International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 11 Issue 12, December-2022, Impact Factor: 7.957

Advanced Audio Signal Processing and Foul Language Detection

Dr. Nisha Auti[1], Anagha Desai[2], Atharva Pujari[3], Shreya Patil[4], Sanika Kshirsagar[5], Rutika Rindhe[6]

[1,2,3,4,5,6] Department of Computer Engineering, Jspm Narhe Technical Campus,Maharashtra, India

ABSTRACT

Speaker recognition is the task of recognizing the identity of someone based on the speaker's speech signal. Tonal Speech Recognition refers to identifying the emotions or sentiments in the voice of the speaker. We have developed a system that makes use of certain keywords in analyzing the speech of the speaker. The advanced system will dissect the audio signal to classify the given speech as happy, sad, aggressive, etc. The system implements advanced audio signal processing and foul language detection. The system with the help of an automatic speech recognition system identifies the voice of the speaker and analyzes the audio signal to detect any use of foul language and aggressive speech. This system is developed to analyze the Indian Regional Language Marathi. We have developed a system to recognize the tonal speech of Marathi language and the foul words spoken and the aggressive speech in the language.

Keywords: Foul Language Detection, Sentiment Analysis, Speaker Recognition, Tonal Speech Recognition

INTRODUCTION

Speech Recognition is the technology that allows mortal beings to use their voices to speak with a computer interface in a way that, in its most sophisticated variations, resembles normal mortal discussion. It is a technology that allows users of information systems to speak entries rather than punching figures on a keypad. Speaker recognition is an important bio-feature recognition method. Speaker recognition is the task of recognizing the identity of an individual based on his/her speech signal. Speaker recognition is a highly important biometric identification technology and this particular technology has been applied in various fields such as secure access to extremely secure areas, also used in machines like voice dialing, banking, database, and computers. The study of speaker recognition can be considered as the use of employing statistical methods to identify the individual based on their unique voice signal properties, which are encoded in a sequence of successive samples in time. Speaker identification is the process of identifying a speaker from the speech information using the characteristics of the sound. The method developed by Bell and Newson in 1975 is used for speaker identification in audio processing. The proposed model uses acoustic data from the speaker's voice and generates a model of their vocal tract. The model is then used to identify the speaker from the collected inputs of audio files. The first patent for speaker recognition was filed back in 1983 by Michele Cavazza and Alberto Ciaramella from the telecommunication research in CSELT to improve the noise reduction techniques across the communication network.

Tonal speech recognition is a process that analyzes audio files to identify the various tones in each word or phrase. For example, if someone said "I'm sorry" three times at the end of every sentence in a recording of their speech, they may be apologizing for something—but how do we know? Tonal speech recognition can help us determine what these apologies are for. Tonal speech recognition is the process of understanding how speakers use intonation to convey meaning. It is used by companies like Amazon Prime to improve its voice assistant Alexa. Sentiment analysis is a process that analyzes audio signals to determine the overall tone of the speech. Sentiment analysis is the process of determining whether an audio file or the speech input by an individual contains positive or negative emotions. Sentiment analysis compares the extracted keywords from the given speech input of an individual to a set of rules that indicate how to recognize the emotions or tones in the speech used. Sentiment analysis has been around for a long time.

Sentiment analysis and tonal speech recognition these two systems are often used together, since they both rely on trained word embeddings to determine how a word is meant. Sentiment analysis analyzes the sentiment of a speech signal by looking at the keywords that appear in it and comparing them to other words that are associated with positive or negative emotions. To do this more effectively, you need to take into account how people use those

words in their speech. For example, if an individual uses aggressive words in their speech. You need to figure out what they mean by that and then use that information to improve your sentiment analysis results. Tonal speech recognition uses similar techniques but instead focuses on tone rather than individual words. This means that instead of trying to identify every word as having either positive or negative connotations so that it can parse out whether something is positive or negative, it tries to identify the overall meaning of the speech. It also takes into account how many times someone says something with a particular tone over time—for example, if they say something three times and then don't say anything else all day, it might classify that statement as being very negative when really it was just an off day for them.

