Article

Speech-to-Text Conventional Myanmar (Burmese) Language Recognition System

Halawati Binti Abd Jalil , Pyae Sone Phyo , Md Amin Ullah Sheikh , Riskhan Basheer *

Posted Date: 24 June 2024

doi: 10.20944/preprints202406.1696.v1

Keywords: Speech to Text; NLP; Language Recognition

Article

Speech-to-Text Conventional Myanmar (Burmese) Language Recognition System

Halawati Binti Abd Jalil, Pyae Sone Phyo, Md Amin Ullah Sheikh and Riskhan Basheer *

School of Computing and Informatics, Albukhary International University, Alor Setar, Malaysia;
halawati@aiu.edu.my (H.B.A.J.); pyae.phyo@student.aiu.edu.my (P.S.P.);
amin.sheikh@student.aiu.edu.my (M.A.U.S.)
* Correspondence: b.riskhan@aiu.edu.my

Abstract: Only languages with abundant resources are given the opportunity to build speech recognition systems, whereas low resource languages are disregarded. This project attempts to create a system that can normally convert speech to text in the Myanmar (Burmese) language, a language part of low resource languages. The Burmese language is extremely intricate and is made up of vowels, consonants, tones, and stress. There are 10 numbers, 12 vowels, and 33 letters in the Myanmar alphabet. Deep learning allows for the evolution of numerous algorithms while eradicating ignorance and even low-resource languages may be turned into speech-to-text systems with a small quantity of data. The procedures required to construct a successful speech recognition system that can assist both the corporate sector and society will be covered in this project, including data collecting, preprocessing, modeling, prototyping, assessment, and ultimately deployment.

Keywords: speech to text; NLP; language recognition

1. Introduction

1.1. Overview

A program's capacity to convert spoken language into written language is known as speech recognition, also known as Automated Speech Recognition (ASR), computer speech recognition, or speech-to-text. Despite being sometimes mistaken with voice recognition, speech recognition focuses on converting speech from a verbal to a written format whereas voice recognition simply aims to distinguish the voice of a certain person (IBM, 2018). The use of speech recognition is essential because it can be used to search the data and information easily with speech (IEEE Signal Processing Society, 2018). The purpose and objective of this experiment are to establish a prototype that can recognize Burmese speech and conventionally convert it into text.

1.2. Background of Study

This project aims to obtain a prototype that can conventionally transcribe the speech spoken in the Burmese language into text. This research intends to present a well-developed prototype of the speech recognition system mainly focused on syllable-based language, by using syllable-based recognition since Myanmar is a monosyllabic and syllable-timed language (Acharjya, et.al., Wunna Soe & Yadana Theins, 2015). This project will be researched on many research papers, conference papers, and articles, by building an extensive pronunciation dictionary, and enhancing the originality of Myanmar speech synthesis with linguistic information to make the recognition more conventional. The final output of this project will be a prototype that can transcribe Myanmar speech into Burmese words conventionally.

1.3. Problem Statement

The problem is that only rich resource languages get the chance to develop a speech recognition system, for other low resource languages like Burmese, it does not get a proper speech recognition system. Due to the language's complexity, none of the publically accessible language recognition programmes for Myanmar (Burmese) perform as well as they should. So, a prototype that can conventionally transcribe the speech into text is necessary to develop for ease of use. For example, Youtube supports the Burmese language but not speech recognition. Whenever someone tried to search for a video with the Burmese speech, YouTube cannot transcribe properly and shows the wrong videos to the users.

1.4. Research Objectives

- To learn how speech recognition works with the use of a deep learning model and its execution.
- To develop a prototype that is capable of transcribing Myanmar(Burmese) Language speech into text.
- To build a website that is implemented with the prototype to use the system easily.

2. Literature Review

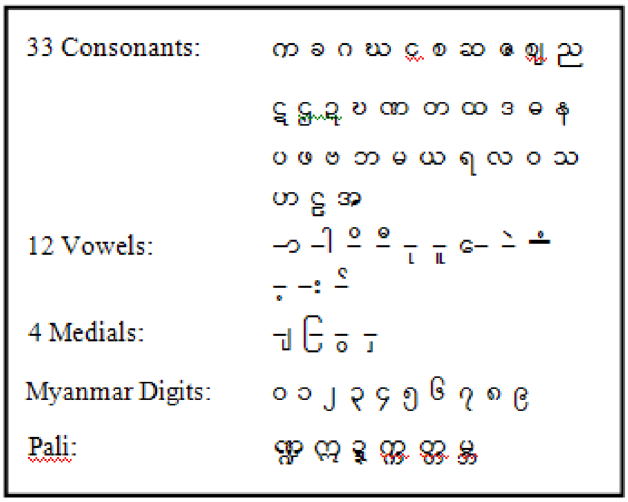
2.1. Overview

This is not the first time someone has tried to build a speech-to-text system for the Burmese language. Many websites try to achieve this goal, but sadly, they all quit halfway. Some systems can only detect specified sentences and print out the results but sometimes it does not even work. It is because Burmese grammar is complex and some words and pronunciations are the same but have different meanings. So, building this system is not an easy task.

Speech recognition is a growing field and many young talented people are trying to establish a new record. Sooner or later, there will be many speech recognition systems for unnoticed languages around the world. If you dig more into this field and you will find more interesting things along the way. Many big companies are investing a lot of money into their speech recognition system to stand out from each other. That is why there will be more work opportunities for young enthusiastic people who are interested in natural language processing.

2.2. Burmese Language

The Burmese language is so complex and consists of Tones, Stress, Vowels, and Consonants. Where the tones are referred to as the lexical meaning of the syllables. There are two groups of tones consists of static and dynamic. Where the static has three different categories low, middle, and high. Also two different categories in dynamic as falling and rising. The stress is like louder, longer, and higher speech that is opposite to normal speech. The Myanmar alphabet consists of 33 letters, 12 vowels, and ten numerals and is written like other languages, from left to right. Unlike English, Burmese sentences do not need any space in sentences but for beauty and cleanliness, sometimes space is used (Wikipedia, 2023). Figure 2.1 shows the Myanmar basic consonants, also the Myanmar vowels, and the Myanmar basic numerals.



33 Consonants:	က ခ ဂ ဃ င စ ဆ ဇ ဈ ည ဋ ဌ ဍ ဎ ဏ တ ထ ဒ ဓ န ပ ဖ ဗ ဘ မ ယ ရ လ ဝ သ ဟ ဠ အ
12 Vowels:	ာ ါ ိ ီ ု ူ ေ ဲ ံ ့ း ်
4 Medials:	ျ ြ ွ ှ
Myanmar Digits:	၀ ၁ ၂ ၃ ၄ ၅ ၆ ၇ ၈ ၉
Pali:	ဿ ဣ ဍ္ဍ ဏ္ဍ ဏ္ဌ ဍ္ဎ

Figure 2.1. Burmese language(Sann Su Su Yee, 2010).

2.3. Related Work

There are some papers that already discussed the Myanmar ASR. Wunna Soe presented Myanmar's syllable-based voice recognition technology (Fatima, et.al., Wunna Soe & Yadana Theins, 2015). Syllable segmentation and the syllable-based n-gram approach were used in this system to build the language model. Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) were used to construct an acoustic model. The speech is recorded utilizing recording software like wave surfer in the news domain area.

