Grenze International Journal of Engineering and Technology, June Issue

Developing a Speech Recognition System for Educational Settings

Rina Damdoo[1], Shweta Choulwar[2], Amod Nagbhidkar[3], Aditya Yadav[4] and Pratham Vyawahare[5]
[1-5]Shri Ramdeobaba College of Engineering and Management, Nagpur, India
Email: damdoor@rknec.edu, nagbhidkarav@rknec.edu, yadavao@rknec.edu, vyawaharepv@rknec.edu, choulwarsv@rknec.edu

***Abstract*— The field of education is not an exception to the revolutionary developments brought about by the swift integration of Artificial Intelligence and Machine Learning. In this work, we propose a Digital Twin to Assist Teaching Learning (DTATL) model, a ground-breaking method intended to automate and improve the teaching-learning process using intelligent speaker recognition, in response to the rapidly growing trend of digital learning. The main goal of our research project is to use the Support Vector Classification (SVC) model to construct a speaker recognition system. This concept is incorporated into an approachable Flask-based user interface, making it available to both instructors and students. We used the Mel-frequency Cepstral Coefficients (MFCC) for the feature extraction technique and achieved remarkable results by utilizing the enormous potential of machine learning, with speaker identification training, and testing accuracies of 95%, and 84% respectively. This study also offers deep insights into the application and impact of the DTATL in the educational sector by offering a thorough overview of its creation, results, and prospective advantages.**

***Index Terms*— Artificial Intelligence, Machine Learning, Digital Twin, Intelligent Speaker Recognition, Support Vector Classification, Speaker Identification, Mel-Frequency Cepstral Coefficients, Dataset Size, Audio Augmentation.**

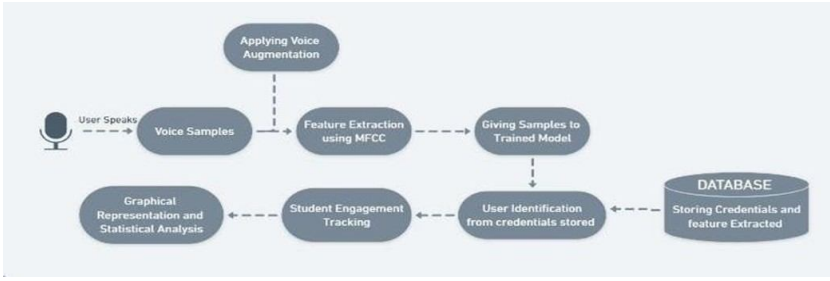

Figure 1: Proposed block diagram for DTATL

I. INTRODUCTION

The main goal of our research project is to use the Support Vector Classification (SVC) model to construct a strong

Grenze ID: 01.GIJET.10.2.205_1

speaker recognition system. This concept is easily incorporated into an approachable Flask-based user interface, making it available to both instructors and students. We have achieved remarkable results by utilizing the enormous potential of machine learning, with testing and training accuracy of 84% and 95%, respectively, in accurately identifying speakers based on their distinctive voice features. Our model's main advantage is its skilful application of the Mel-frequency Cepstral Coefficients (MFCC) feature extraction technique. This technique successfully extracts about 35 unique voice characteristics from speech samples, which greatly adds to the remarkable accuracy that our system has been able to achieve. Our findings highlight how important dataset size is in influencing model performance. We intentionally selected the SVC model because of its performance under such limitations, even though we were working with a rather small dataset. The remarkable training and testing accuracies attained have supported this decision and demonstrated the system's resilience in practical situations. Subsequent initiatives are deliberately focused on improving speech feature extraction and verification procedures. In addition, we want to make our method more applicable in educational contexts so that it may be used to gauge students' understanding and engagement. To further improve accessibility and usability, we also intend to build a Progressive Web Application (PWA) and integrate semantic textual/sentence similarity models into our framework. By providing them with an advanced voice-based identification mechanism, the DTATL model gives instructors more authority. This approach makes it possible to apply teaching strategies that are optimized and makes it easier to evaluate student engagement in real-time. As a result, our study makes a significant contribution to the ever-changing field of AI-driven educational technology, creating a stimulating and dynamic learning environment that is well-positioned for the future. This study offers deep insights into the application and impact of the DTATL model in the educational sector by offering a thorough overview of the model's creation, results, and prospective advantages. With a history of intellectual involvement and richness of knowledge, this research endeavour seeks to provide answers, ideas, and conclusions beyond traditional bounds. This work is a demonstration of our dedication to this research as well as a spark for more conversations and advancements in the field. As we continue our joint quest for knowledge and promote our common intellectual interests, we provide our most recent contributions in the following pages.

