Speech to Text by Using the Sindhi Language

Naadiya Khuda Bux[1], Ambreen Khan[1], Khuda Bakhsh[1]
[1]Research Lab of Information Security and AI, Benazir Bhutto Shaheed University, Karachi, 75660, Sindh, Pakistan.
E-mail: naadiya.khudabux@yahoo.com, ambreen.khan@bbsul.edu.pk, khudabukhsh1992@gmail.com

Corresponding Author: Naadiya Khuda Bux, naadiya.khudabux@yahoo.com

Received: 02-07-2022; **Accepted:** 23-08-2022; **Published:** 29-08-2022

Abstract: We live in the era of technology, and advancement in technology is growing exponentially. In Pakistan, especially in Sindh, many people prefer to speak than write. People are not well aware of computational and other global languages. So, that's why they face so many difficulties typing and then converting it into the Sindhi language. Especially in offices/organizations where Sindhi is the first language used in speaking and typing. There drafting huge consumes too much time. They face many difficulties such as finding spelling correct words and so on. People with medical deficiencies and disabilities will also get a beneficial source of help from this tool. This tool can handle all these difficulties and solve all the discussed problems. This project aims to develop a web-based application that tries to overcome the disadvantages of the other available applications. The application is generic, meaning it may work solely for a specific regional language speaker in any country in the world. The main objective of the work presented throughout this report is to develop an enterprise and open platform for the nation. We are using the Convolutional Neural Network and API in the development phase. Advance python libraries detect the user's speech, and then the conversion will take one into text.

Index Terms: Speech to Text, Sindhi Natural Language Processing, Native Language, Convolutional Neural Network.

1. Introduction

Over 2300 languages are spoken in Asia, yet the environment for natural language processing does not adequately account for these languages. Speech to text by using the native language is an advancement in technology to evaluate and digital language divide [1]. One of the most revolutionary and significant technologies for languages worldwide is Speech Recognition [2], which helps advance language computing. Most languages have either already developed or are in the process of developing speech recognition systems [3, 4]. Because the number of characters and sounds in the Sindhi language is greater than in other languages, making it unique, most existing approaches cannot be used for the Sindhi language [5]. As a result, for designing and developing Sindhi Speech Recognition tools, a different approach in correspondence Sindhi sounds and characters is required. Speech-to-text conversion is a demand of this era where people worldwide are now a part of technology and using it for daily work. Typing text in other languages is a significant issue as people are not well aware of English [6] and other languages but have sound knowledge of their native language [7].This project aims to use Sindhi speech as input and then convert that speech into written text.

Speech to text project targets only specific Sindhi language and civilization. Native people are good at the sindhi language, but if we talk about other languages, most people don't know. Thus, the applications help you to use your language and get the result of that speech in written form. The main objective of the work presented throughout this report is to develop a web-based application for the users to access the application's services to fulfil their desired results according to their needs. The speech recognition engine is a software component that performs the recognition process. The primary function of the speech recognition engine is to process spoken input and convert it into text that an application can understand. This application will be a nationwide educational & civilizational revolution as the Sindhi language will nurture in the technology field. The Sindhi language is already known worldwide, but this project is very fruitful for improving the language.

Numerous researchers have researched different native languages to develop native language apps in previous research [8, 9]. Speech to text using the native Sindhi language has some limitations as we consider only one language in our project. The developed language model is valid for most use of that specific Sindhi language, but in future, we will be adding more regional languages. Since the model is based on the Sindhi word dataset, it will not support continuous speech.

Our application attempts to make a user-friendly system and provides a platform where the native language will grow in this era of technology. We are intent on making an application that will help the users understand things in their regional language. NLP has a broad scope, with so many uses in customer service, grammar check software, business marketing, etc. The system will be designed considering the users' needs and the psychology of regional people so everyone can easily access and use it. Furthermore, this has documented social and environmental benefits that include:

- It saves typing time as the app directly gets input in voice and interprets it into text
- It will affect a million lives of regional people as no one will depend on others to help them type.
- As the system requirements involve software and other utilities that are freely available as open source, the completion of this system is feasible for the team to complete project.
- The app will be accessible at a low price so everyone can benefit from it.
- As the system aims to provide a platform for regional speech-to-text conversion, it can be used in drafting to give employment chances to regional people.

1.1 Motivation and Contribution

The project is stated with the sole aim that nearly 20% of people suffer from various disabilities. Many of them are blind or unable to use their hands effectively. So, these kinds of people always stay dependent on others. In Pakistan, so many cities are still way backwards, living lives who are still illiterate and unable to write, so they prefer to speak as compared to writing. That's why we are developing this system to help Sindhi students and Sindhi people.