Thin Thin Nwe showed how to recognize speech in Myanmar using a hybrid Artificial Neural Network (ANN) and HMM (Gopi, 2022, et.al., Gouda, 2022, et.al., Thin Thin Nwe & Theingi Myint, 2015). This technique employed segmentation based on syllables. Techniques for extracting features included Mel Frequency Cepstral Coefficient (MFCC), Linear Predictive Cepstral Coding (LPCC), and Perceptual Linear Prediction (PLP). The words were recognized using hybrid ANN-HMM.

Hay Mar Soe Naing presented a continuous voice recognition system for the travel industry with a vast vocabulary for Myanmar (Humayun, 2022, et. al., Hay Mar Soe Naing et al., 2015). Deep Neural Networks (DNN) were employed in this system to model acoustics. The acoustic model was expanded with tonal features. For DNN training, sequence discriminative criteria such as Cross-Entropy (CE) and State-level Minimum Bayes Risk (sMBR) were applied.

Ingyin Khaing presented continuous speech recognition for Myanmar using Dynamic Time Wrapping (DTW) and HMM (Lim, 2019, et.al., Ingyin Khaing, 2013). The feature extraction process in this system used Linear Prediction Coefficients (LPC), MFCC, and Gammatone Cepstral Coefficients (GTCC) approaches. Additionally, DTW was applied to feature clustering to address the Markov model's lack of discrimination. For the recognition procedure, HMM was utilized.

Aye Nyein Mon, Win Pa Pa, and Ye Kyaw Thu changed Convolutional Neural Network (CNN) hyperparameters for Myanmar ASR, and the performance of Myanmar ASR is enhanced (Majid, et.al., 2021 & Aye Nyein Mon et al., 2018). The localization attribute of the CNN can lower the number of neural network weights that must be taught, hence decreasing overfitting. Additionally, the pooling process is highly helpful in addressing the little frequency changes that speech signals frequently have. As a result, CNN's various parameters, such as feature map counts and pooling sizes, have been altered in order to improve ASR accuracy for the Burmese language. Because Myanmar is a tonal and syllable-timed language, a large data collection was used to construct the syllable-based Myanmar ASR. Following that, two open test sets, web data, and recorded data, are used to assess the effectiveness of word-based and syllable-based ASRs.

(Saeed, et.al., 2019, Shah, et.al., 2019 & Yong et al., 2023) discusses the implementation of efficient smart street lights that are capable of monitoring crime and accidents. It can be inferred that the system likely uses some form of recognition technology to detect and report incidents. This could potentially include image or pattern recognition to identify accidents or criminal activity. (Mallick et al., 2023) present a solution to the transportation problem related to drug delivery from drug factories to different warehouses. The aim is to minimize the delivery time as well as the cost of transportation. The paper uses the Stepping Stone method for optimization of the cost and compares it with Vogel's method. This involves the use of mathematical and computational recognition to identify the most efficient routes and schedules. In another paper, (Tabbakh, et.al., 2021 & Mallick et al., 2023) discuss the minimization of costs in airline crew scheduling using an assignment technique. The paper presents a case study in which the airline schedule and crew schedule are optimized to minimize transportation cost. This involves the use of recognition in the form of identifying the optimal assignment of crew members to flights to minimize costs.

(Zaman et al., 2021) present an ontological framework for information extraction from diverse scientific sources. This involves the use of recognition in the form of identifying relevant information from various sources and extracting it in a meaningful way. The paper uses deep learning to evolve numerous algorithms and even low-resource languages may be turned into speech-to-text systems with a small quantity of data. (Hussain et al., 2021) discuss performance enhancement in wireless body area networks with secure communication. The paper presents a solution to the problem of secure communication in wireless body area networks, which likely involves the use of recognition systems to identify and mitigate potential security threats. The Hierarchically Aggregated Graph Neural Networks (HAGNN) proposed by (V. Singhal et al., 2020., & Xu et al., 2023) is used to capture different granularities of high-level representations of text content and fusing the rumor propagation structure. This approach could potentially be adapted for speech-to-text systems to capture different levels of linguistic features in the speech data and use them for more accurate transcription.

Lawrence demonstrated that whole-word reference-based linked word recognition algorithms have advanced to the point where they can now achieve great recognition performance for tiny or syntax-restricted, moderate-sized vocabularies in a speaker-trained mode (RABINER et al., 1990, p. xx). In particular, it has been shown that very high string accuracy for a vocabulary of digits may be achieved in a speaker-trained mode utilizing either HMM or templates as the digit reference patterns.

A voice processing and identification method for individually spoken Urdu language words were introduced by S K Hasnain (Beg & Hasnain, 2008, p. xx). A collection of 150 unique samples obtained from 15 distinct speakers served as the foundation for the speech feature extraction. Matrix Laboratory (MATLAB) was used to create the feed-forward neural networks for voice recognition. In this study, the author made an effort to detect spoken Urdu words using a neural network. The obtained data's Discrete Fourier Transform (DFT) was utilized to train and test the voice recognition neural network. The network was quite accurate in its predictions.

According to Vikhe, accurate endpoint identification is crucial for isolated word recognition systems for two reasons, it is crucial for reliable word recognition and it reduces the amount of computing required to analyze speech when the endpoints are accurately found (Vikhe, 2011). The experimental database consisted of zero to nine English-language numbers.

This project aims to build a prototype that can conventionally convert Burmese speech into text. This project will be a novice because this project is based on a deep learning model and a small amount of data. The more data feed to the model, the better the prediction result. Almost all existing Burmese speech-to-text models are left halfway and neither tries to maintain it nor make the system better. These kinds of models need a huge amount of data and need to be maintained over time. Some models are built with very old designs and cannot perform very well these days. This project will be built with up-to-date techniques and always maintain based on the feedback of the users. Possible challenges faced are collecting the data, a huge amount of data storing and pre-processing, modeling with up-to-date techniques, and finally implementation on the website. However, the biggest challenge is to run the prototype and get a satisfactory result. This prototype can be developed better in the future with more good quality data and using Natural Language Processing (NLP) to get more accurate prediction results.

3. Methodology

3.1. Overview

Before anything is started, the first thing to do is research the related project ideas and methodology. In order that this project will be done quickly and efficiently. However, researching is not an easy task because there are a ton of papers and articles that discuss this idea, and most of them are written with the same principles and ideas and some are not that effective. The same methodology and same techniques are applied in most of the research papers. In addition, there are quite a few papers that really stand out from the others. Therefore, reading one paper after another is really tough because of the analysis. The issue is analyzing every paper to know whether these papers are useful for this project or not. Some of the papers used completely the same method, which is why it is a waste of time. After all the hard work, there are very few papers that really can be helpful for the project.