Foul language detection is the process of identifying and classifying profanity in the speech of an individual. Foul language detection is another type of computer-based speech analytics that detects when people use certain words or phrases outside of their normal vocabulary. For example, if an individual uses certain abusive words in their speech then it could be considered as an instance of foul language. Speech recognition systems can also be used to identify foul language from text-based communications such as emails, memos and social media posts. The most common way to do this is by using a deep neural network (DNN) or a convolutional neural network (CNN). A DNN works by looking at the sounds in the sentence and trying to find patterns in them. It then applies those patterns to words that are similar in sound, but not necessarily spelled the same way. The result is a more accurate recognition of foul language. However, there are still some problems that arise with this approach. For example, it would not work well if you have long pauses between words or if your accent changes during a sentence. The ability to recognize foul language is an important skill for many people, and it's getting harder to do it with real-time speech recognition.

We have developed a machine learning model for advanced signal processing and foul language detection. The model is designed to analyze one of the Indian Regional Languages, Marathi. There are instances when you need to detect foul language in Marathi. The main reason for this is that some people use foul language in their exchanges when they want to make their point or express themselves, but they do not want their discussion to be eavesdropped by others. Also, numerous people who are not native speakers of Marathi use words like "bloody" or "hell" during normal conversations. In other cases, these words may be used as compliments or expressions of aggression. The first step of the system will be to dissect the audio speech signal to recognize the speaker, after which the model will analyze the speech to perform sentiment analysis and tonal speech recognition, further proceeding to detect the foul or vituperative words spoken by an individual during his/her speech. This analysis will be carried out for marathi language. The deep neural network and convolutional neural network have been the most effective at detecting foul language in Marathi. The deep neural network has been able to detect a variety of foul words, including swear words. The convolutional neural network has been able to detect all types of foul language.

BACKGROUND

Speaker recognition is the process of identifying a person by their voice, in this case, a speaker.In audio signal processing, we can use various techniques and algorithms to extract speech from an audio signal. Sentiment analysis of audio signals is a method for determining whether an audio message has positive or negative sentiment. This can be used to improve speech recognition systems by providing a more accurate transcription of the original audio messages.

Speaker recognition algorithm is one of several techniques used in our work which includes feature extraction methods like Mel-frequency cepstrum coefficients (MFCC) and Mel-frequency cepstrum transform (MFCT).Another method for recognizing speakers is by using neural networks. This technique has been shown to outperform acoustic modeling for most applications because it allows for more complex models than acoustic models without sacrificing recognition accuracy.

Speaker recognition methods have two important modules which are Enrolment and Verification. At the time of enrolment, the subject's audio signals are captured to extract important features from them and finally, a model is created which is also termed as voice print. For the next phase that is verification , a voice sample or commonly termed as utterance is balanced against the previous model. The important difference between identification methods and verification methods is that , for verification only one voice print is considered to be balanced against a single voice pattern whereas identification methods tend to balance the utterance against multiple models.

In order to perform speaker recognition and sentiment analysis over audio signals, there are several technologies that are used such as speech recognition systems, text-to-speech synthesis engines and acoustic models.These technologies can be used for different purposes such as voice command systems where it speaks commands or reminders to users or for entertainment purposes where it reads books or plays music.

Speaker recognition is a very powerful tool for identifying speakers in any given environment, but it does not just provide accurate results; it also has an impact on the outcome of an application's performance. That's because speaker recognition depends on many factors: environment (temperature, lighting, background noise), acoustics

(room size, acoustic properties), and even age (acoustics change as we get older). All of these variables can have an impact on how well your system performs when it comes time to identify speakers.

Limitations of Existing System-

The major limitation is that the system can only differentiate between the voices as male or female, but it cannot distinguish the identity of the person who the voice belongs to. The tonal frequencies of marathi Language are studied at a very small scale which constitutes a major limitation that we are trying to overcome. Another Limitation is that the system cannot implement recognition as well as sentiment analysis of the speech at the same time and there is no mechanism of deployment for the existing models.