II. Literature Review

A. Background Of Speech Recognition

The Bell Labs Audrey's speech recognition system was the first to recognize the ten English digits when speech recognition research got underway in the 1950s. However, it advanced and became a significant topic for research in the late 1960s and early 1970s. Artificial neural networks (ANNs) and the HMM model were effectively applied to speech recognition in the 1980s. 1988 Majoree Kai et al. achieved a speaker-independent continuous voice recognition system-SPHINX, with 997 vocabulary items using the VQ/ Iü IMM approach. This is the first voice recognition system in the world, it is a continuous, high-performance, non-specific system with a broad vocabulary. Individuals eventually overcome the three main challenges of having a wide vocabulary, speaking continuously, and speaking in general terms. Additionally, it recognized the majority of statistical models and techniques used in language processing and speech recognition. Speech recognition technology has already moved from the lab to the real world, with more developed devices available. Many wealthy nations, including the US, Japan, and South Korea, as well as well-known corporations like AT&T, Apple, Microsoft, IBM, and others, heavily invest in the study and development of useful voice recognition systems. Over the years, several models and techniques have been created on the subject of voice recognition. In addition to Hidden Markov Models (HMMs), two more prominent models utilized in voice recognition are Connectionist Temporal Classification (CTC) and Dynamic Time Warping (DTW).

B. Dynamic Time Warping (DTW)

DTW is a pattern-matching and speech recognition method based on dynamic programming. By determining the best alignment between two temporal sequences, it calculates how similar they are. DTW was incorporated into some of the earliest speech recognition systems and is effective for single-word recognition. Unfortunately, because of its processing complexity and lack of modelling for speech variability, it frequently has trouble with continuous speech recognition. The DTW algorithm can be mathematically represented as follows: Given two sequences A={a1,a2,. an} and B={b1,b2,
bm}where n and m are the lengths of sequences A and B respectively, the DTW distance between these sequences is calculated using the following dynamic programming recurrence relation:

D(i,j)=Distance(ai, bj)+min(D(i-1,j), D(i,j-1), D(i-1,j-1)) (1) where D(i,j) represents the DTW distance between the ith element of sequence A and the jth element of sequence B. The Distance D(ai, bj) function calculates the distance between ai and bj, which can vary based on the specific application, such as Euclidean distance or Manhattan distance.

C. Connectionist Temporal Classification (CTC)

For speech recognition tasks, CTC is a neural network- based method that has grown in favour recently. For end-to-end automatic speech recognition (ASR) systems, it is especially helpful. CTC models are appropriate for jobs involving variable-length sequences since they do not require explicit alignments between the input and output sequences. Their performance in both small and large vocabulary ASR tasks has been encouraging.