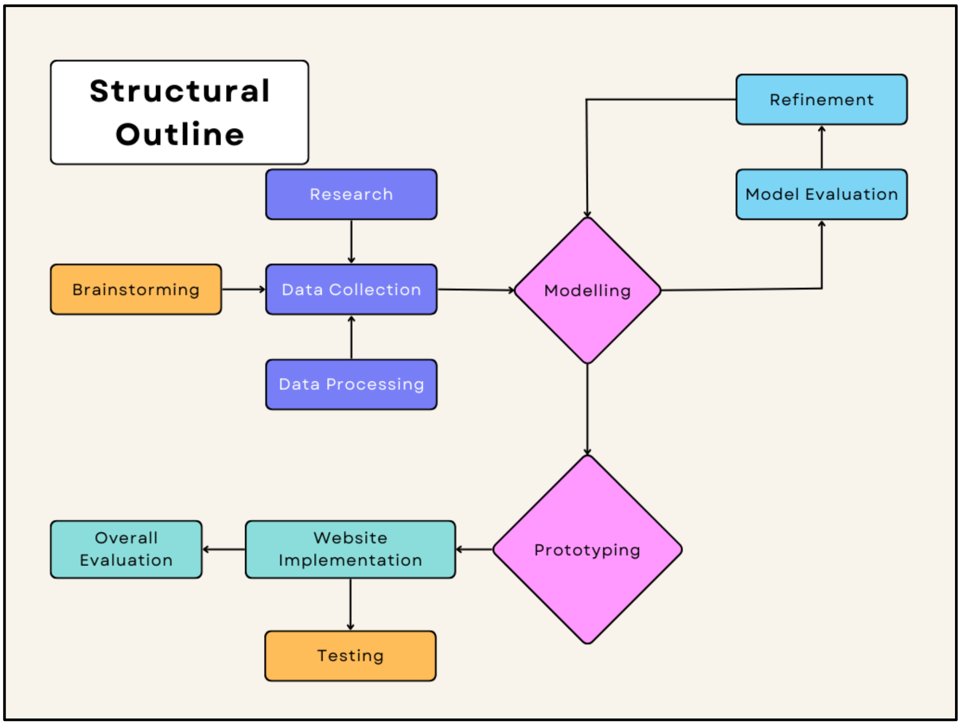


Figure 3.1. Methodology Structural Outline.

After researching and analyzing, another thing is to collect the data sample, in this case, speech data samples. As an advantage of research, there are some websites that provided free, already clean, speech data samples. Additionally, another source is the YouTube Burmese podcasts and audiobooks. The audiobooks are really long and need to be clean for usage but they are really effective data samples because they speak in the clearest as possible to make others people understand. In this way, these audiobooks are excellent materials for data samples.

Data processing is the process of cleaning the data samples to maximize the output of the algorithm. For this case, the audiobooks need to clean because they will be downloaded in MPEG Audio Layer 3 (mp3) format and they need to be converted into Waveform Audio File Format (wav). Another thing to do is trim the data samples because there are times that people don't speak for a few seconds not only in conversation but also in audiobooks. Then, trim and cut the timeframes to reduce the size of the dataset to boost the algorithm run time.

The modeling stage is the most important and difficult part of the project. There are tons of deep learning models that can help with ASR for any language. But the only problem is that it takes a lot of time to write lines by lines of code which is very confusing. Because these models need to be finely tuned depending on the language that is going to be predicted. To overcome this problem, by using the pre-built models. For this project, Wav2vec2 by MetaAI (Meta AI, 2020). Due to a self-supervised training method, which is a relatively novel idea in this industry, Wav2Vec 2.0 is one of the most advanced models for automatic speech recognition currently available. With this method of training, we may pre-train a model using unlabeled data, which is always easier to acquire. The model may then be adjusted for a particular purpose on a given dataset (Sus, 2021). This model used Connectionist Temporal Classification (CTC) and is pre-trained for multiple languages and can be implemented in any ASR project for training with new datasets by users (Platen, 2021). Training the model takes a lot of time and it depends on how much training data will be used also. After training, evaluate the model performance, and train again for more accuracy with more training data. If the model is giving out a decent result then the model is ready for prototyping.

After the modeling stage, now it is time to prototype the system.

To prototype the system, the best way to do it is to upload the trained model to the server and reuse it anytime when needed it. By doing this, the model stays on the server with the latest training performance and it makes the model test in various ways. This way it will be easier to implement in any project or even easily implement on the website later.

After all the hard work, there will be another hard work and the confusing stage is only left which is to implement the system into the web interface, a website. Furthermore, making and hosting that website on the internet in order to access by the whole world.

The last stage of the project is to evaluate the prototype through the website. Adjusting the appearance of the website added more data samples for better accuracy, trained the model for longer, and optimized the models in the modeling stage. These actions will be repeated to maintain the prototype and website.

3.2. Hardware Requirements

Table 3.1. Hardware Specifications.

Component	Specifications	
Laptop or Desktop	Operating System	Windows 10 Home
	Central Processing Unit	AMD Ryzen 5
	Memory	16 GB
	Hard Disk Drive	1TB
	Solid-state Drive	256 GB

3.3. Software Requirements

Table 3.2. Software Specifications.

Software	Version
Anaconda	2021.11
Microsoft Visual Studio	1.74.2
cmd	10.0.19045.2546

3.4. Programming Languages

Table 3.3. Programming Language.

Programming Language	Usage
Python	Main programming language use from start to finish, model building, training, evaluation, and deployment, all are done in a single language.

3.5. Project Timeline

Table 3.4. Timeline.

Week	DESCRIPTION
Week 1	Article Collection
Week 2	Article Analysis
Week 3	Literature Review
Week 4	Methodology
Week 5	Proposal Submission
Week 6	Project Pitching
Week 7	Report Submission
Week 8	Speech Dataset Collection
Week 9	Speech Dataset Collection
Week 10	Data Pre-processing
Week 11	Data Pre-processing
Week 12	Modelling
Week 13	Modelling
Week 14	Prototyping
Week 15	Website Creation
Week 16	Prototype Implementation into Website
Week 17	Prototype Implementation into Website

Week 18	Evaluation
Week 19	Project Pitching
Week 20	Overall Refining
Week 21	Final Report Submission

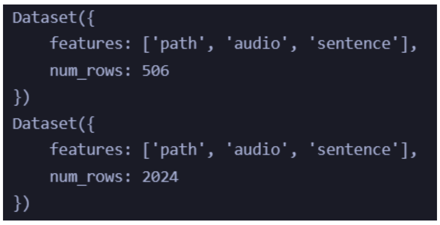
3.6. Project Development

3.6.1. Data Collection

This Burmese ASR project is using the dataset called SLR80 from the popular dataset provider OpenSLR (Oo et al., 2020). This dataset contains a total of 2530 clean speech samples by 20 speakers and all the speakers are female. The dataset also provides a clean transcription of each speech sample. The combination of all speech data samples will be around 5 to 6 hours. For an ASR model, this amount of hours is not ideal but considering the computer specifications, this amount is good enough to at least train and get some output. As the dataset is small, the training time will be more for better accuracy of prediction.

3.6.2. Data Processing

Before model training, data processing will set up the data as needed. Divide the dataset into two groups, train, and test, as the initial step in the data processing. The remaining 2024 samples, or 80% of the 2530 samples, are utilized for training, while 506 samples, or 20%, are used for testing.



```
Dataset({
    features: ['path', 'audio', 'sentence'],
    num_rows: 506
})
Dataset({
    features: ['path', 'audio', 'sentence'],
    num_rows: 2024
})
```

Figure 3.2. Train and Test Split.