Features of Neural Network based Speaker Recognition models-

Improved security

Speaker recognition can be used as a form of authentication to grant or deny access to secure systems or locations. This can help prevent unauthorized access and increase the overall security of a system.

Increased convenience

Speaker recognition can be used to automate tasks such as logging into accounts or making phone calls, which can save time and make everyday tasks more convenient for users.

Better accuracy

Recognition technology can be more accurate than other forms of authentication, such as passwords or PIN numbers.

Enhanced customer experience

Speaker recognition can be used in customer service environments to improve the user experience by allowing customers to quickly and easily identify themselves and access their account information.

Cost savings

In some cases, speaker recognition technology can help organizations reduce their operational costs by automating tasks and reducing the need for manual labor.

SYSTEM OVERVIEW

Speaker recognition, also known as voice or speech recognition, is the process of automatically identifying an individual by their unique voice characteristics. This technology is commonly used in voice-controlled virtual assistants, such as Amazon's Alexa or Apple's Siri, and in security systems to authenticate individuals by their voice.

The system architecture of speaker recognition typically involves several components, including a microphone or other audio input device, a feature extraction module, a machine learning algorithm, and a database of speaker models. Furthermore, we are adopting Sentiment Analysis, a subfield of natural language processing that focuses on identifying and extracting the sentiment or emotion behind a given piece of text.

The audio input is first captured by the microphone and converted into digital form by undergoing multiple processes such as Sampling, Quantization and Framing. The resulting digital signal is then processed by the feature extraction module, which extracts relevant characteristics from the signal that can be used to identify the speaker. These characteristics may include the pitch, the spectral content, and other features of the speaker's voice.

The extracted features are then fed into a machine learning algorithm, which compares them to a database of speaker models to identify the speaker. The machine learning algorithm may use a variety of techniques, such as clustering or classification, to determine the most likely speaker.

Once the speaker has been identified, the system further proceeds for Sentiment Analysis on that piece of tone. This will classify the speech into Neutral, Aggressive, Negative or Abusive context to generate a Graded Sentiment Analysis

report.

