firstly, if the audio is in two different tracks as stereo, then they need to be converted in to one track as mono audio. Then the silence sections should be removed from the audio and slice and split the audio into those segments. This is because there is a noticeable silence between sentences so that that gap can be used to break each sentence in the subtitles. Based on the language, speaking style and speaker the time between sentences may vary

3.6 Processing

Then the preprocessed files can be run through the deepspeech model and scorer to generate the relevant text for each audio segment. The deepspeech model will handle the acoustic properties of the sound and then the scorer will refine the generated output by considering the language properties.

3.7 Language Translation (Optional)

Because currently the default DeepSpeech model only support English in large scope the generated text will be work well mostly for audio in English. However, the generated text can be translated from the model to any desired language. There are few ways of archiving this goal. Most accurate way of processing this is to build a separate language translation model from a relevant data set so that it will perform accurately in that particular context but as for a general purpose using a API will be enough because in this stage text is used as inputs and outputs so that the data usage will be rally low and most online translation APIs like Google Cloud Platform (GCP) Translator will work fine with any regular grammar on any context.

3.8 Writing to File

Those generated text can be then written into a .srt (subtitle) file. The important thing here is that words in .srt files need time stamp for each so that it can be displayed at specific time in the media. To archive this, proposed solution uses a format called as VTT (Video Time Text), where words may be associated with a timestamp cue.

Figure 4 is a high-level overview of the system design. You can see the different modulus and the processing pipeline together with the GCP Language Translation module.

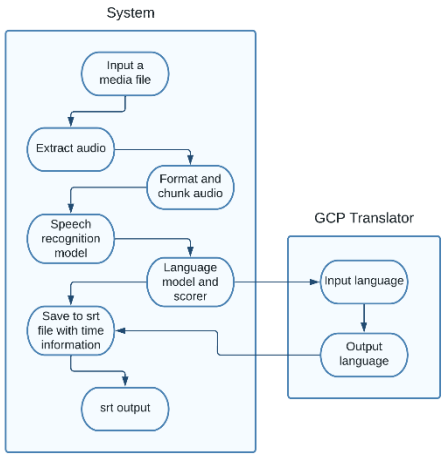

Fig 4: Processing Pipeline of the System

4. IMPLEMENTATION

For the implementation the solution can be deployed via 3 formats. In each case there will be pros and cons but based on the scenario and the user's requirement all 3 implementations are usable to perform the processing to get the desired outputs.

4.1 Separate Subtitle Generator

In this implementation only the STT translation and the scorer will be built into the application where it performs its primary function and generate the relevant subtitle file. So that the user can either use it as a separate .srt file or bake directly into the media file as embedded subtitles. Then the end user can open the media file and the subtitle file in their preferred media player and playback. But the downside is that this method takes

few steps from users' side to finally play the output and although the development and the processing part is simple the user experience is not much present.

4.2 Extension to an Existing Media-player

By using an existing media player as a base and starting development from there. Open-source media players like VLC player already supports 3rd party extensions and they provide the necessary documentations to build extensions. Using this implementation will bypass most of the steps of the processing pipeline. It will take much less developing time at the cost of customizability because when build an extension, it needs to follow the original developers' guidelines and their structure so that the extension doesn't have the full control over the code.

4.3 Standalone Media-player with Inbuild Subtitle Generator

This method gives the full control over the code base and the processing pipeline so the flow can be optimized to best preform in desired use case. But also, this takes much developing time because there is the need to build a fully functioning media player ground up. To show the proposed solution's implementation's full potential this is the recommended method because there will be no external bottlenecks to the system so that user can experience the model's full potential.

As for the interface the basic system is runnable as a separate subtitle generator as the first implementation method trough a CLI interface using Python commands shown in figure 5.

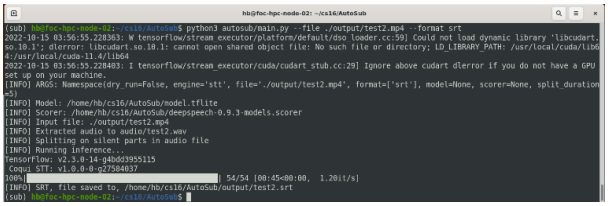

Fig 5: Runnable CLI

To make this more user-friendly implementation building a GUI is suggested that can provide the function to the end user so that user can interact with the software without tackling with the code or software environment. Figure 6 shows the current design of the GUI.