As this model is supervised by label data, it is required to delete some characters that do not really have a sound, such as special characters and spaces, after separating the dataset into train and test. Another step is to provide all the unique characters from the Burmese language to the CTC algorithm. From the transcription of datasets, the total unique characters in the transcription are 60 characters, and also manually added two more characters that are unknown characters and padding to the start and end of the transcription. These two characters are specially for the CTC algorithm to improve performance. Eventually, the total of unique characters is 63 characters. After that, load all the characters into Wav2Vec2CTCTokenizer, this will tokenize the transcription sentence into characters. Now, preprocessing for text data is done and now it is time for audio data. For the audio data, the data set is already trimmed and short but the sampling rate is 48KHz. The sampling rate required by the CTC algorithm is 16KHz. So, change the sampling rate of the whole dataset.

3.6.3. Modelling

For the modelling, there are a few steps needed to prepare before starting the training process. First of all, step up the Wav2Vec2FeatureExtractor, this feature extractor will do the jobs such as pre-processing audio files to Log-Mel Spectrogram features, padding, normalization, and conversion to PyTorch tensors. After that, set up the Wav2Vec2Processor, it is a wrapper of tokenizers and feature extractors.

As the most important process for almost all deep learning models, the Word Error Rate (WER) function is defined and created. Finally setting up the Wav2Vec2ForCTC model, this model is fined tuned and pre-trained by MetaAI. Since this model is so big that trained with a large amount of data from multiple languages, for this project downscaling the parameters is an important thing to do. In downscaling, the layer drop parameter is set to zero since the dataset for training is so small. And assign the Wav2Vec2Processor to the algorithm.

It's time for training now that the model has been defined. There are a few factors that must be specified before training. The first configuration involves dividing the entire training data by the batch size and setting it to 100. Instead of using epochs, the Wav2Vec2ForCTC model uses steps. The epoch number must still be included, though. After setting up the parameters, call the train function and it will start the training process. The training process will take a huge amount of hours even if the dataset is small. After 2 days worth of training the model, the model is start to recognize the speech and give some output text. The final result of the training process is shown in Table 4.1.

Table 4.1. Training Result.

Epoch	Training Loss	Learning Rate	Evaluation Loss	WER	Training Time
1.58	10.3152	6.00E-05	4.660822	1	8 - 9 hours
3.17	3.9002	0.00012	3.500513	1	8 - 9 hours
4.76	3.4987	0.00018	3.450787	1	8 - 9 hours
6.35	3.4524	0.00024	3.372131	1	8 - 9 hours
7.93	3.128	0.0003	1.946949	1	8 - 9 hours
9.52	1.3368	0.000233	0.684924	1	8 - 9 hours

3.6.4. Prototyping

After the modelling is done, the next step is to prototype the model so that the model is useable to do the prediction. In order to do that, there is a good and reliable website called Hugging Face that pretty much looks like a GitHub. Upload the model there and anyone can download the model, not the code, by downloading the model, it does not need to train the model again, it can be used instantly after downloading the model from the local directory and only take a few lines of codes to predict the speech.

Download the model and create a function to predict the speech. To predict the speech, the speech file needs to be put inside one folder then call the path of that speech file and change the sampling rate of the speech file. And changing the speech file into PyTorch tensors and then finally prediction is taken place and as a result, the text will be returned.

3.6.5. Deployment

Now everything is ready and the only thing left to do is to make the prototype useable through the website. Implementing the Python code and function into the website is a bit confusing and there are a ton of programming languages that do that job but for this project, Streamlit is used. Streamlit is a python library that makes the python code into a web interface but the downside is that Streamlit does not have that much of a visually appealing function. So, the website will be very simple visual.

4. Result and Discussion

4.1. Overview

In Deep learning or neural networks, supervised learning is better than unsupervised learning because of its controllability. But at the same time, the model needs detailed parameters to handle all the processes. A very good ASR needs thousands of hours of speech data with good transcription. This is a downside of ASR because these data are so big and to get a clean transcription of each speech data is a great effort that can not be done in a given time. Even after that, to handle these data and model training, a very powerful computer is needed because the training process of ASR models takes days or weeks. A normal personal computer cannot perform these powerful processes.

After long hours and harsh process of training, the overall training result is gathered as shown in Table 4.1. As the table shows, there is only a total of six training and evaluation epochs, however, to get that six epochs, the computer is going through a very powerful process for more than 48 hours. Even though the dataset is small and the training time is only 48 hours, the training loss, evaluation loss, and learning rate are getting higher and higher results. But the only thing that did not change throughout the epochs is WER, it is consistently staying at one.

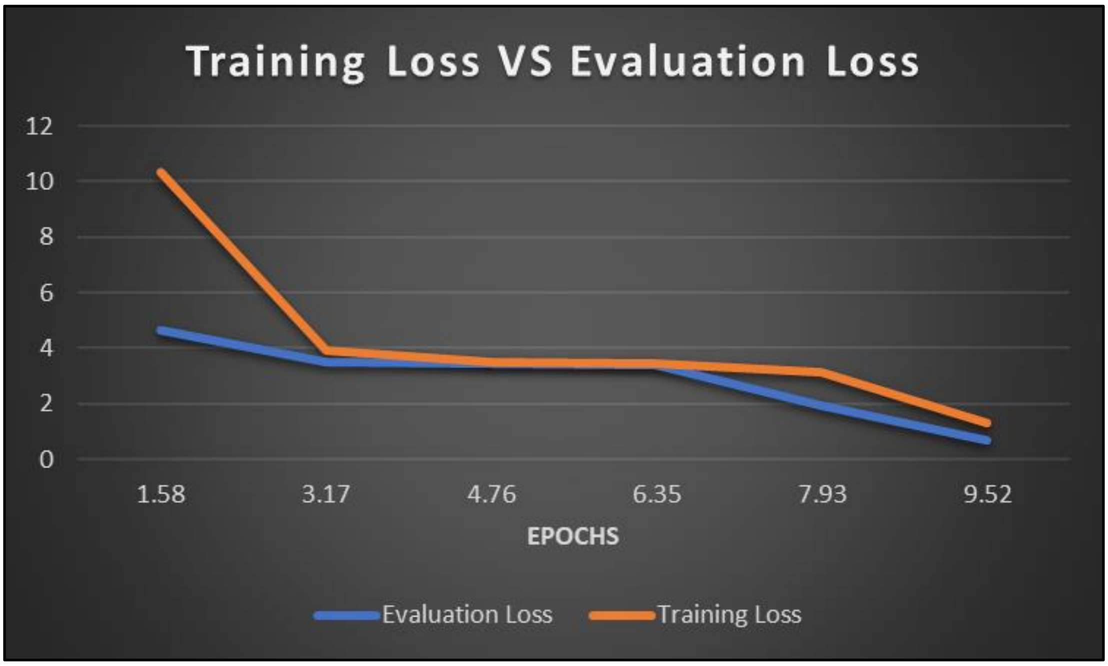


Figure 4.1. Training Loss VS Evaluation Loss.