D. HMM in Speech Recognition

Using HMM in Speech Recognition involve

1) *Modelling Temporal Sequences*: A key component of voice recognition is the modelling of temporal sequences, which is a task best left to HMMs. The dynamic quality of speech, including phonetic changes and transitions, can be captured by them. Probabilistic Framework: To model uncertainty in speech, HMMs offer a probabilistic framework. This is crucial since speech is naturally changeable, and speech patterns and pronunciation variations can be accommodated by HMMs.
2) *State-Based Modelling*: HMMs model speech using a state-based methodology, in which a given state corresponds to a specific phoneme or sound. This makes it possible for HMMs to represent speech's underlying phonetic structure.
3) *Flexibility*: HMMs are adaptable to a range of voice recognition tasks, such as discrete word recognition and continuous speech recognition, because they can handle both discrete and continuous input.
4) *Historical Performance*: HMMs have a successful history in speech recognition. Many years of study and improvement have produced extremely efficient speech recognition systems, like the SPHINX system mentioned in the literature review.
5) *Integration with Other Models*: HMMs can be combined with other models and methods, like language models for more advanced language understanding and Gaussian Mixture Models (GMMs) for acoustic modelling.

In addition to these, there are numerous other voice recognition models, such as

E. Gaussian Mixture Models (GMMs)

These statistical models are employed in voice recognition systems for acoustic modelling. Using a combination of Gaussian distributions, GMMs depict the probability distribution of characteristics taken from speech signals. A unique speech pattern is captured by each Gaussian component. When modelling speech sound variability, GMMs work very well, enabling accurate recognition even in the presence of various accents, speaking velocities, and ambient factors.

F. Convolutional Neural Networks (CNNs)

These are a type of deep learning model that is particularly useful for deciphering intricate patterns found in speech signals. CNNs process spectrogram pictures obtained from audio sources in speech recognition. The frequency content of speech transmissions throughout time is represented by spectrograms. Convolutional layers in CNNs enable precise phoneme detection, speaker identification, and speech-to-text conversion by capturing local characteristics and hierarchical patterns in spectrograms.

The convolution operation in CNNs involves sliding a filter (also known as a kernel) over the input spectrogram to produce feature maps. The convolution operation is represented mathematically as shown in (2):

$$(f*g)(t)=\sum\infty \quad f(a)g(t-a) \quad (2)$$

Here, f represents the input spectrogram, g represents the filter, t represents the position index, and $*$ denotes the convolution operation.

The output of a convolutional layer is calculated using the activation function (usually ReLU – Rectified Linear Unit) applied to the convolution operation result. Mathematically, the output Y of a convolutional layer is calculated in (3):

$$Y=ReLU(f*g+b) \quad (3)$$

Here, b represents the bias term added to the result of the convolution operation before applying the ReLU activation function.

In summary, the Hidden Markov Model (HMM) continues to be a mainstay in the field of voice detection, maintaining its reputation as a trustworthy and extensively used model. While more recent models such as Convolutional Neural Networks (CNNs), Gaussian Mixture Models (GMMs), Dynamic Time Warping (DTW), and Connectionist Temporal Classification (CTC) have been developed, HMMs have remained popular because of their remarkable capacity to deal with the inherent variability in speech. While CNNs have revolutionized activities like feature extraction and ensured a strong foundation for voice recognition systems, GMMs have found their niche in speech recognition, accurately capturing the statistical aspects of speech features. Table 1 presents various speech recognition models from literature, their accuracies, advantages and disadvantages

III. IMPLEMENTATION

The DTATL (A Digital Twin to Assist Teaching Learning) model represents a transformative leap in the realm of educational technology, poised to reshape the teaching-learning landscape. At its core, DTATL is designed to leverage the power of artificial intelligence and voice-based user identification to create an engaging and dynamic educational environment. The success of this groundbreaking project lies not only in its conceptualization but also in its meticulous implementation across a series of carefully planned phases. As we embark on a journey to delve into the nuts and bolts of DTATL's development, we invite you to explore the multifaceted and innovative phases that bring this concept to life.