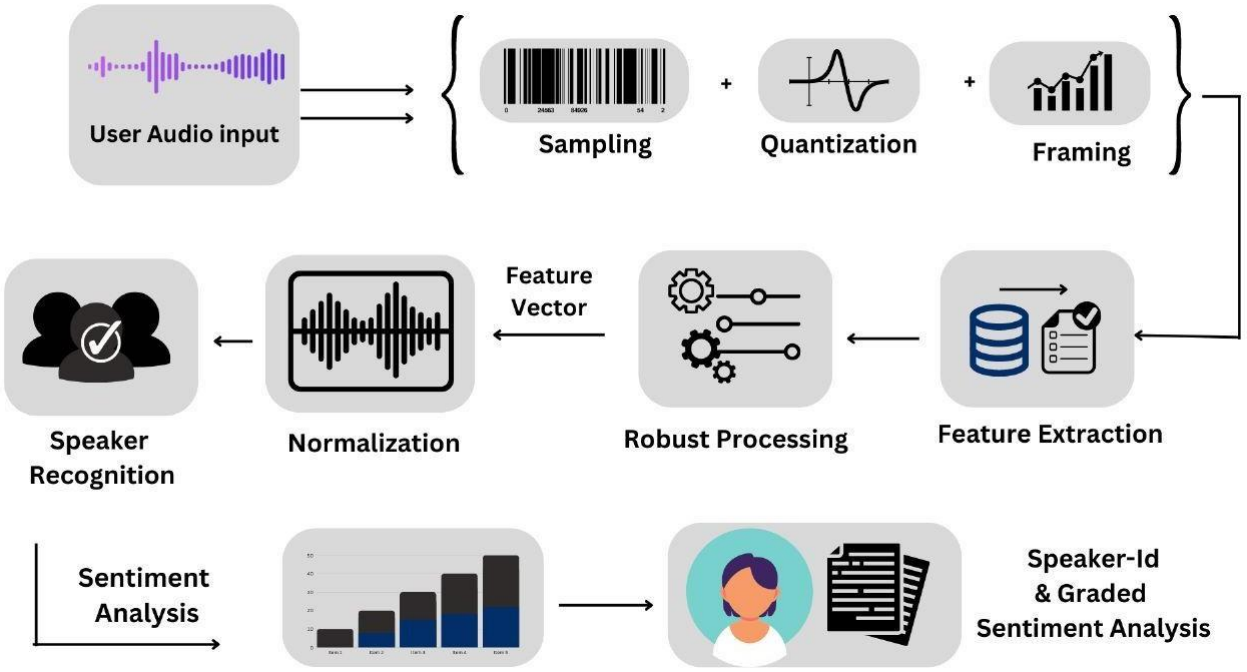

Figure 1: System Architecture

Overall, the system architecture of speaker recognition involves the combination of audio input and signal processing, machine learning, and a database of speaker models to automatically identify individuals by their unique voice characteristics.

TECHNOLOGY STACK

Librosa
Librosa is a Python library for music and audio analysis that could potentially be used in speaker recognition applications. Speaker recognition involves identifying an individual based on their unique voice characteristics, and Librosa provides a range of tools for extracting and analyzing these characteristics from audio data. In addition to spectral analysis, Librosa also provides a number of other features that could be useful in speaker recognition. These include tools for segmenting audio data into distinct segments, such as individual words or phrases, as well as functions for extracting various features from the audio data, such as rhythm and tempo.

Numpy
NumPy is a Python library for scientific computing. It provides a high-performance multidimensional array object, as well as a large collection of functions for working with these arrays. NumPy also provides a large collection of functions for working with arrays, including mathematical functions, linear algebra functions, and random number generation. These functions are implemented in a way that is highly efficient, making it possible to perform complex numerical computations on large datasets quickly and easily.

OS
The os module in Python is a built-in library that provides functions for interacting with the operating system. This module provides a portable way of using operating system dependent functionality, such as reading or writing to the file system, starting a new process, or working with directories and paths.

PyTorch
PyTorch is an open-source machine learning library for Python. PyTorch provides a range of tools and libraries for building and training machine learning models, including support for deep learning and neural networks. It also includes a number of useful features, such as automatic differentiation, which makes it easy to implement and train complex machine learning models.

Multiprocessing
Multiprocessing could potentially be used to speed up the processing of audio data and improve the performance of the speaker recognition system.One way that multiprocessing could be used in speaker recognition is by dividing the

audio data into multiple smaller segments and processing each segment in parallel using separate processes. This could potentially reduce the overall processing time, as each process can work on a different segment of the data simultaneously. Another potential use of multiprocessing in speaker recognition is to distribute the workload across multiple computers or processing units. This could be useful in systems that require a high degree of computational power, such as systems that use machine learning algorithms to identify speakers.

Glob

The glob module in Python provides a convenient way to match file and directory names using wildcards. It is part of the standard Python library and can be used to quickly and easily find files and directories that match a given pattern.

Neural Net

Neural nets, or artificial neural networks, can be used in speaker recognition systems to identify individuals based on their unique voice characteristics. In a speaker recognition system, the neural net would be trained on a large dataset of audio samples from different speakers, learning to identify the unique features of each speaker's voice. The use of a neural net in speaker recognition allows the system to learn and adapt to the unique characteristics of each speaker's voice, improving its performance over time. Neural nets are also able to handle complex and noisy data, making them well-suited to the task of speaker recognition.

METHODOLOGY

Enrollment Stage

The enrollment stage for speaker recognition is the first step in the process. At this stage, speakers must be enrolled in the system as well as their name and speech. The speaker's name is entered into the system using a text box that contains the speaker's name and information about them, such as their gender, age, occupation and other demographics. The speech is entered into the system using the speech recognition feature of the platform. Once both are entered, they are analyzed by a machine learning algorithm in order to recognize patterns that indicate that certain words or phrases are likely coming from a human being rather than a computer program or another type of non-human entity.