Fig 6: Media Player GUI

This media player (SUBGEN Player) demonstrates the third implementation which is to make a standalone media player with the mentioned system. This implementation excludes most of the errors from the system because developer has the control over the full system from inputs to outputs. Figure 7 is the high-level system architecture of the media player.

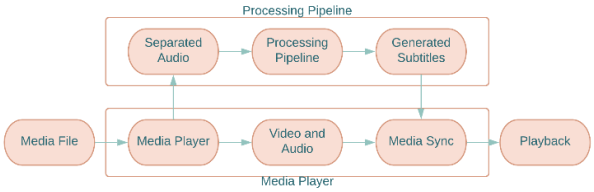

Fig 7: Media Player Architecture

As for inputs this method can eliminate unsupported files going through the model by simply validating input formats. And developer can give the user to select the translation language of the subtitles right away via a drop down in a user-friendly manner as shown in figure 8.

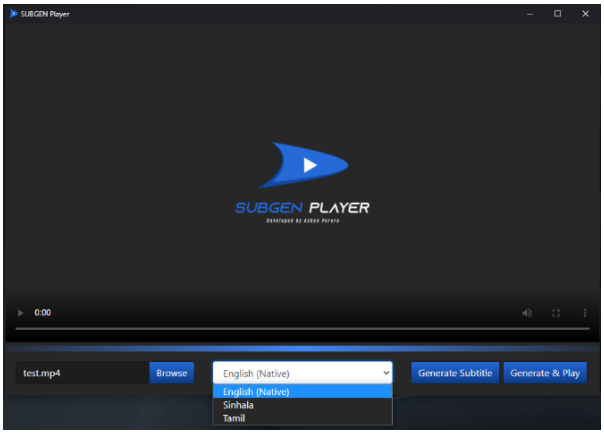

Fig 8: Language Select

After that the STT subtitle generation and language translation can be performed and playback the video right away via the media player interface as shown in figure 9. In this approach as mentioned before user will get a pleasant experience yet get the chance to user the full functionality of the system without and compromises seamlessly.

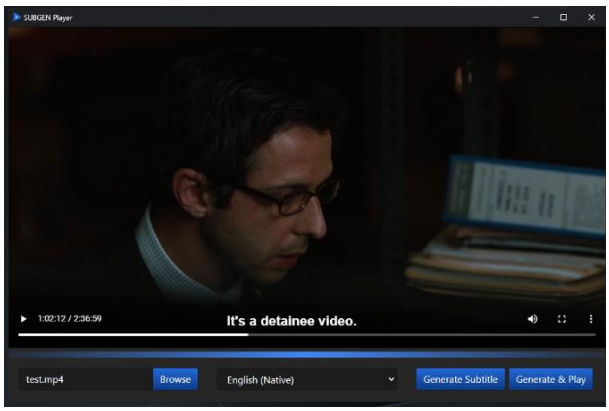

Fig 9: Video Playback with Subtitles

5. TESTING AND EVALUATION

For the testing purposes the solution was implemented as a separate subtitle generator which is the first method that suggests in the previous section of the paper. Then took 100 video samples from movies, speeches and tv shows with different speakers all covering kids, teens, adults, elders both male and female in different emotions sad, angry, happy, disappointed in different environments with different background noises from different sources like recordings,

announcements and phone calls and ran each one through the system multiple times. Below are some of the test samples with their relevant outputs as a summery.