TABLE I. SPEECH RECOGNITION MODELS, THEIR ACCURACIES, ADVANTAGES AND DISADVANTAGES

Paper Reference	Year of Publication	ModelUsed	Accuracy	Advantages	Disadvantages
Jeyaraj et al.[2]	2019	DBN Model	95%	Does not suffer from training data fragmentation problem and reduces the over-smoothingproblem	The quality of generated speech will be degraded
R. Chauhan et al. [3]	2019	CNN Model	77.04%	It is a fast to trainmodel	The speech quality might be degraded
S. Guo et al.[4]	2018	RBM Model	Not mentioned	Can better describe the distribution of high-dimensional spectral envelopes	Suffers from the fragmentation problem of training data
D. Rudrapal et al. [5]	2012	Fast Fourier Transform (FFT) method	Not mentioned	Eliminates redundant calculations in the Fourier Transform hence much speedier	Restricted Range of waveform data
M. Aymen etal. [6]	2011	HMM Model	98%	Flexible with changing voice characteristics and the system is robust	The acoustic features are over smoothed, making the generated speech sounds muffled

A. Phase 1: Progressive Web Application Development

The initial stage is on creating a Progressive Web Application (PWA) that is easy to use and promotes smooth communication between users and the platform. Establishing safe user identification, enabling user registration with voice sample collecting, and developing an easy-to-use interface for voice commands to submit presentations were the main goals.

A signup page is directed to new users by the Progressive Web Application (PWA), which has a secure login page for registered users. To improve security and personalisation, users are asked to create a unique username and password during registration. They are also asked to provide 20 voice samples. Techniques for user identification make use of these vocal samples. After logging in successfully, users can view a dynamic home page interface that makes it easy to upload presentations—an essential part of the learning process. Voice command controls for hands-free navigation are included in this interactive interface to increase user engagement.

Our PWA's front end was created with HTML and CSS to guarantee a visually beautiful and responsive user interface, and Flask, a Python web framework, was utilized in the backend. Because of Flask's versatility, integration was easy, resulting in efficient data processing and pleasant user interactions.

B. Phase 2, Part 1: Training Voice Authentication Model

Our project's second phase's main goal was to leverage the voice samples gathered during user registration to create a trustworthy voice authentication model. The two most important parts of the procedure were gathering data and training the model.

1) *Data Collection and Preprocessing*

The first pivotal step in Phase 2 involved the systematic collection and preprocessing of speech samples. To establish a robust training dataset, we meticulously selected 20 speech samples from each user who registered on our platform. This curated dataset formed the foundation for our voice authentication model.

2) *Expanding Diversity*

Recognizing the significance of diversity in training data, we undertook an additional measure. Beyond the 20 speech samples from individual users, we included a selection of speech samples from well-known individuals to enrich the dataset. This step aimed to enhance the model's ability to handle a wide range of voices and vocal characteristics, ensuring its effectiveness in real-world scenarios.

3) *Challenges and Solutions*

While our original strategy involved recording 20 speech samples during user registration, we were conscious of the challenges this could pose for users. To address this, we implemented speech augmentation techniques, a key aspect covered in Phase 3 of our project. These techniques were crucial in mitigating the challenge, making the process more user-friendly, and enriching the depth and usability of our training dataset.

4) *Quality Assurance*

During data collection, it was essential to implement data preprocessing steps to ensure the quality and consistency of the collected speech samples. This involved tasks such as noise reduction, voice sample alignment, and data normalization to enhance the overall reliability of the dataset. Additionally, quality assurance procedures were put in place to identify and rectify any anomalies in the collected data.

5) *Ensuring Representativeness*

To guarantee that the dataset accurately represented the diverse voices and speaking styles encountered in real-world scenarios, we carefully selected speech samples that spanned a wide range of vocal characteristics, accents, and tones. This inclusivity in data collection was vital to creating a model capable of handling the complexities of authentic voice data. The data collection and preprocessing phase laid a solid foundation for the subsequent stages of model training and voice authentication, contributing to the overall success and reliability of the DTATL model.