2. Literature Review

Sindhi is one of Pakistan's major languages, spoken by 30-40 million people [10]. Sindhi is a language that is frequently used on the internet. Sindhi blogs, literary websites, online newspapers, and discussion boards are proliferating on a daily basis. Sindhi is Pakistan's second most widely spoken written language, after Urdu. Sindhi is one of the world's oldest languages, with a history spanning over 1000 years, from 600 BC to 500 AD [11]. According to research, the general impression that the Sindhi language was extracted directly from Sanskrit is incorrect. According to a 1998 assessment by Pakistan's legislature, the education rate in Sindh is 45.29%. For centuries, the Sindhi voice had advanced. The voice of Sindh's general public interacted with Aryan, and it evolved into Indo-Arayan (Prakrit). The strong presence of Sanskrit and Prakrit in Sindhi dialects is referred to as India's voice. Because it contains some lexis from Dravidian from the Mediterranean, Persian, and Arabic, it is known as Moen-Jo-Daro human advancement. Sindhi is also one of India's perceived authority voices, as it is spoken by approximately 1.2 million people, the majority of whom moved from the region of Sindh. Approximately 4000000 people speak Sindhi as their first language [12]. Speech recognition is a subfield of computational linguistics that develops methodologies and technologies to enable computers to recognise and convert spoken language into text [13, 14]. When compared to Modern Standard Arabic, corpus design for speech synthesis is a well-researched topic in languages such as English [15]. The emphasis is frequently on methods for automatically generating the orthographic transcript to be recorded (usually greedy ways). This paper investigates Modern Standard Arabic (MSA) phonetics and phonology in order to develop criteria for a low-cost method of creating a speech corpus transcript for recording. The dataset size is reduced multiple times using these optimization methods with different parameters, resulting in a much smaller dataset with identical phonetic coverage as before the reduction. This output transcript has been chosen for recording. This is part of a larger project to create a fully annotated and segmented speech corpus for MSA. The suggested approach is to develop a set of language-dependent grapheme-to-allophone rules that can anticipate such allophonic fluctuations and offer a phonetic transcription that is sensitive to the local context to an automatic voice recognition system. The originality of this technique is that each word's pronunciation is taken directly from a context-sensitive phonetic transcription rather than a predetermined dictionary, which may not necessarily reflect the real pronunciation of the wordThe research also tries to enhance acoustic modelling by using the stress feature as a supra-segmental element of speech. The efficacy of the suggested rules was assessed by comparing the performance of a dictionary-based system to that of an automatically generated phonetic transcription. The study discovered that deleting the fixed lexicon and learning the phone probabilities via induced phonetic transcription enhanced system performance by 9.3% on average. Marking the stressed vowels with separate stress markers results in an additional 1.7% improvement [16, 17]. In this work [18], we review the accomplishments done in the field of Arabic voice recognition so far, covering corpora, phonemes, language models, acoustic models, and some intriguing research paths. According to the survey results, the shortage of publicly available continuous speech corpora justifies additional study focus in this field. It also illustrates the need of big corpora or a benchmark in boosting Arabic language research for good human-computer interaction.A Sindhi Unicode-8-based linguistics data set is also multi-class and multi-featured [19]. It was created to address natural language processing (NLP) and linguistic issues in the Sindhi language. The data set contains information on the grammatical and morphological structure of Sindhi language texts, as well as the sentiment polarity of Sindhi lexicons. As a result, data sets can be used for information retrieval, machine translation, lexicon analysis, language modelling analysis, grammatical and morphological analysis, and sentiment and semantic analysis [20].

3. Sindhi Gaelic Speech to Text Keyboard

Sindhi Gaelic Speech to Text Keyboard converts your Sindhi Gaelic speech into text with built-in speech recognition technology. Sindhi Gaelic Speech to Text Typing App is a great tool for typing quickly and easily by saying what you want to write in text by voice and sending or saving it. When people are pressed for time or unable to type, they can simply activate this Sindhi Gaelic speech to text or voice to text keyboard and say whatever they want. Convert Sindhi Gaelic speech to text quickly and easily with this free Sindhi Gaelic speech-to-text tool [21].

3.1 Features of Sindhi Gaelic Voice pad - Speech to Text

In the numerous past features of the Sindhi Gaelic Voice pad - Speech to Text has been derived. For simplicity, these are elaborated here.

- 50+ user interface languages, including Sindhi Gaelic.
- Note down & remind you later at the time you set.
- Also, one-touch quickly to share Sindhi Gaelic Speech in the text to Your friends.
- Integrated with your Android calendar, you don't need to maintain another one.
- It can work when the phone screen is turned off.
- Automatically save your voice notes in a storage file easily to back up the cloud.
- With only one touch, It can constantly receive your speech and convert it to text.
- Headset button to control Start & Stop voice recognition.
- Speech recognition multi-languages.

4. Methodology

Speech to text using a regional language system is a dynamic system that relies on two underlying sources of information, including user voice registration through the mic. The user (uploader) who will convert speech to text will speak, and the app will detect the voice. Then, the app will segment words to get an accurate voice conversion into text. This will be done with the help of a pre-trained dataset stored in the database. The app will be free for all to get access.

We started from scratch. First, we divided all the work among members as the project is based on computation linguistics, so we started collecting all the previous literature work by the scholars. While reading published papers, we got to know about different design methodologies. Nevertheless, the most important thing is that we collected massive data sets of Sindhi words and sentences. More than 40 papers were read, and finally, we got the whole idea cleared about project designing, development and implementation. The proposed architecture is elaborated in Fig 1.