Table 2. Original vs Generated Subtitles

Original content	Generated subtitle
Oh, what about my hair	what about my hair
My name is cliff I'm rick dons stunt double	My name's cliff I'm rick done stunt double
I just read it getting this for my husband	I just read it getting into a gift from my husband
I would do it you know me I would like to help you out	I would do it you know me I would like to help you out.
this feels too easy	this feel too easy
for far too long we have been enemies but today is a new day	for for long we have been enemies . but today is a new day
doctor calm me down clam down Kevin	doctor come down come down even.
I couldn't go to my local station of course so I chose a different line all together	I couldn't go to my local station of course so I was a different man altogether.
but there is really a case where someone's handwriting telling me something that I need to know	but there's really a case where someone's handwriting doesn't tell me something I need to know
there is no one that you can think of might buy the entire amount	there is no one that you can think of might buy the entire amount
that sir is with my blessing what you are paying for	that is with my blessing for you a pain for
what you talking about now	what you talking about now
what you doing here anyway	what you doing here anyway
what I fail to recognize is why Michel should be motivated to write you a cheque for 20 million	what I failed to recognise it why Michael should be motivated to write you a check for twenty million
no, this time it is different. when I look into that babies' eyes something changed	now this time was therefore when i look into that little baby's eyes so changes
this is personal for me too. let's bring everything we got	his personal for me to a spring every single thing we got
you go out there for vengeance you gonna get someone killed	to go out there for venters you won't get some one killed
I know that you want you want to apologize and I hate to tell you I'm not interested. you think I'm supposed to apologize you?	know what you want you want to apologize and i hate to tell you i'm not interested you think i'm supposed to apologize
each place in the map is a different level of the game and levels get harder as we go	I place on the opposite of level of the game and the levels get harder as we go.
the ostrich is a flyless bird. one of 60 species	the ostrich is a flyless bird. one of sixty species
I'm a former member of the middle class raised by two	a former member of the middle class raised by two
accountants in a tiny apartment in base side queen	accountants in a tiny apartment in base side queen
money doesn't just buy you better life better food better cars	one doesn't just buy you a better life than a food that a cause on
wanna know what money sounds like? go to a trading floor of wall street	annamoe sounds like go to a trading for westerly
your only responsibility is to put meat on the table	responsibility is to put me on the table
name of the game. move the money from your client's pocket into your pocket	name the game moved the money from your clients pocket in the your pocket
reason for the call today john is something just came across my desk john it is perhaps the best thing that I ever seen in last 6 months if you have 60 seconds I would like to share the idea with you gotta minute?	reason for the cold to day john is something just came across my desk john it is perhaps the best thing I've seen in the last six months seconds I'd like to share the idea what do you got a man.
I've been thinking about quitting singing ever since I was young	I've been thinking about putting saying ever since I was young.
you can't be serious. I am stand down	you can't be serious I am stan to
listen his heart rate is slowing	listen his heart race is slowly
show me his territorial routes	some as territorial robs
the mass extension we fear has already begun and we are the cause we are the infection	the man extinction we feared has already begun we cause we are the affection
email from Los Angeles times weather	mail from Los Angeles times whether
hey guys how's it going. hi Theo, hey why didn't you call me back last week	hey guys how's it going. hi there why didn't you call me back last wake
always the fruit. don't you know what people say you gotta eat your fruits and juice your vegetables	always the fruit but to don't you know and people say you've got I eat your fruit and juice your bench
its mandatory I'm a board member	it mandatory I'm a board member
it's like seeing a unicorn in 5000 dollar suite	like saying a unicorn in five thousand dollar suite
bane if you are watching we have your money don't heart the boy	bain if you're watching we have your money don't hurt the boy.
who is the girl? my new assistant	who the girl my no. assistant
look at that sword. all guns are outlawed city punishable by death.	I could that word all guns are outlawed in city. punishable by death.

Most for the times results were accurate enough and most importantly they were consistent across the board in multiple runs. There were few times that some words were miss read by the model and scorer wasn't able to correct them but every time it was the same word that the system mistaken, and the error was the same too however those mistakes were made when the speaker speaks verry fast or there are more than one person speaking or sometimes in environments with too much background noise or pronunciation was close with homophones

or rhyming words and although there were mistakes the final output sentence was grammatically correct. So that by adjusting the scorer by a custom scorer file or adding some context-based data to the training data set and retraining the DeepSpeech model will bias the neural network to correct those errors itself.

As for further analysis there was a test which cross checked the data form the inputs and outputs to demonstrate the word count of each sentence and how close is the distribution. To this test some random sentences were taken, and their word counts then run them through the model and plot them against the generated output sentences word count as a distribution. Below figure 10 and figure 11 are word distributions of the inputs and outputs respectively.