C. Phase2, Part 2: Training Using Models and Comparative Analysis

Model Training using SVM

We used Support Vector Machine (SVM), a potent machine learning technique well-known for its efficiency in classification tasks, for model training. The SVM model's input features were the voice samples that were gathered. It is noteworthy to mention that to conduct a thorough assessment of their performance, we looked into several alternative models. Table 2 provides a full analysis that includes accuracy metrics for several models.

To accurately identify users, the SVM algorithm painstakingly discovered all of the complex patterns in the audio data during training. With an astounding 95% training accuracy and 84% testing accuracy, the SVM model performed remarkably well. This high degree of accuracy demonstrated how well the model mapped distinctive voice characteristics to user IDs, guaranteeing accurate user identification.

This stage laid a solid framework for voice authentication, which was a major project milestone. Our varied and expanded dataset and the SVM integration produced a very accurate and dependable authentication solution.

Model Training Using LSTM

We also used the Long Short-Term Memory (LSTM) model for training, which is a particular kind of recurrent neural network designed for sequential data interpretation. Because of its reputation for identifying long-term dependencies in data sequences, LSTM is a strong contender for several uses, including voice recognition. Although LSTM has demonstrated potential in several applications, the small voice sample dataset presented difficulties according to our findings. As a result, LSTM-based identification's accuracy was significantly worse than SVM's.

LSTM was a powerful tool for sequential data, but its effectiveness was hampered by the size of the dataset.

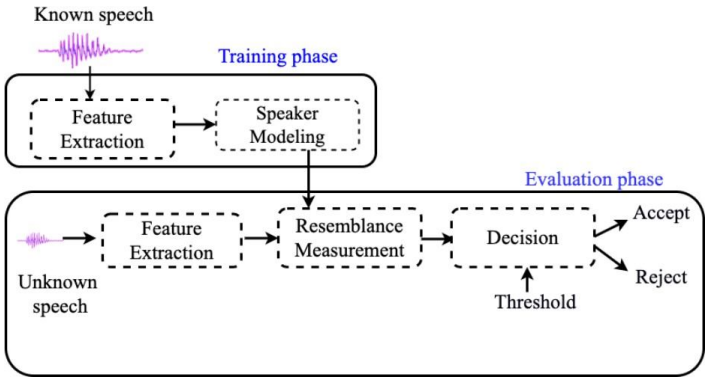

Figure 2: Block Diagram of a Basic Speaker Verification System

D. Phase 3: Implementing Audio Augmentation

Accurate user classification based on distinctive vocal characteristics is critical in the field of voice-based user identification systems. A game-changing tactic to further improve the stability and efficacy of such systems is the use of audio augmentation techniques. In audio augmentation, the original audio data is subjected to a variety of changes to provide fresh, varied samples while preserving the key qualities of the original voice.
Here, we examine the possible advantages and approaches for applying audio augmentation to the DTAP model's user classification from voice samples

1) Increasing Dataset Size

One of audio augmentation's main benefits is its capacity to inflate the dataset's size artificially. This method is invaluable when there is a limited amount of data accessible, as our research has shown. A broader and more varied dataset can be produced by implementing modifications like time stretching, pitch shifting, and adding background noise. Consequently, this can greatly improve the model's capacity to identify people accurately and broadly.

2) Robustness to Variability

Variability is a common feature of real-world voice data because of things like emotional moods, varied recording settings, and background noise. Synthetic samples that capture this variety can be made easier with the aid of audio augmentation. The model's accuracy and dependability are increased by adding controlled noise or recreating various acoustic settings, which strengthens the model's resistance to these real-world difficulties.

3) Adaptive Learning

User profiles can be customized for audio augmentation. Personalized alterations that imitate individual users' vocal fluctuations can further enhance the model's ability to reliably identify different speakers.