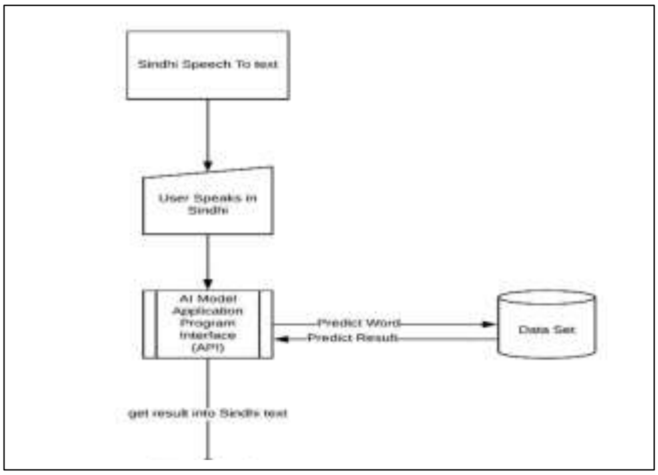


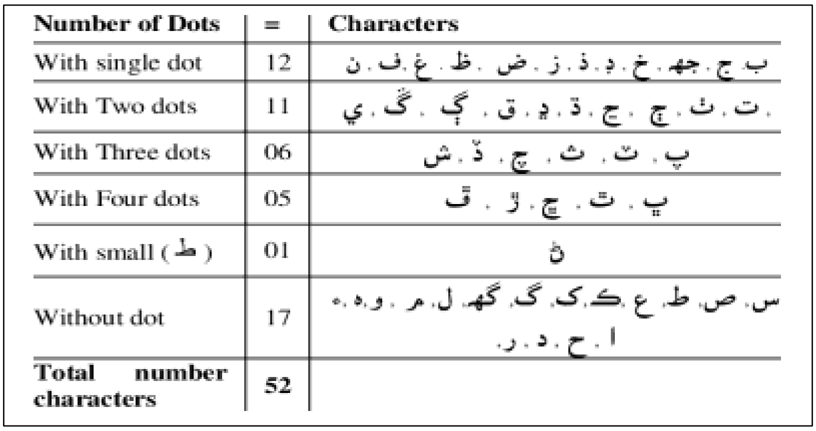
Fig 1. Proposed Architecture of API

4.1 Dataset

The Corpus of any language may be an essential and crucial component of a speech on which research is done since practical investigations of language can be done using that language's corpus. A significant collection of written text assembled for linguistic examination is known as a corpus. Because of this, it is essential information that gives lexicographers, grammarians, and other interested parties a complete grasp of a language. Data on morphology, syntactic parsing, lexicon organisation, semantic, pragmatic, and other linguistic elements can be found through corpus research. The language corpus can be spoken or written, open or closed. The open Corpus is unsecured and current, whereas the close Corpus is specific and constrained. The Corpus, on the other hand, can be used for a variety of tasks, including information retrieval, machine translation, pattern recognition, speech recognition, text-to-speech and speech-to-text recognition and synthesis, feature extraction and analysis, vectorization, word-to-vector analysis, dictionary development, thesaurus and Word Net development, word tokenization, text tagging, text parsing, morphological analysis, machine learning process, classification, and cluster analysis. Our idea is built on word-by-word conversion, with the APP focusing on Sindhi words rather than sentences.

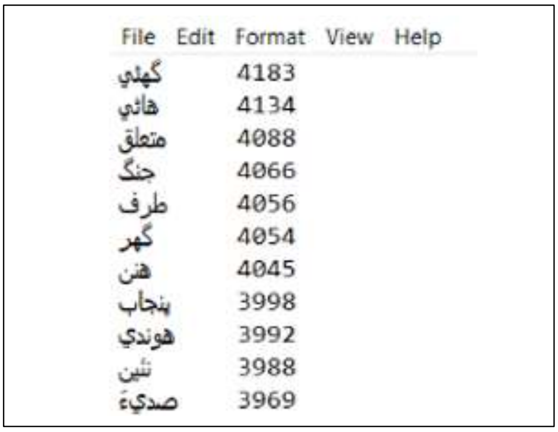
4.2 Collection of the dataset

Sindhi letters are further divided for easy recognition and segmentation, as shown in Fig.2. Whole letters are categorized with several digits, dots and symbols. In this study, speech to text conversion app, the small dataset consists of sentences to train the machine to recognize the voice and detect the right and accurate word for easy conversion. . Where the data is stored in various files and each file is represented in tabular format as shown in Fig 3. The audio data is stored in a folder, as shown in Fig 4, where each folder contains files with different accents Fig 5.