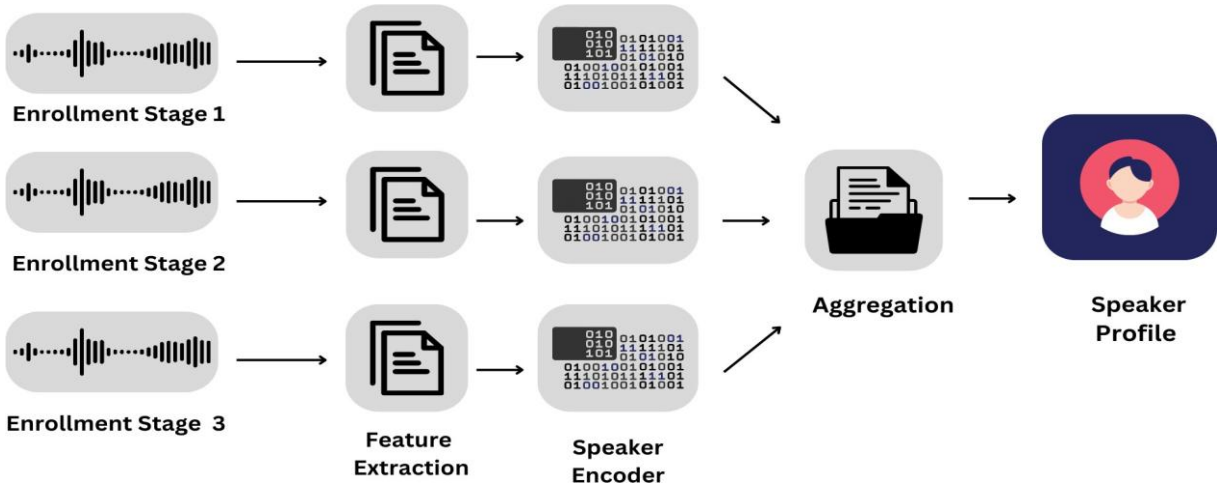

Figure 2: Speaker Enrollment Stage

Feature Extraction

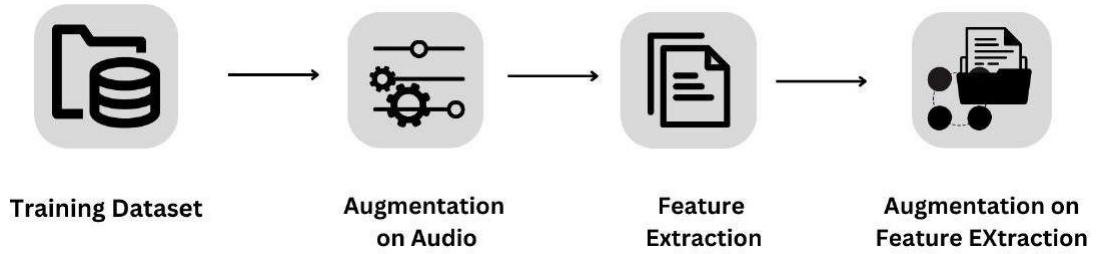

Figure 3: Feature Extraction Phase

Feature extraction is a process that breaks down information into smaller pieces in order to maximize the amount of data that can be used. This is done by identifying what characteristics or properties make up a larger piece of information. Audio signals often contain information that can be used to extract features. Some of these features are as simple as the presence or absence of a sound, but others might include more complex information about the sound, like its length or pitch. Feature extraction is a process by which a digital audio signal is analyzed to identify and extract useful information from it. It helps in removing unwanted voice prints or noise and maintains the time-frequency boundaries by conversion of digital or analog signals. In our model, unwanted noise, abusive words, sounds not from Articulatory organs such as Clapping, Whistling, sounds that are not socially meaningful such as coughing, sneezing are extracted and the data is passed for further frame stacking.