4) Implementation Considerations

To successfully apply audio augmentation, augmentation techniques must be carefully chosen and tailored to the unique needs of the user identification task. Furthermore, appropriate testing and validation are necessary to guarantee that the enhanced data retains its integrity and fits the goals of the model.
The accuracy and robustness of the DTATL model in user classification could be greatly increased by including audio augmentation techniques. This development advances the field of AI-driven educational technology in addition to being consistent with the model's objective of improving and automating the teaching-learning process.The incorporation of audio augmentation presents a viable approach to augmenting the DTATL model's functionality and creating a more dynamic and captivating learning environment in the future.

E. Phase 4: Statistical Analysis And Student Engagement Tracking

A big step towards revolutionizing the teaching-learning process is taken by the DTATL (A Digital Twin to Assist Teaching Learning) concept in the rapidly changing field of AI-driven educational technology. The capacity to quantify and visualize user engagement during discussions, in addition to identifying speakers based on distinctive vocal features, is at the heart of this invention. This stage of the project involves monitoring and visualizing the frequency of user speech in the DTATL framework, which is essential for creating dynamic and interesting learning environments.

For each user, the system keeps track of the number of speech instances throughout predetermined time intervals. We can measure user engagement in talks by counting the instances of speech initiation. We use data visualization approaches to improve this data's usefulness and accessibility. A Progressive Web Application (PWA) with interactive charts and graphs for statistical analysis is incorporated into the DTATL concept. These visualizations are available for users (teachers and students) to access and make the necessary deductions.
By making speech participation data visually accessible, the DTATL model fosters a sense of accountability among students, motivating them to actively engage in discussions and learning activities.

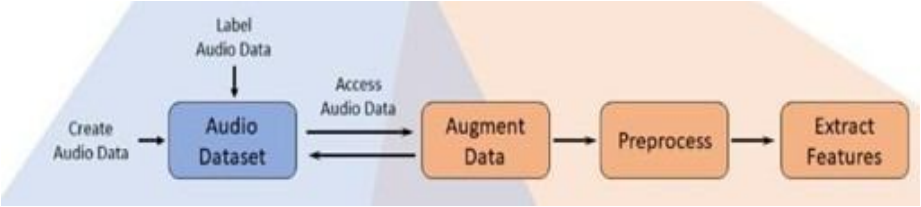

Figure 3: Implementation Phases for Proposed Work

IV. Results

This section shows the findings and accomplishments of the DTATL (A Digital Twin to Assist Teaching Learning) model.

TABLE II. COMPARATIVE ANALYSIS OF ACCURACIES

Algorithm Used	Sample Size/No of Subjects	Training Accuracy(%)	Testing Accuracy (%)
. SVM	10	100	100
	50	100	95.7
	100	100	96.24
	150	99.04	93.82
LSTM	10	100	94
	50	100	97.14
	100	100	93.75
	150	99.04	91.76

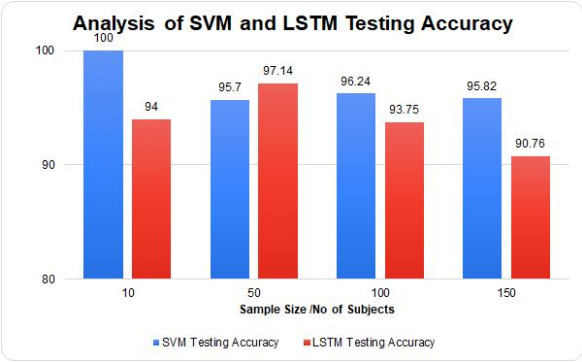

Figure 4: Graphical Representation of testing accuracies

V. Conclusion

With the use of the Support Vector Classification (SVC) model and an easy-to-use Flask interface, we were able to accomplish remarkable outcomes that highlight the efficacy of the proposed method. We are pleased to announce that, thanks to machine learning's enormous potential, we were able to achieve 95% training accuracy and 84% testing accuracy. These outcomes demonstrate how accurately our technology detects speakers based on