Number of Dots	=	Characters
With single dot	12	ب . ج . جھ . خ . ڊ . ذ . ز . ض . ظ . غ . ف . ن
With Two dots	11	. ت . ث . ڃ . ڄ . ڌ . ڍ . ق . ڳ . ڱ . ي
With Three dots	06	پ . ٽ . ث . چ . ڎ . ش
With Four dots	05	ڀ . ٿ . ڇ . ڙ . ڦ
With small (ط)	01	ڻ
Without dot	17	س. ص. ط. ع. ڪ. ک. گ. گھ. ل. م . و. ه. ھ ا . ح . د . ر.
Total number characters	52	

Fig 2. Sindhi Language Characters Summarization

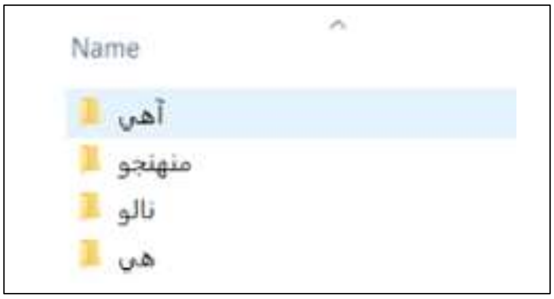


گهلي	4183
هاٿي	4134
متعلق	4088
جنگ	4066
طرف	4056
گهر	4054
هنن	4045
پنجاب	3998
هوندي	3992
ٿين	3988
صديءَ	3969

Fig 3. Text Dataset representation

5. Machine Learning in Speech to Text by Native language

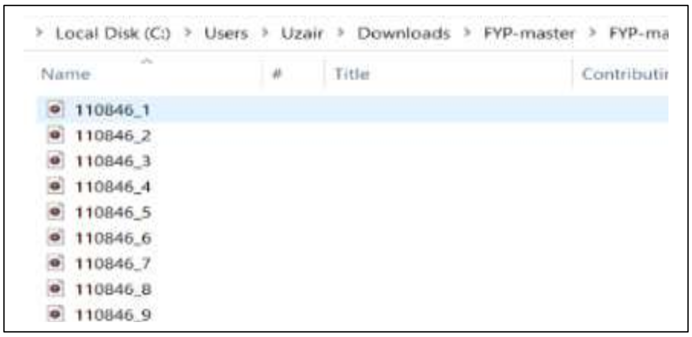
It would be more accurate to say that machine learning groups use speech recognition and voice synthesis to leverage the power of input recognition for all benefits. Speech is powerful, and it adds a human dimension to various electronic devices. People nowadays use cloud-based computers that can be controlled by voice and provide conversational responses to a wide range of queries. Speech recognition training enables AI models to recognise distinct input in recorded audio data. In many cases, machine learning still has a long way to go before reaching perfection this application will be programmed to cover all nuances present in human speech, such as speech length, voice pattern, tone frequency, and so on. However, in order to properly train our speech recognition system, quality information is gathered for processing the available input. The created system is extremely beneficial to people with disabilities. Assume a person has lost the use of his hands or is blind. To make natural voice recognition work in that case, they can use automatic speech recognition or advanced voice recognition.



Name

- آهي
- منهجو
- نالو
- هي

Fig 4. Voice dataset



› Local Disk (C:) › Users › Uzair › Downloads › FYP-master › FYP-ma

Name	#	Title	Contributi
110846_1			
110846_2			
110846_3			
110846_4			
110846_5			
110846_6			
110846_7			
110846_8			
110846_9			

Fig 5. Each variable of the Sindhi audio dataset contains multiple file representation

6. Experiments and Results

The project's goal is to develop a model that can correctly identify a human-spoken word. To obtain a final model, we trained neural networks on a set of data that was tailored to the target data. When you don't have access to a large sample of target data, this method comes in handy. The audio signal is encoded in wav format, with a standard duration of one second. Each entry had only one word. We use Sindhi words that are easy to understand.

6.1 Speech-to-Text API recognition

A Speech-to-Text API synchronous recognition request is the most basic approach for recognising speech audio data. Spoken-to-Text can handle up to one minute of synchronous speech audio data.But our project is a design that will be working on words, so it will hardly take a few seconds to convert speech into text. After Speech-to-Text processes and recognizes all of the audio, it returns a response. The proposed API architecture for the project is presented in Fig 5. Where the NLP API is a machine learning model for analyzing text. NLP APIs can analyze syntax, extract entities, and evaluate the syntax of the text. We developed our own API and integrated the trained model to get the accurate output when the user gives input.

6.2 Speech to Text deployment

Speech to Text deployment is done using python's framework, which will integrate python's script of NLP programming to the web application's front end. The project front end is designed using HTML, CSS and JavaScript, where user input is brought to python's environment with the help of the Angular framework, which will integrate the project code into its interface.

6.3 UI/UX Web technologies

The user interface is designed using HTML, CSS and JavaScript where a button is placed to start the process through the mike icon user will give input and get the desired text as per the input in the text field design on the front screen user interface.