Recognition

In this phase, the identification of speaker / subject will take place. From the given set of samples , initially , raw audio files will be analyzed to gain some insight about the audio frequencies of each speaker. Later, important features from those audio files or voice prints will be observed and considered for further processing. A pattern matching algorithm extracts the pattern of voice prints and It involves comparing the periodic nature of the signal with the given pattern, and finding the best match. Pattern matching is often used as a way to assess what someone else has heard or to recognize sounds that are similar to ones you've encountered before. After this stage, all the subjects can be uniquely identified / recognized and the same will be reflected in the output.

There will also be a normalization stage.Normalization is the process of setting the volume of a signal to its maximum level.The normalization process is used to ensure that the audio data in a file has been properly quantized, and that it can be played back without distortion. This is important because it ensures that there will be no clipping or overloading of the audio device, which can cause distortion or other issues with playback.

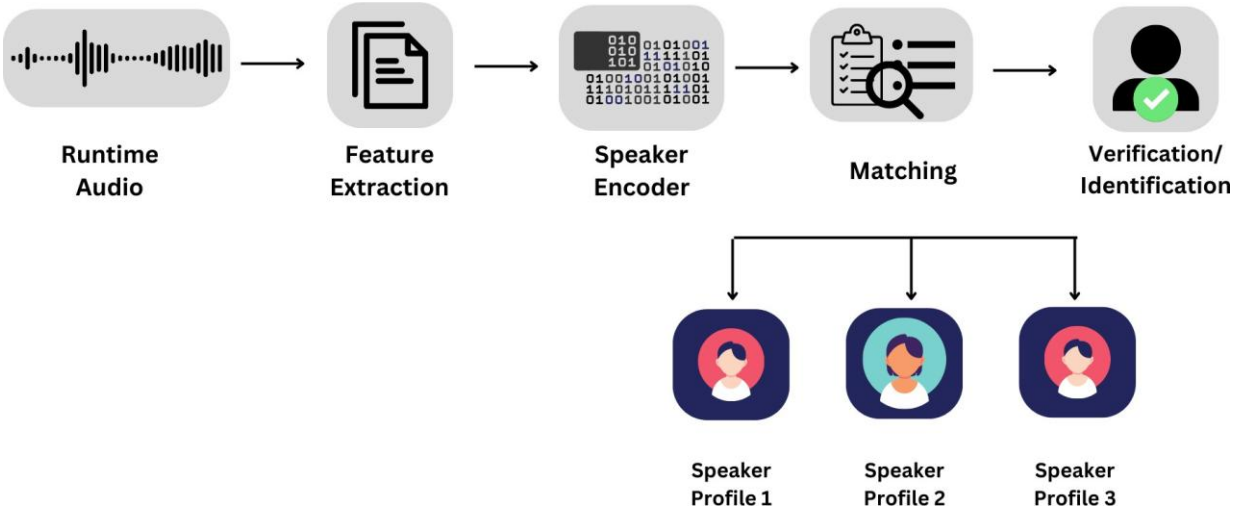

Figure 4: Speaker Recognition Stage

Based on the data collected the sentiment analysis check is performed to identify the emotional tone behind a body of Input file. Our system will define whether it is a positive, negative, aggressive or neutral behavior. This helps us better analyze the user response. The output will be in the form of speaker id which will be of different users, speech sentiment analysis displaying the emotion of the speaker.

CONCLUSION

We proposed a deep neural network based speech recognition model for tonal speech recognition and foul language detection. The developed system recognizes the individual based on his/her speech signals and after the recognition, the system analyzes the speech of the speaker and searches for keywords that may help to determine the tones of the speaker i.e. it helps analyze the joy, sorrow, aggression, etc emotions in the speech of the speaker. After speaker recognition and tonal speech recognition, the system analyzes the speech signal to search for keywords that help detect the foul speech that the speaker used to communicate with. The proposed system is specifically developed for one of the Indian Regional Languages , Marathi. It has been observed that a lot of work has been done in the tonal speech recognition system of Asian Continental Languages and Indo-European Languages, but very little work has been done for Indian Regional Languages in the tonal speech recognition system. This was developed to recognize the tonal speech of Marathi language and the foul words spoken as well as the aggressive speech in the language. The developed speaker recognition system provides a secure and protected environment for speaker authentication and voice identification. The area of speech

recognition and speaker identification is continually changing and perfecting. In the future, there are vast possibilities to enhance the area of speech recognition technology. Speech recognition can provide a secure terrain to the various services that we utilize in our day-to-day life by making use of voice authentication. In the future, the correctness of speech recognition and the quality of speech will be more advanced which will make communication easy and reliable for everyone.