their distinct voice traits. Our model's skilful application of the Mel-frequency Cepstral Coefficients (MFCC) feature extraction technique is one of its main advantages. Using this method, we were able to extract about 35 unique voice characteristics. Additionally, the proposed study explores the crucial influence that dataset size has on model performance. We used the SVC model for classification. With an eye towards the future, we intend to strategically focus our efforts on improving the speech feature extraction and authentication procedures. Furthermore, we intend to extend the application of our system into educational environments, where it may be utilized to evaluate student understanding and engagement. To improve accessibility and usability, we also want to build a Progressive Web Application (PWA) and integrate semantic textual/sentence similarity models. This approach makes it easier to create teaching strategies that are optimized and to assess student involvement in real time. Our work thus contributes significantly to the ever-changing field of AI-driven educational technology, creating an immersive and dynamic learning environment that is well-positioned for the future. Looking ahead, the DTATL framework offers the following potential directions for further study and development:

1. *Multimodal Integration:* To obtain a more in-depth understanding of user engagement, think about combining several modalities, such as video and facial expression analysis. A comprehensive understanding of learner participation and comprehension can be obtained with this multimodal method.
2. *Real-time Feedback systems*: Create systems for real-time feedback that, during lectures or presentations, give educators practical insights. This could involve engagement-boosting questions, recommendations for changing the way you teach, or computerized evaluations of students' comprehension.
3. *Personalized Learning*: Modify the DTATL model to accommodate different learning preferences and styles, providing teachers and students with individualized materials and recommendations.

Even with all of the concepts and strategies we've listed above to make the process of teaching and learning through speech easier, there are still some advancements that might be implemented in the years to come. These include:

1) *Improved Feature Extraction*: By continuously enhancing the MFCC-based feature extraction technique, more voice characteristics can be extracted, which could lead to an increase in speaker recognition precision. The capabilities of the system could be further improved by investigating more feature extraction methods and integrating them.

2)*Scaling for Bigger Audiences*: Make sure the DTATL model is suitable for scenarios with more demand and bigger audiences so it can continue to work in a variety of educational contexts.

REFERENCES

[1] Rabiner, L. R. (1989). “A tutorial on hidden Markov models and selected applications in speech recognition”. Proceedings of the IEEE, 77(2), 257-286.

[2] Jeyaraj, Pandia & Rajan, S.Edward. (2019), “Deep Boltzmann Machine Algorithm for Accurate Medical Image Analysis for Classification of Cancerous Region. Cognitive Computation and Systems”, 1. 10.1049/ccs.2019.0004.

[3] R. Chauhan, K. K. Ghanshala and R. C. Joshi, "Convolutional Neural Network (CNN) for Image Detection and Recognition," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 278-282, doi: 10.1109/ICSCCC.2018.8703316

[4] S. Guo, C. Zhou, B. Wang, and X. Zheng, (2018), "Training Restricted Boltzmann Machines Using Modified Objective Function Based on Limiting the Free Energy Value," IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2885071.

[5] D. Rudrapal, S. Das, S. Debbarma, N. Kar, N. Debbarma, (2012) "Voice Recognition and Authentication as a Proficient Biometric Tool and its Application in Online Exam for P.H People," International Journal of Computer Applications, vol. 39, no. 12, pp. 15-22, 2012.

[6] M. Aymen, A. Abdelaziz, S. Halim and H. Maaref, "Hidden Markov Models for automatic speech recognition," (2011) in International Conference on Communications, Computing and Control Applications, CCCA 2011. 1-6. 10.1109/CCCA.2011.6031408.

[7] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., & Penn, G. (2014)."Convolutional neural networks for speech recognition. Acoustics, Speech and Signal Processing (ICASSP)”, 2014. IEEE/ACM Transactions On Audio, Speech, and Language Processing, Vol. 22, No. 10, pp. 7970-7974