6.3 Python & Angular

Python programming language is the base of our web application, where the initial and the functional programming script is written on. The custom data set and the algorithms are integrated for the pattern matching to generate desired output for the given input from the user. The audio data/ input given by the user is shown in Fig 6. The developed web application provides web access to the speech-to-text conversion process, shown in Figures 7 and 8, respectively. So, the Angular framework is used to integrate the voice script from python to the web interface, which will provide user input. In Fig 9, the learning algorithm Convolutional Neural Network results have been shown. Where the activation function RELU is used, and the epoch is set to 20

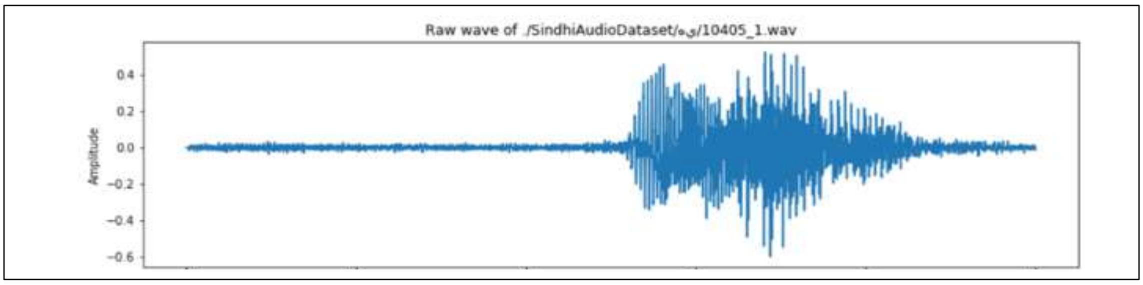


Fig 6. User Voice Input

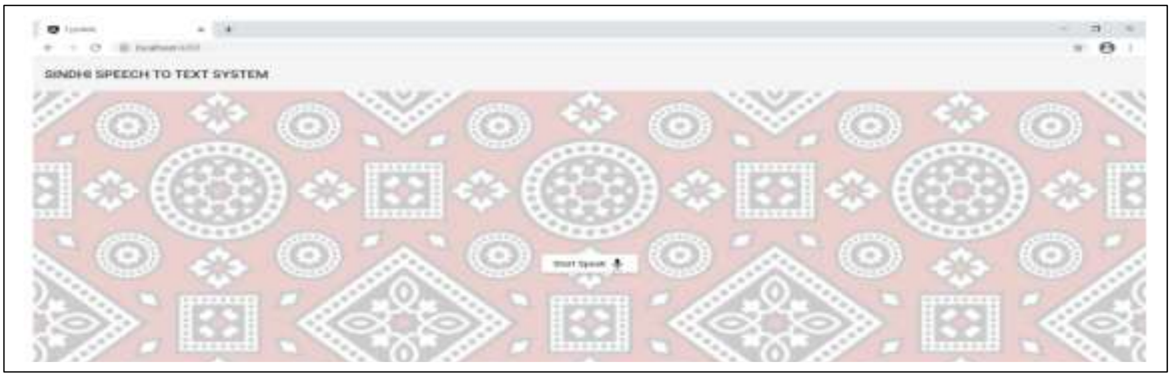


Fig 7. Developed Sindhi Speech to text System Input

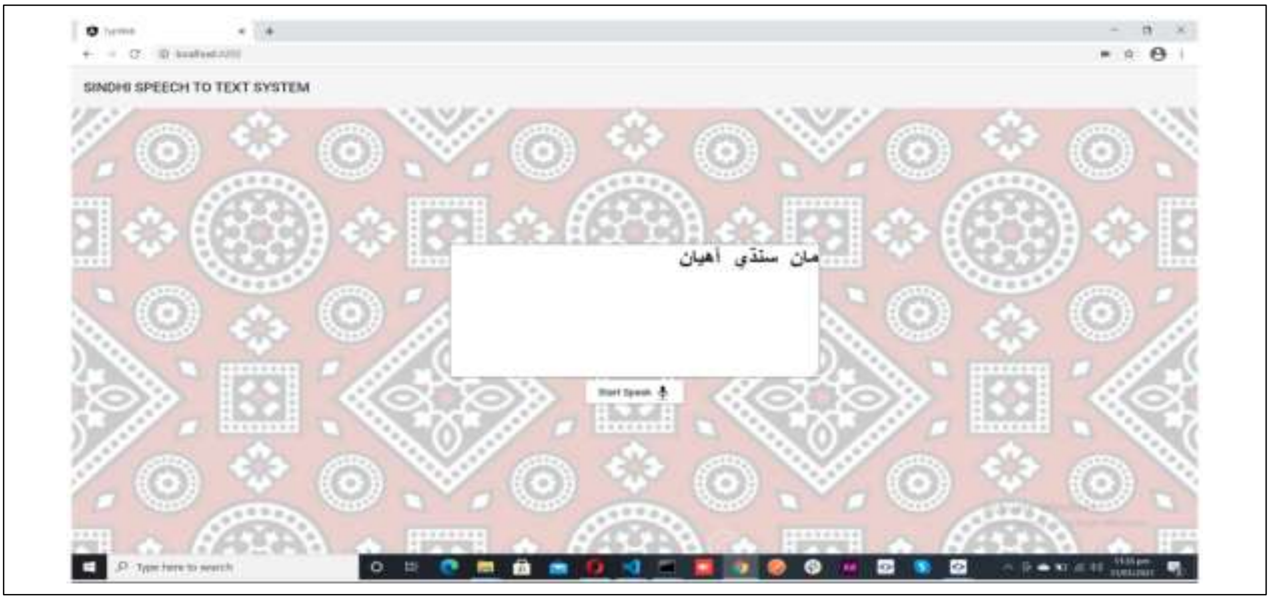


Fig.8 Output Application Programming Interface

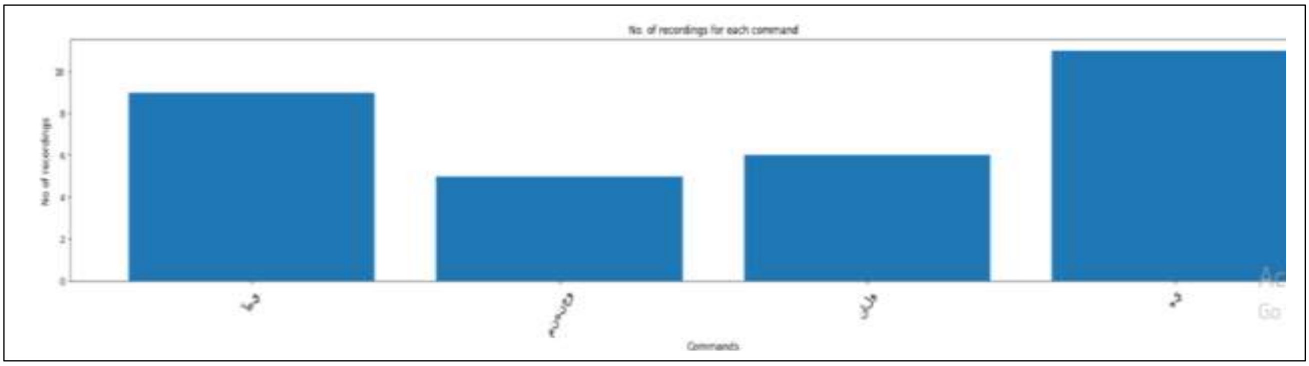


Fig 9. Sentence is fragmented into words

7. Results Discussion

7.1 Convolutional Neural Network

Utilizing a Convolutional Neural Network (CNN), the suggested technique. Neurons in CNN have biases and weights that can be learned with multiple layers. One or more convolutional layers may be present in a CNN, which is a feed-forward neural network. Like in a straightforward multiple layer neural network, CNN is followed by one or more fully linked layers. The back-propagation approach is being used in the neural network setup that has been suggested to train a CNN. Traditional CNN is made up of the layers seen in Fig. 10. While Fig. 11 describes the layering architecture of a configured CNN in detail.

The whole speech to text is based on the neural network, basically purpose of using this method to get the high accuracy rate of detection of speech. As it supports small amount of dataset so that`s why we have used a technique in which user will give input in words. It does not support continuous speech. Preprocessed voice dataset will be identified by the model as then it will recognize the voice data through signal processing and voice attributes which gives final output in text.