REFERENCES

[1]. Douglas O'Shaughnessy , " Automatic Speech Recognition ", Plenary in IEEE Chilecon 2015, 978-1-4673-8756-9/15/$31.00 © 2015 IEEE.
[2]. Srujana K , R.Ramesh , Kiran G., Ch. Manikanta, " Artificial Intelligence Speech Recognition System using MATLAB "978-1-5386-3243-7/17/$31.00 ©2017 IEEE.
[3]. Feng Ye 1,2 and Jun Yang 1,* , " A Deep Neural Network Model for Speaker Identification " , Appl. Sci. 2021, 11, 3603.
[4]. Sakshi Dua 1, Sethuraman Sambath Kumar 1, Yasser Albagory 2, Rajakumar Ramalingam 3, Ankur Dumka 4,5, Rajesh Singh6, Mamoon Rashid 7,*, Anita Gehlot 6, Sultan S. Alshamrani 8 and Ahmed Saeed AlGhamdi 2 , " Developing a Speech Recognition System for Recognizing Tonal Speech Signals Using a Convolutional Neural Network " , Appl. Sci. 2022, 12, 6223.
[5]. Habib Ibrahim, Asaf Varol, "A study on Automatic Speech Recognition Systems", IEEE, June 2020
[6]. Ghai W, Singh N (2013) Continuous speech recognition for Punjabi language. Int J Comput Appl 72(14):23–28
[7]. Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C, Casper J, Catanzaro B, Cheng Q, Chen G, Chen J, Chen J, Chen Z, Chrazanowski M, Coates A, Diamos G, Ding K, Du N, Elsen E, Engel J, Fang W, Jiang B, Ju C, Jun B, Legresley P, Lin L, Liu J, Liu Y, Li W, Li X, Ma D, Narang S, Ng A, Ozair S, Peng Y, Prenger R, Qian S, Srinet K, Sriram A, Tang H, Tang L, Wang C, Wang J, Wang K, Wang Yi, Wang Z, Wang Z, Wu S, Wei L, Xiao B, Xie W, Xie Y, Yogatama D, Yuan B, Zhan J, Zhu Z (2016) Deep speech 2: end-to-end speech recognition in English and Mandarin. In: Proceedings of the 33rd international conference on machine learning (ICML), New York, vol 48, pp 173–182
[8]. Besacier L, Le VB, Boitet C, Berment V, 2006 ASR and translation for under-resourced languages. In: Proceedings of the international conference on acoustics, speech and signal processing, Toulouse, France, vol 5, pp 1221–1224
[9]. Dua, M.; Aggarwal, R.K.; Biswas, M. Optimizing Integrated Features for Hindi Automatic Speech Recognition System. J. Intell. Syst. 2019, 29, 959–976.
[10]. Ghai, W.; Singh, N. Continuous Speech Recognition for Punjabi Language. Int. J. Comput. Appl. 2013, 72, 23–28
[11]. Huang, J.T.; Li, J.; Gong, Y. An analysis of Convolutional Neural Network for Speech Recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April2015; pp. 4989–4993.
[12]. Reynolds, D.A. An Overview of Automatic Speaker Recognition Technology. In Proceedings of the 2002 IEEE International Conference Acoust. Speech Signal Process, Orlando, FL, USA, 13–17 May 2002; Volume 4, pp. 4072–4075.
[13]. Ankur, M.; Divya, K.; Agarwal, R.K. Speaker recognition for Hindi speech signal using MFCC-GMM approach. Procedia Comput. Sci. 2018, 125, 880–887.