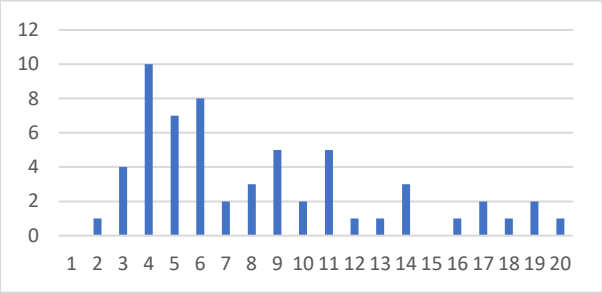

Fig 10: Word Distribution of Input Sentences

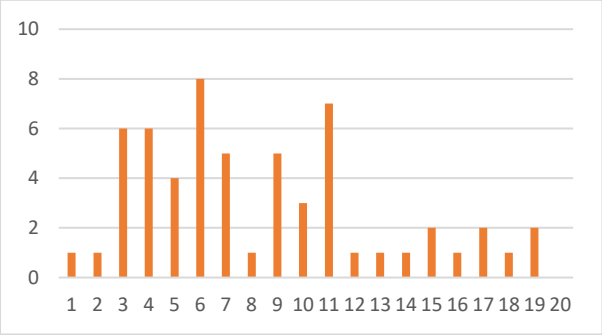

Fig 11: Word Distribution of Output Sentences

Also, randomly selected 20 sentences was taken and calculated their Adequacy and Fluency metrics on a scale of 1-5 for more detailed and robust analysis. Those parameters are known as standard representations of Machine Translation (MT) accuracy. For better visualization below is a plot of the Adequacy and Fluency scores against the sentence count. Figure 12 and 13 represents the results of Adequacy and Fluency testing respectively.

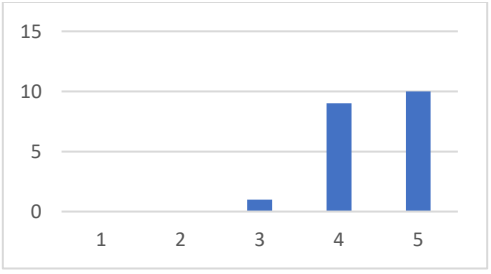

Fig 12: Adequacy Testing

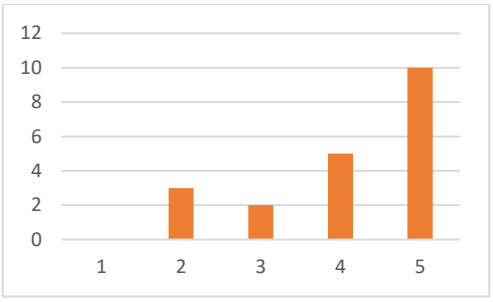

Fig 13: Fluency Testing

The BLEU is the most popular inexpensive, fast, language-independent, and automated evaluation matrix for machine translation[26]. The BLEU evaluation matrix provides results in between zero and one, which indicates how similar the candidate text is to the reference text. If the BLEU value is much closer to 1, it represents similar texts.

Consider the process of the BLEU, which calculated scores for individual segments in a sentence. The final score takes considering the average of these scores over the whole corpus. The BLEU score can be calculated using below equation shown in figure 14.

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}.$$

Then,

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

Fig 14: BLUE score

The freely available BLUE score calculation script was used to run the testing sentences and result sentences and got BLEU scores listed below[27].

BP: 0.9687850933540854

P(1) = 336/473 = 0.7103594080338267
P(2) = 224/434 = 0.5161290322580645
P(3) = 161/395 = 0.40759493670886077
P(4) = 116/356 = 0.3258426966292135

BLEU: 0.4550883804338765

Next for testing language translation portion by running the same clip without the language translation which generates subtitles in native language (English) and then GCP translator enabled translated results (Sinhala) and compared them like shown in the figure 15.

Fig 15: BLUE score

For more in detailed look, table 3 displays the output from both as .txt files and compare them. Surprisingly they were accurate to a level that its generally usable.