As from the figure above, at the first epochs of training, the loss during training is really high which is above 10, but at the same epoch, the loss during the evaluation is only over four. That is because the training samples are four times higher than the testing samples. Overall, as the epochs go higher, the loss for both training and evaluation is getting decreased. At the last epoch of six, both training loss and evaluation loss are under two.

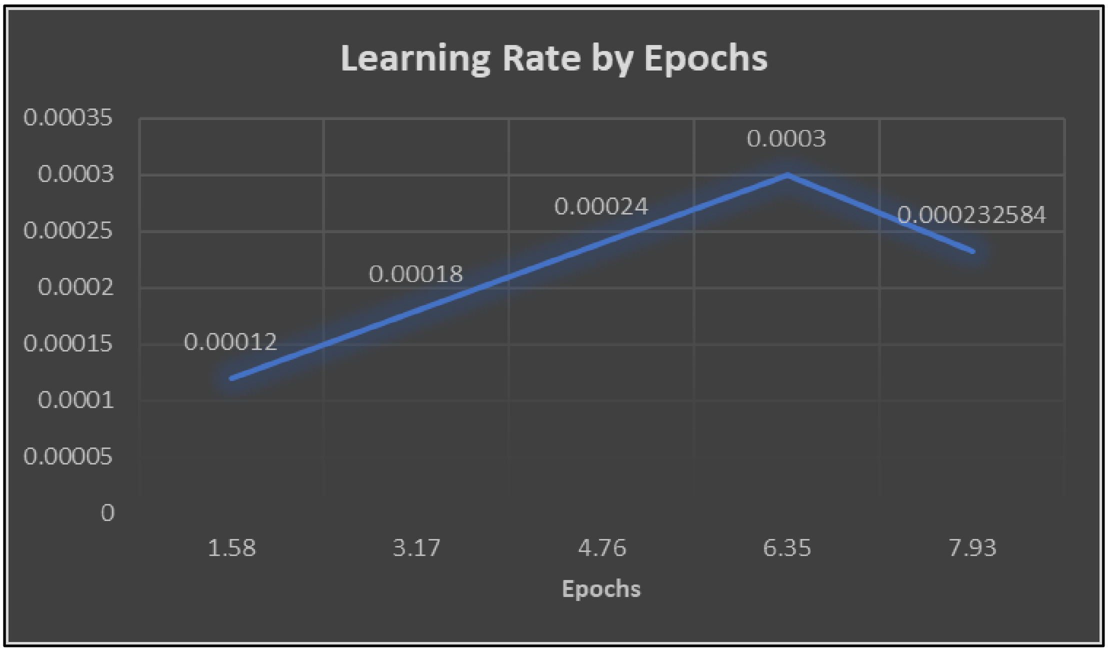


Figure 4.2. Learning Rate by Epochs.

The learning rate of the models is also a great factor that determines the accuracy of the model. In this case, the learning rate is not very significant but slowly it gets a good result as the epochs go higher. The learning rate is very small that is under zero, it is because the small dataset is getting at the very big model, Wav2Vec2ForCTC. The learning rate is rather unstable compared to the training loss and evaluation loss since at epoch six, the learning rate is going down in values compared to the values at epoch five.

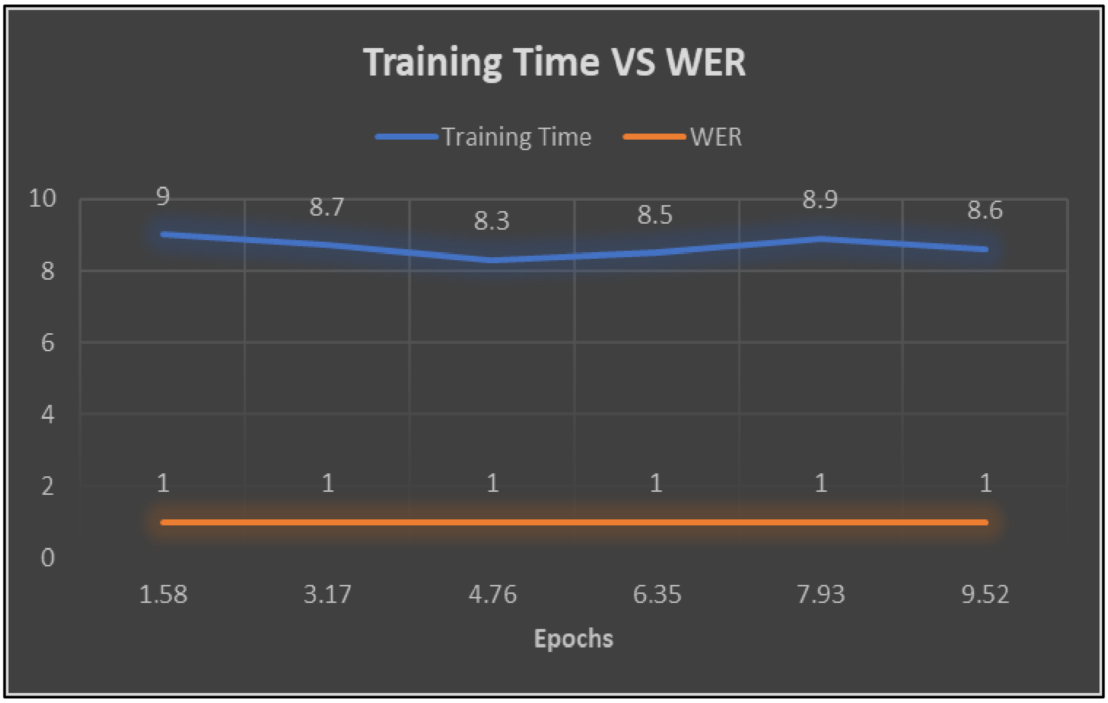


Figure 4.3. Training Time Vs WER.

After training the model for more than 48 hours, the model is getting in good shape from every aspect except the WER. The WER is staying at one consistently through the epochs. The training time required between eight to nine hours to finish one epoch, it is a very time-consuming and very powerful huge task for a computer to handle. However, as the fact from table 4.1, it is good enough to assume that the model is going to a be good model but slowly.

4.2. Website Deployment

After all the processing, modelling, training and result analysis, the last step and the best way to showcase all the work that has been done along the way is through the website deployment. Abstract of the website deployment is all about creating the website and integrating the model and maintaining that website over time. It is mostly making the website work faster, perform well, and give the user a seamless experience. To create such kind of website through the Python programming language, the streamlit is the best and easy tool to use. Streamlit is a Python library that provides easy solutions and implementation for creating websites.

Streamlit has many advantages using as ease of implementation, detailed documentation and a huge community but there are also disadvantages such as limited features and poor visuals. The basic streamlit visual main interface is easy to create according to the need of the user and Figure 4.4 shows the simple and basic user interface for this project.