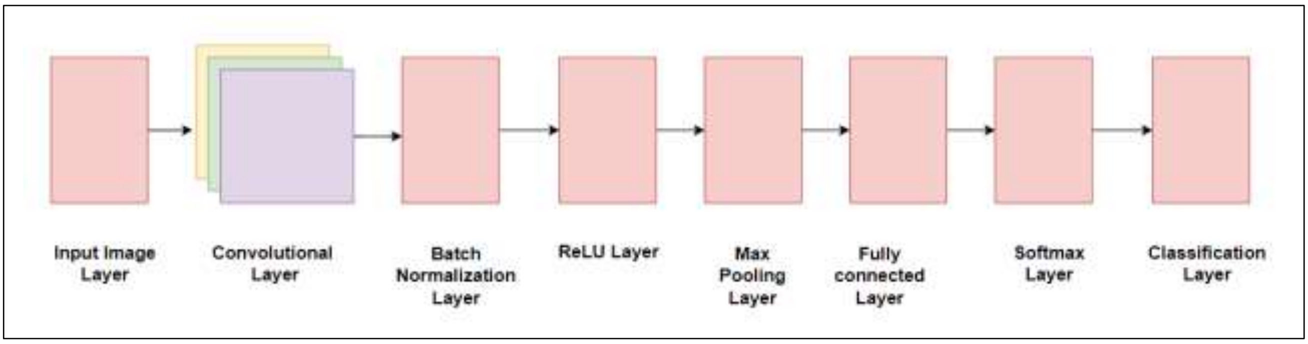


Fig 10. Layers of convolutional neural network

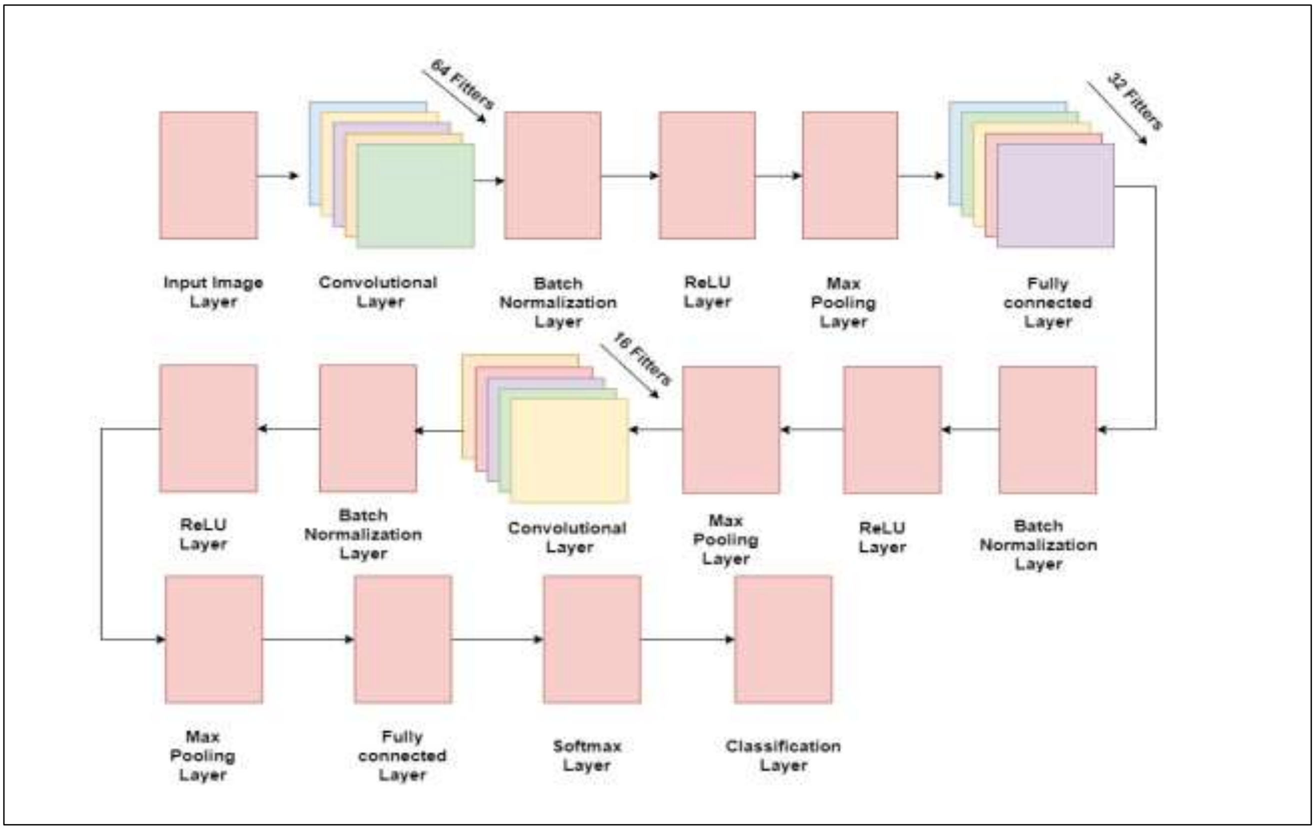


Fig 11. Configured layers of convolutional neural network

7.2 Convolutional Neural Network Training

The training of the neural network is then required by the neural network configuration. To accomplish this step, the overall dataset set is divided into two sections: training data and validation data. The training sample data consists of 75% of the data, and the remaining 25% is used for the test. The neural network is trained with an initial learning rate of 0.005 and a 0.3 dropout rate. The final results are represented in Fig.12. It can be clearly seen that with the increasing epoch that training and testing loss decreases.

8. Conclusion

Speech to text project targets only specific Sindhi language and civilization. Native people are good at the Sindhi language, but if we talk about other languages, most people do not know. Thus, the applications help you to use your language and get the result of that speech in written form. The main objective of the work presented throughout this report is to develop a web-based application for the users to access the application's services to fulfil their desired results according to their needs. The speech recognition process is performed by a software component known as the speech recognition engine. The primary function of the speech recognition engine is to process spoken input and translate it into text that an application understands.

This application will be a nationwide educational & civilizational revolution as the Sindh language will nurture in the technology field. Speech to text using native language is a project based on a web application using core python language. We have been using flask and angular framework along with API. The tool will be working on the spiral model technique. The complete project will be a user-friendly open platform for all native users. User voice will be detected, segmented into words, and then converted into text. This is an advancement in computational linguistics. Tools have been developed in the NLP field, but the Sindhi language and its development are still not as advanced as they should be. We are giving just an idea of speech to text using the Sindhi language.