Table 3: Original vs Translated Subtitles Comparison

Original subtitles	Translated subtitle
three years ago in our state of wisconsin that in the united states a man went into a seat temple	මීට වසර තුනකට පෙර අපේ විස්කොන්සින් ප්‍රාන්තයේ එක්සත් ජනපදයේ මිනිසෙක් ආසන පන්සලකට ගියේය
in a terrible act of violence killed six innocent people americans and indians	දරුණු ප්‍රචණ්ඩ ක්‍රියාවකින් අහිංසක ඇමරිකානුවන් සහ ඉන්දියානුවන් හය දෙනෙකු මිය ගියේය
and in that moment of share grief are two countries reaffirmed a basic truth as we must	එම දුක බෙදාගන්නා මොහොතේ රටවල් දෙකක් අප විසින් කළ යුතු මූලික සත්‍යයක් යළි තහවුරු කර ඇත
in his youth and for representing this nation's energy and its optimism	සහ ඇය වෙනුවෙන් කරන සියලුම කැපී පෙනෙන වැඩ
or to practice no faith at all and to do so free of persecution and fear and discrimination	සෑම පුද්ගලයෙකුටම තම ඇදහිල්ල තමන් තෝරා ගන්නා ආකාරය හෝ කිසිදු ඇදහිල්ලක් පිළිපැදීමට අයිතියක් ඇති බව
great honor to be the first american president to join you for republic day	ජනරජ දිනය සඳහා ඔබ හා එක්වන පළමු ඇමරිකානු ජනාධිපතිවරයා වීම මහත් ගෞරවයකි
when the tricolor waving above us we celebrated the strength of your constitute	ත්‍රිවර්ණ ධජය අපට ඉහළින් ලෙළදෙන විට අපි සැමරුවේ ඔබතුමාගේ බලයේ ශක්තියයි
and many times i believe that the relationship between india and the united states can be one of the	ඉන්දියාව සහ එක්සත් ජනපදය අතර සම්බන්ධය ඉන් එකක් විය හැකි බව බොහෝ විට මම විශ්වාස කරමි

Also, while testing although there is a use of a third-party API for translation so that there was continuous data sending and fetching through internet the system was fast enough that there wasn't any lag, or the model couldn't keep up the speed. Considering the language count that GCP Translator API supports it's a fair trade to perform the translation this way compromising few seconds of the processing time.

6. CONCLUSION AND FURTHER WORK

This paper has described design and development of the Mozilla DeepSpeech model based Media player, that can be used to generate subtitle automatically. According to the results obtained, this proposed system is reliable to a level where this is generally usable. Since if took the media players implementation method the only third-party entity in the system will be the GCP translator so that controlling anything inside the processing pipeline from inputs to outputs is possible. This let developers with the space to develop the system furthermore. To improve and refine the performance, accuracy and usability of the system below changes can be conducted as further work.

Since there is a third-party API in the system that can affect to the performance at any time because it connects to the API through network which is an attribute that can't be controlled it's better to develop a language translator for at least the three most used languages. This method will let the media player to fully function as a standalone software.

The other major development that authors propose is to develop few more models to perform the STT translation for other languages than English. Since the model architecture and the code is already available preparing a data set enough to train the model in few different languages and implement those models to the system can hugely improve the user base of the system.

7. REFERENCES

[1] A. Ramani, A. Rao, V. Vidya, and V. B. Prasad, "Automatic Subtitle Generation for Videos," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Mar. 2020, pp. 132–135. doi: 10.1109/ICACCS48705.2020.9074180.

[2] N. Radha and R. Pradeep, "Automated subtitle generation," vol. 10, pp. 24741–24746, Jan. 2015.

[3] B. Xu, C. Tao, Z. Feng, Y. Raqui, and S. Ranwez, *A Benchmarking on Cloud based Speech-To-Text Services for French Speech and Background Noise Effect*. 2021.

[4] P. R. Hjulström, "Evaluation of a speech recognition system," 2015. https://www.semanticscholar.org/paper/Evaluation-of-a-speech-recognition-system-Hjulstr%C3%B6m/49c1997d54811c7eb79463260f3513c2a89b7235 (accessed Oct. 10, 2022).

[5] J. Huang *et al.*, "The IBM Rich Transcription Spring 2006 Speech-to-Text System for Lecture Meetings," May 2006, pp. 432–443. doi: 10.1007/11965152_38.

[6] R. D. Sharp *et al.*, "The Watson speech recognition engine," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Apr. 1997, pp. 4065–4068 vol.5. doi: 10.1109/ICASSP.1997.604839.

[7] F. Filippidou and L. Moussiades, "A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems," *Artificial Intelligence Applications and Innovations*, vol. 583, pp. 73–82, May 2020, doi: 10.1007/978-3-030-49161-1_7.