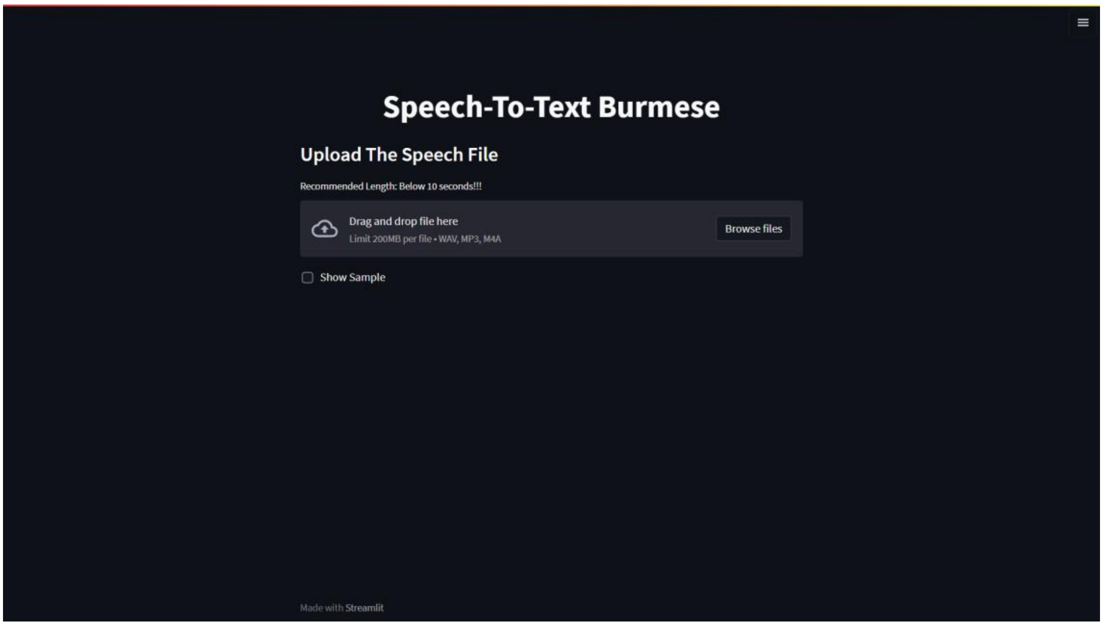
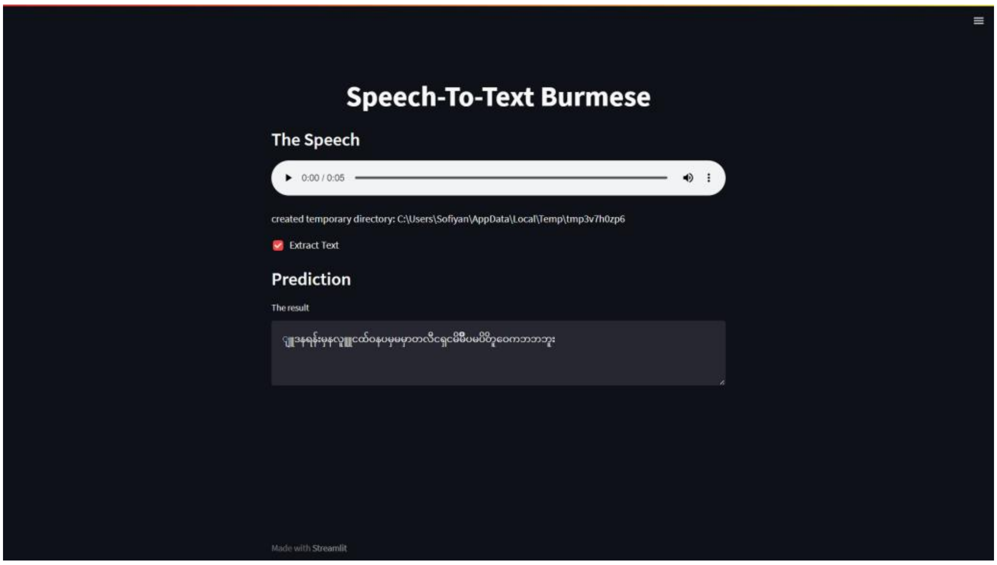


Figure 4.4. Basic User Interface.

According to the Figure 4.4, there are two main options that are available to the user, the first one is to upload the speech file as needed and the second one is to see the examples. The first one is dedicated to the main function of the website, there, users can upload the speech file and listen to the uploaded speech file and an option to extract the text from the uploaded speech file as shown in Figure 4.5.



Speech-To-Text Burmese

The Speech

created temporary directory: C:\Users\Sofiyan\AppData\Local\Temp\tmp3v7h0zp6

☑ Extract Text

Prediction

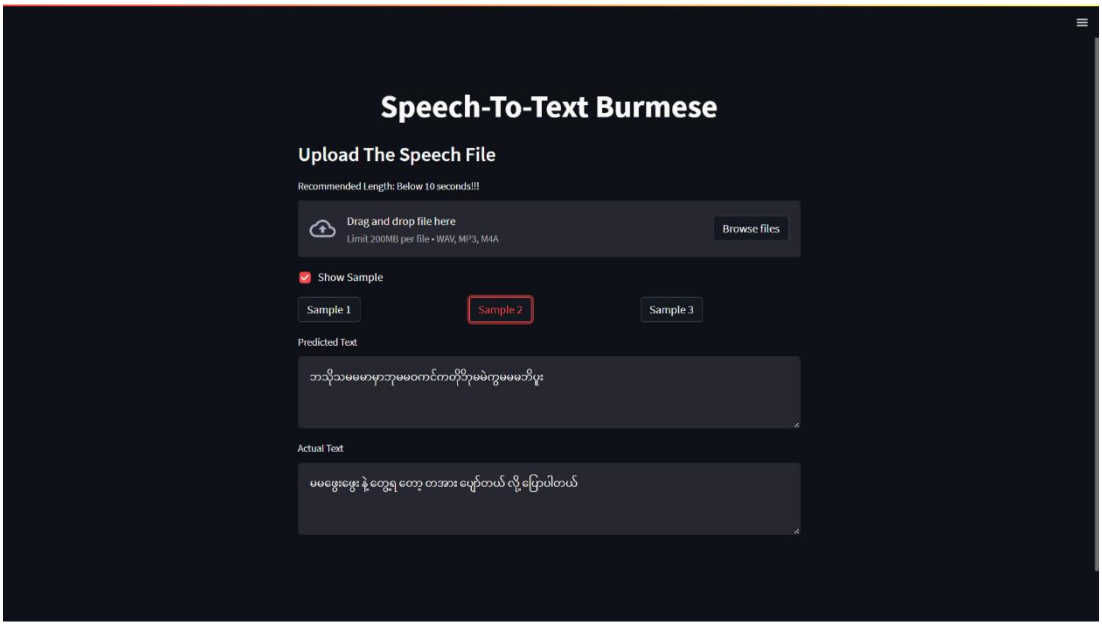
The result

ျ့ဒနရန်းမှနလူျငထ်ဝနပမှမမှာတလီငရှငမိမိပမပိတုဝေကဘဘဘူး

Figure 4.5. Uploaded Speech File and Extracted Text.

When the users upload the speech file for extracting, the uploaded speech will go through the data pre-processing step basically changing the uploaded speech file sampling rate to 16000 Hz because the model is trained on the speech samples of 16000 Hz. There is another condition to change the uploaded speech file that is the speech file must be stored in a temporary folder on the local computer. This local temporary folder will be stored just for a few seconds, as soon as the processing and extracting of the uploaded speech file is complete, this local temporary folder will be deleted by itself.