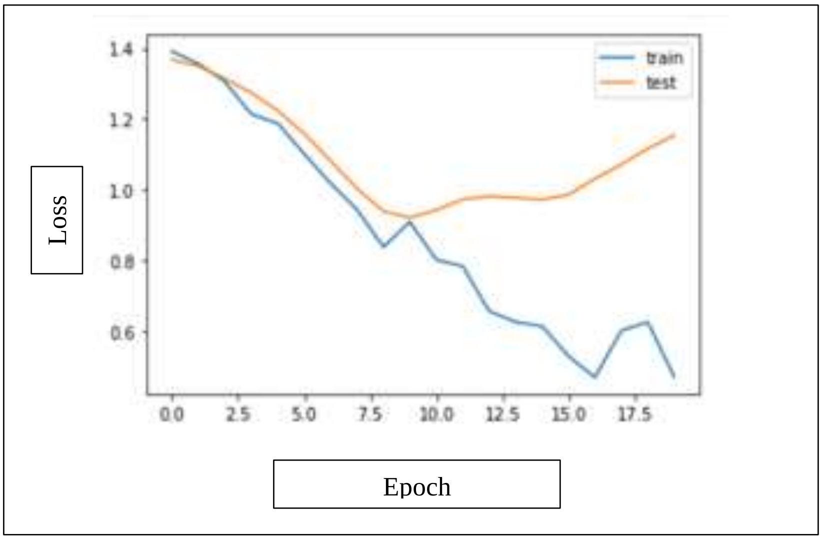


Fig 12. Convolutional Neural Network training and testing results

References

[1]. Humayun, Mohammad Ali, Hayati Yassin, and Pg Emeroylariffion Abas. "Native language identification for Indian-speakers by an ensemble of phoneme-specific, and text-independent convolutions." Speech Communication 139 (2022): 92-101.

[2]. Ma, Pingchuan, Stavros Petridis, and Maja Pantic. "Visual speech recognition for multiple languages in the wild." Nature Machine Intelligence 4, no. 11 (2022): 930-939.

[3]. Takenouchi, Tomoko. "The Effects of Pronunciation Instruction Using Speech Recognition Software for Adult Learners of English."

[4]. Bhaskar, Shabina, and T. M. Thasleema. "LSTM model for visual speech recognition through facial expressions." Multimedia Tools and Applications 82, no. 4 (2023): 5455-5472.

[5]. Abbasi, Abdul Malik, Mansoor Ahmed Channa, Imtiaz Husain, and Ms Ahlam Khan. "Temporal Patterns of Voice Onset Time of English-Sindhi Stops." (2022).

[6]. Tatipang, Devilito P. "William Shakespeare and Modern English: To What Extent the Influence of Him in Modern English." Journal of English Language Teaching, Literature and Culture 1, no. 1 (2022): 61-71.

[7]. Lee, Sangmin-Michelle. "A systematic review of context-aware technology use in foreign language learning." Computer assisted language learning 35, no. 3 (2022): 294-318.

[8]. Kalyanathaya, Krishna Prakash, D. Akila, and P. Rajesh. "Advances in natural language processing–a survey of current research trends, development tools and industry applications." International Journal of Recent Technology and Engineering 7, no. 5C (2019): 199-202.

[9]. Gatt, Albert, and Emiel Krahmer. "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation." Journal of Artificial Intelligence Research 61 (2018): 65-170.

[10]. Mahmood, Shahid, Ghaffar Ali, Rashid Menhas, and Muazzam Sabir. "Belt and road initiative as a catalyst of infrastructure development: Assessment of resident's perception and attitude towards China-Pakistan Economic Corridor." PloS one 17, no. 7 (2022): e0271243.

[11]. HWAR, MHPA. "LA GUAGES OF SI D BETWEE RISE OF AMRI AD FALL OF MA SURA ie 5000 YEARS AGO TO 1025 AD."

[12]. Ghori, Ammar Farid, Aisha Waheed, Maria Waqas, Aqsa Mehmood, and Syed Abbas Ali. "Acoustic modelling using deep learning for Quran recitation assistance." International Journal of Speech Technology 26, no. 1 (2023): 113-121.

[13]. Arora, Siddhant, Siddharth Dalmia, Pavel Denisov, Xuankai Chang, Yushi Ueda, Yifan Peng, Yuekai Zhang et al. "Espnet-slu: Advancing spoken language understanding through espnet." In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7167-7171. IEEE, 2022.

[14]. Gupta, Akshat. "On building spoken language understanding systems for low resourced languages." arXiv preprint arXiv:2205.12818 (2022).

[15]. Halabi, Nawar, and Mike Wald. "Phonetic inventory for an Arabic speech corpus." (2016): 734-738.

[16]. Alsharhan, Eiman, and Allan Ramsay. "Improved Arabic speech recognition system through the automatic generation of fine-grained phonetic transcriptions." Information Processing & Management 56, no. 2 (2019): 343-353.

[17]. Sawa, Yuya, Ryoichi Takashima, and Tetsuya Takiguchi. "Adaptation of a Pronunciation Dictionary for Dysarthric Speech Recognition." In 2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech), pp. 631-635. IEEE, 2022.

[18]. Al-Anzi, Fawaz, and Dia AbuZeina. "Literature survey of Arabic speech recognition." In 2018 International conference on computing sciences and engineering (ICCSE), pp. 1-6. IEEE, 2018.

[19]. Poncelet, Jakob, and Vincent Renkens. "Low resource end-to-end spoken language understanding with capsule networks." Computer Speech & Language 66 (2021): 101142.

[20]. Dootio, Mazhar Ali, and Asim Imdad Wagan. "Unicode-8 based linguistics data set of annotated Sindhi text." Data in brief 19 (2018): 1504-1514.

[21]. Ursani, AHSAN AHMAD, BHAWANI SHANKAR Chowdhry, and M. A. Unar. "A Speech To Text System for Sindhi Language." Mehran University Research Journal of Engineering and Technology 20, no. 3 (2001): 139-146.