[8] M. Stenman, "Automatic speech recognition An evaluation of Google Speech," *undefined*, 2015, Accessed: Oct. 10, 2022. [Online]. Available: https://www.semanticscholar.org/paper/Automatic-speech-recognition-An-evaluation-of-Stenman/69dab8bf2f729ed94f53a2dd5df03799258b34a8

[9] N. Anggraini, A. Kuniawan, L. Wardhani, and N. Hakiem, "Speech Recognition Application for the Speech Impaired using the Android-based Google Cloud Speech API," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 16, pp. 2733–2739, Dec. 2018, doi: 10.12928/TELKOMNIKA.v16i6.9638.

[10] J. Y. Chan and H. H. Wang, "Speech Recorder and Translator using Google Cloud Speech-to-Text and Translation | Journal of IT in Asia," Dec. 2021, Accessed: Oct. 10, 2022. [Online]. Available: https://publisher.unimas.my/ojs/index.php/JITA/article/view/2815

[11] A. Agarwal and T. Zesch, *Robustness of end-to-end Automatic Speech Recognition Models -- A Case Study using Mozilla DeepSpeech*. 2021.

[12] A. Agarwal and T. Zesch, "LTL-UDE at Low-Resource Speech-to-Text Shared Task: Investigating Mozilla DeepSpeech in a low-resource setting," p. 5.

[13] E. Nacimiento-García, C. S. González-González, and F. L. Gutiérrez-Vela, "Automatic captions on video calls, a must for the elderly: Using Mozilla DeepSpeech for the STT," in *Proceedings of the XXI International Conference on Human Computer Interaction*, in Interacción '21. New York, NY, USA: Association for Computing Machinery, Sep. 2021, pp. 1–7. doi: 10.1145/3471391.3471392.

[14] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, Mar. 2020, doi: 10.1016/j.physd.2019.132306.

[15] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition." arXiv, Feb. 05, 2014. doi: 10.48550/arXiv.1402.1128.

[16] G. E. Dahl, Dong Yu, Li Deng, and A. Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012, doi: 10.1109/TASL.2011.2134090.

[17] A. Amberkar, P. Awasarmol, G. Deshmukh, and P. Dave, "Speech Recognition using Recurrent Neural Networks," Mar. 2018, pp. 1–4. doi: 10.1109/ICCTCT.2018.8551185.

[18] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)." arXiv, Feb. 07, 2019. Accessed: Oct. 10, 2022. [Online]. Available: http://arxiv.org/abs/1803.08375

[19] C. K. On, P. M. Pandiyan, S. Yaacob, and A. Saudi, "Mel-frequency cepstral coefficient analysis in speech recognition," in *2006 International Conference on Computing & Informatics*, Jun. 2006, pp. 1–5. doi: 10.1109/ICOCI.2006.5276486.

[20] L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques." arXiv, Mar. 22, 2010. doi: 10.48550/arXiv.1003.4083.

[21] R. Vergin, D. O'Shaughnessy, and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 525–532, Sep. 1999, doi: 10.1109/89.784104.

[22] A. Graves, "Connectionist Temporal Classification," in *Supervised Sequence Labelling with Recurrent Neural Networks*, A. Graves, Ed., in Studies in Computational Intelligence. Berlin, Heidelberg: Springer, 2012, pp. 61–93. doi: 10.1007/978-3-642-24797-2_7.

[23] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks," presented at the ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning, Jan. 2006, pp. 369–376. doi: 10.1145/1143844.1143891.

[24] H. Scheidl, S. Fiel, and R. Sablatnig, "Word Beam Search: A Connectionist Temporal Classification Decoding Algorithm," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Aug. 2018, pp. 253–258. doi: 10.1109/ICFHR-2018.2018.00052.

[25] A. Hannun, "Sequence Modeling with CTC," *Distill*, vol. 2, no. 11, p. e8, Nov. 2017, doi: 10.23915/distill.00008.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. doi: 10.3115/1073083.1073135.

[27] D. Dmello, "donnabelldmello/nlp-bleu." Nov. 17, 2019. Accessed: Oct. 22, 2022. [Online]. Available: https://github.com/donnabelldmello/nlp-bleu