As Figure 4.4 shows, there is another option to choose from if the users do not want to upload the speech file, they can view the pre-existed speech files and their transcriptions. Users can click the "Show Sample" checkbox and three samples will show on the screen. Users can choose one of the samples and view the transcriptions. This process is shown in Figure 4.6.



Speech-To-Text Burmese

Upload The Speech File

Recommended Length: Below 10 seconds!!!

☑ Show Sample

Predicted Text

ဘသိုသမမမာမှာဘုမမဝကင်ကတိုဘိုမမဲကွမမမဘိပူး

Actual Text

မမဖွေးဖွေး နဲ့ တွေ့ရ တော့ တအား ပျော်တယ် လို့ ပြောပါတယ်

Figure 4.6. Pre-existed Samples.

5. Conclusion

5.1. Overview

The Burmese language is so complex and consists of Tones, Stress, Vowels, and Consonants. The Myanmar alphabet consists of 33 letters, 12 vowels, and ten numerals. Even though this language is too complex, the model is trying to convert the speech to text with just the characters and speech mapping. The model is not depending its accuracy from the lexicon or dictionary to predict the text. It is obvious that the model is needed of more clean training samples. The amount of 2530 samples is not a good amount for the model to solely depend on the characters and speech mapping.

5.2. Social Innovation

In several industries, speech recognition technology is transforming the way companies do business. Speech recognition is becoming more prevalent in everyday life, including the workplace. Businesses' adoption of chatbots and virtual personal assistants, as well as consumers' increased usage of voice-enabled devices, has fueled the incorporation of speech-to-text technologies at work. The main advantage of voice recognition software is increased productivity. Users may dictate papers, email answers, and other content without having to manually enter them into a machine. Using speech-to-text technology eliminates one barrier between a user's ideas and their digital output, which may help to improve corporate operations, save time, and ultimately enhance productivity. Employees may be more productive in their professions and focus on higher-value activities with the help of voice technologies. It means your company will obtain critical information sooner, enhancing overall productivity.

Businesses that use speech processing technology in their products might appeal to a broader audience. For example, many business applications now demand users to interact with the app using their hands to some extent, which might be challenging for particular users with specific impairments. Businesses that invest in advanced voice recognition technology automatically have a competitive advantage over competitors with less accessible products. For most individuals, talking is still faster than typing or crossing the room to perform a task. As a result, Amazon, Apple, and Google have integrated voice recognition into their digital assistants. Giving a voice command is faster than getting out of bed and turning up the thermostat. The incorporation of voice commands into an increasing number of smart products simplifies the user experience (again, with the potential to break into underserved markets, as previously indicated), and companies profit from that simplicity of use. Voice recognition may even have the ability to assist various departments to execute work more quickly at the company level.

Voice recognition helps in communicating between persons who speak various languages. Speaking in one's language and having it processed by voice recognition software and then translated audibly or graphically opens up a whole new world of communication possibilities. This can assist multinational enterprises in engaging in more meaningful interactions without the requirement for a translation. It might also be used in foreign hotels and business centers to assist passengers overcome language barriers and having easier access to information.

In Myanmar, there are a lot of elderly people who cannot type the word they want to communicate with their family members or relatives. This is because they are used to using keypad mobile phones and smartphones are not an option for them. However, technology is changing and the keypad is less and less popular in everyday life. Most elderly people change from their keypads to smartphones but they can only call and use social media, the features of smartphones are too complicated for them, and typing is one of them. In most cases, texting is much better than calling, which is why they can benefit from speech-to-text systems. They just need to talk about what they want and the system will give out the text.

5.3. Future Enhancement

Not any model is perfect, it is true for this model also but moreover this model has more room for development. As from Table 4.1, the numbers are not lying as the model is to train more, and then the model will become better. There are only two ways to improve this model, the first way is to gather more data samples. At the same time, the data samples must be very clean and the transcription of that samples must be clear of noise and accurate. The more data feed into the model for training the that a true fact that cannot deniable.

Another way to train the model is by giving tons of time training. It is recommended on a powerful computer to save some time. It is also recommended to train on the GPU because it has more computational power than the normal CPU. If the model is going to train on the GPU then there is one parameter to turn on before the training. In Figure 4-5, change the value of fp16 into True. This is simply changing the training and evaluation process to GPU.

Overall, the model is performing really well with just small amounts of training samples. At last, the model is able to predict the text from a speech input even though the WER is still high. Investigating how speech recognition works with the use of deep learning models and execution is achieved by building the Burmese ASR model. The model is capable of recognizing the speech and conventionally converted into text. Not only that the model is available to everyone through the website for the public.

References

Acharjya, D. P., Mitra, A., & Zaman, N. (2022). Deep learning in data analytics. Springer International Publishing.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33, 12449-12460.

Beg, A., & Hasnain, S. K. (2008). A speech recognition system for Urdu language. Wireless Networks, Information Processing and Systems, 118-126. https://doi.org/10.1007/978-3-540-89853-5_14

Chit, K. M., & Lin, L. L. (2021). Exploring CTC based end-to-end techniques for Myanmar speech recognition. Advances in Intelligent Systems and Computing, 1038-1046. https://doi.org/10.1007/978-3-030-68154-8_87

Fatima-tuz-Zahra, N. Jhanjhi, S. N. Brohi, N. A. Malik and M. Humayun, "Proposing a Hybrid RPL Protocol for Rank and Wormhole Attack Mitigation using Machine Learning," 2020 2nd International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2020, pp. 1-6, doi: 10.1109/ICCIS49240.2020.9257607.

Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2010). A review on speech recognition technique. International Journal of Computer Applications, 10(3), 16-24.

Gopi, R., Sathiyamoorthi, V., Selvakumar, S., Manikandan, R., Chatterjee, P., Jhanjhi, N. Z., & Luhach, A. K. (2022). Enhanced method of ANN based model for detection of DDoS attacks on multimedia internet of things. Multimedia Tools and Applications, 1-19.

Gouda W, Almurafeh M, Humayun M, Jhanjhi NZ. Detection of COVID-19 Based on Chest X-rays Using Deep Learning. Healthcare. 2022; 10(2):343. https://doi.org/10.3390/healthcare10020343

Gu, N., Lee, K., Basha, M., Ram, S. K., You, G., & Hahnloser, R. H. (2024, April). Positive Transfer of the Whisper Speech Transformer to Human and Animal Voice Activity Detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7505-7509). IEEE.

Humayun M, Ashfaq F, Jhanjhi NZ, Alsadun MK. Traffic Management: Multi-Scale Vehicle Detection in Varying Weather Conditions Using YOLOv4 and Spatial Pyramid Pooling Network. Electronics. 2022; 11(17):2748. https://doi.org/10.3390/electronics11172748

Humayun M, Sujatha R, Almuayqil SN, Jhanjhi NZ. A Transfer Learning Approach with a Convolutional Neural Network for the Classification of Lung Carcinoma. Healthcare. 2022; 10(6):1058. https://doi.org/10.3390/healthcare10061058

Hussain, K., Rahmatyar, A. R., Riskhan, B., Sheikh, M. A. U., & Sindiramutty, S. R. (2024, January). Threats and Vulnerabilities of Wireless Networks in the Internet of Things (IoT). In 2024 IEEE 1st Karachi Section

Humanitarian Technology Conference (KHI-HTC) (pp. 1-8). IEEE.

Hussain, S. J., Irfan, M., Jhanjhi, N. Z., et al. (2021). Performance enhancement in wireless body area networks with secure communication. Wireless Personal Communications, 116(1), 1–22. https://doi.org/10.1007/s11277-020-07702-7

Johnson, M., Lapkin, S., Long, V., Sanchez, P., Suominen, H., Basilakis, J., & Dawson, L. (2014). A systematic review of speech recognition technology in health care. BMC medical informatics and decision making, 14, 1-14.

Kamath, U., Liu, J., & Whitaker, J. (2019). *Deep learning for NLP and speech recognition* (Vol. 84). Cham, Switzerland: Springer.

Khaing, I., & Linn, K. Z. (2013). Myanmar continuous speech recognition system based on DTW and HMM. *International Journal of Innovations in Engineering and Technology (IJIET)*, *2*(1), 78-83.

Lim, M., Abdullah, A., Jhanjhi, N. Z., Khan, M. K., & Supramaniam, M. (2019). Link prediction in time-evolving criminal network with deep reinforcement learning technique. IEEE Access, 7, 184797-184807.

Mallick, C., Bhoi, S. K., Singh, T., Hussain, K., Riskhan, B., & Sahoo, K. S. (2023). Cost Minimization of Airline Crew Scheduling Problem Using Assignment Technique. International Journal of Intelligent Systems and Applications in Engineering, 11(7s), 285-298.

Majid, M., Hayat, M. F., Khan, F. Z., Ahmad, M., Jhanjhi, N. Z., Bhuiyan, M. A. S., ... & AlZain, M. A. (2021). Ontology-Based System for Educational Program Counseling. Intelligent Automation & Soft Computing, 30(1).

Mallick, C., Bhoi, S. K., Singh, T., Swain, P., Ruskhan, B., Hussain, K., & Sahoo, K. S. (2023). Transportation Problem Solver for Drug Delivery in Pharmaceutical Companies using Steppingstone Method. International Journal of Intelligent Systems and Applications in Engineering, 11(5s), 343-352.

Minn, K. H., & Soe, K. M. (2019). Myanmar word stemming and part-of-speech tagging using rule based approach (Doctoral dissertation, MERAL Portal).

Mon, A. N., Pa Pa, W., & Thu, Y. K. (2018). Improving Myanmar automatic speech recognition with optimization of convolutional neural network parameters. *International Journal on Natural Language Computing (IJNLC) Vol*, *7*.

Mon, A. N., Pa, W. P., & Ye, K. T. (2019). UCSY-SC1: A Myanmar speech corpus for automatic speech recognition. *International Journal of Electrical and Computer Engineering*, *9*(4), 3194.

Naing, H. M. S., Hlaing, A. M., Pa, W. P., Hu, X., Thu, Y. K., Hori, C., & Kawai, H. (2015, December). A Myanmar large vocabulary continuous speech recognition system. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (pp. 320-327). IEEE.

Naing, H. M. S., Hlaing, A. M., Pa, W. P., Hu, X., Thu, Y. K., Hori, C., & Kawai, H. (2015, December). A Myanmar large vocabulary continuous speech recognition system. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (pp. 320-327). IEEE.

Oo, Y. M., Wattanavekin, T., Li, C., De Silva, P., Sarin, S., Pipatsrisawat, K., ... & Gutkin, A. (2020, May). Burmese speech corpus, finite-state text normalization and pronunciation grammars with an application to text-to-speech. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 6328-6339).

Rabiner, L. R., Wilpon, J. G., & Soong, F. K. (1989). High performance connected digit recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*(8), 1214-1225.

Saeed, S., Jhanjhi, N. Z., Naqvi, M., Malik, N. A., & Humayun, M. (2019). Disparage the barriers of journal citation reports (JCR). International Journal of Computer Science and Network Security, 19(5), 156-175.

Shah, S. A. A., Bukhari, S. S. A., Humayun, M., Jhanjhi, N. Z., & Abbas, S. F. (2019, April). Test case generation using unified modeling language. In 2019 International Conference on Computer and Information Sciences (ICCIS) (pp. 1-6). IEEE.

Soe, W., & Theins, Y. (2015). Syllable-based Myanmar language model for speech recognition. IEEE Xplore. Retrieved from https://ieeexplore.ieee.org/document/7166608

Sowjanya, A. M. (2021). Self-Supervised Model for Speech Tasks with Hugging Face Transformers. *Turkish Online Journal of Qualitative Inquiry*, *12*(10).

Tabbakh, A. (2021). Bankruptcy prediction using robust machine learning model. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(10), 3060-3073.

Thein, Y. (2010). High accuracy Myanmar handwritten character recognition using hybrid approach through MICR and neural network. *International Journal of Computer Science Issues (IJCSI)*, *7*(6), 22.

Thu, Y. K. (2021, April 17). myG2P. GitHub. Retrieved from https://github.com/ye-kyaw-thu/myG2P

Tucci, L., Laskowski, N., & Burns, E. (2019). A guide to Artificial intelligence in the enterprise.

V. Singhal et al., "Artificial Intelligence Enabled Road Vehicle-Train Collision Risk Assessment Framework for Unmanned Railway Level Crossings," in IEEE Access, vol. 8, pp. 113790-113806, 2020, doi: 10.1109/ACCESS.2020.3002416.

Willyard, W. (2022, January 28). What role does an acoustic model play in speech recognition? | Rev blog. Rev Blog. https://www.rev.com/blog/resources/what-is-an-acoustic-model-in-speech-recognition

Xu, S., Liu, X., Ma, K., Dong, F., Riskhan, B., Xiang, S., & Bing, C. (2023). Rumor detection on social media using hierarchically aggregated feature via graph neural networks. Applied Intelligence, 53(3), 3136-3149.

Yong, C. T., Hao, C. V., Riskhan, B., Lim, S. K. Y., Boon, T. G., Wei, T. S., Balakrishnan, S., & Shah, I. A. (2023). An implementation of efficient smart street lights with crime and accident monitoring: A review. Journal of Survey in Fisheries Sciences.

Yu, D., & Deng, L. (2015). Automatic speech recognition. Signals and Communication Technology. https://doi.org/10.1007/978-1-4471-5779-3

Zaman, G., Mahdin, H., Hussain, K., Atta-Ur-Rahman, Abawajy, J., & Mostafa, S. A. (2021). An ontological framework for information extraction from diverse scientific sources. IEEE Access, 9, 42111-42124. https://doi.org/10.1109/ACCESS.2021.3063181

Zhong, G., Wang, L., Ling, X., & Dong, J. (2016). An overview on data representation learning: From traditional feature learning to recent deep learning. The Journal of Finance and Data Science, 2(4), 265-278. https://doi.org/10.1016/j.jfds.2017.05.001

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.