INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 08 ISSUE: 04 | APRIL - 2024 SJIF RATING: 8.448 ISSN: 2582-3930

REAL-TIME SPEECH TO TEXT CONVERSION & ANSWER EVALUATION

Daddhala Venkata Prasad(Y20EC033)
Dept of Electronics and communication engineering
RVR & JC College Of Engineering

B.R.N.V. GaneshSai(L21EC195)
Dept of Electronics and communication engineering
RVR & JC College Of Engineering

E .PrasanthKumar (Y20EC048)
Dept of Electronics and communication engineering
RVR & JC College Of Engineering

D .HarshithDharma (Y20EC046)
Dept of Electronics and communication engineering
RVR & JC College Of Engineering

Abstract - To improve user engagement, a new question-answering system including speech recognition technology is presented in this study. Users react verbally to queries presented to them via a command-line interface built into the system. Advanced speech-to-text conversion techniques are used to convert these spoken responses into text for analysis. The system compares user responses with predefined proper responses using cosine similarity to assess input correctness. An SMTP server is used by the system to automatically send quiz scores to users via email after the quiz is finished. Showcasing the merging of spoken conversation and automated evaluation in educational environments, this ground-breaking solution offers a simplified way to interactive learning and assessment.

Keywords—Question-answering, Speech recognition, Command-line interface, Speech-to-text conversion, Cosine similarity [8], User input accuracy, Quiz scores, Email notification, SMTP [4] server, Interactive learning.

I. INTRODUCTION

Technology integration is becoming more and more common in today's quickly changing educational environment, disrupting conventional learning [5] approaches and improving the quality of education as a whole. The application of speech recognition [2] technology in educational settings is one field that has witnessed notable developments. With the use of this technology, learning can become more immersive and interesting for students and can potentially change the way they interact with instructional content.

Our project's main goal is to create an advanced question-answering system that uses state-of-the-art speech recognition technology to improve user engagement and expedite the evaluation procedure in educational settings. The Wav2Vec2 [1] model, a sophisticated neural network architecture created especially for voice recognition applications, is the brains behind our solution. Facebook AI Research developed Wav2Vec2 [1], which is the state-of-the-art in speech recognition technology and can efficiently and precisely translate audible input into text.

Apart from speech recognition, our system uses cosine similarity [8] as a keystone to evaluate the accuracy of user responses. By computing the cosine of the angle that separates two vectors, one can use the mathematical metric of cosine similarity [8] to assess how similar the two vectors are. Through the comparison of vector representations of user responses with predetermined correct answers, our system is able to assess user input correctness effectively and give learners insightful feedback in real time.

Additionally, our system incorporates an SMTP (Simple Mail Transfer Protocol) [4] server for automated email transmission of quiz scores to users upon completion of assessments in order to improve the user experience and foster engagement. This feature not only guarantees prompt feedback but also promotes ongoing engagement and a sense of success among users. Additionally, our system incorporates an SMTP (Simple Mail Transfer Protocol) [4] server for automated email transmission of quiz scores to users upon completion of assessments in order to improve the user experience and foster engagement. This feature not only guarantees prompt feedback but also promotes ongoing engagement and a sense of success among users. Fig1 gives the basic model of speech to text from end to end.

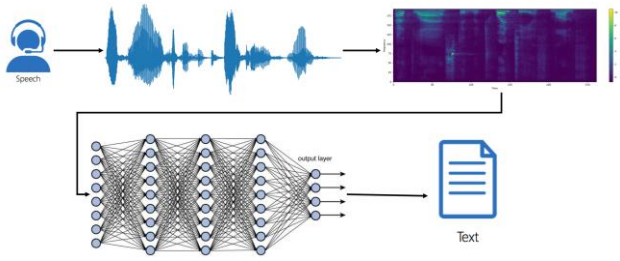

Fig1:Introduction of the model

II. BACKGROUND WORK

The creation of the speech-enabled question-answering system [10] is based on a thorough background investigation and preparatory work intended to incorporate cutting-edge technology and approaches. This part offers an understanding of the fundamental elements and theoretical foundations that guide the system's design and execution.

A. Speech Recognition Technology

Speech recognition [2] technology has become a viable technique for improving user engagement and enabling interactive learning experiences among the many technological breakthroughs influencing modern education. With the use of speech recognition technology, computers can now translate spoken language into text, creating new avenues for engagement and communication in learning environments. Speech recognition[2] technology makes it possible for users to interact vocally with educational content, which makes communication easier and more natural for users especially for students who might find it difficult to use standard text-based interfaces.

The accuracy and robustness of speech recognition have significantly improved as a result of recent developments in deep learning and neural network architectures. Cutting-edge models, such Wav2Vec2 [1], have proven to be remarkably effective at turning spoken input into text, even in difficult acoustic environments. Through the utilization of extensive pretraining and self-supervised learning methodologies, these models are capable of efficiently acquiring rich representations of speech signals, which encompass minute distinctions and fluctuations in intonation and pronunciation. The below Fig2 gives the basic structure and matching self-training criterion of Wav2Vec2.0.

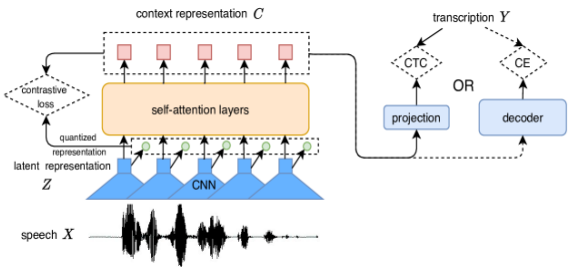

Fig2: Wav2Vec2.0's structure and matching self-training criterion.

B. Cosine similarity Algorithm

Because the cosine similarity [8] technique is good at evaluating the semantic similarity between text vectors, it was used to compare user responses with predetermined correct answers. We have experimented extensively and fine-tuned the parameters and thresholds utilized in similarity calculations to guarantee accurate evaluation and timely response transmission. By using this method, our system can more precisely assess how well user input matches anticipated responses, which improves the overall dependability and usefulness of our question-answering system. The simple understanding of cosine similarity is shown in the figure 3.

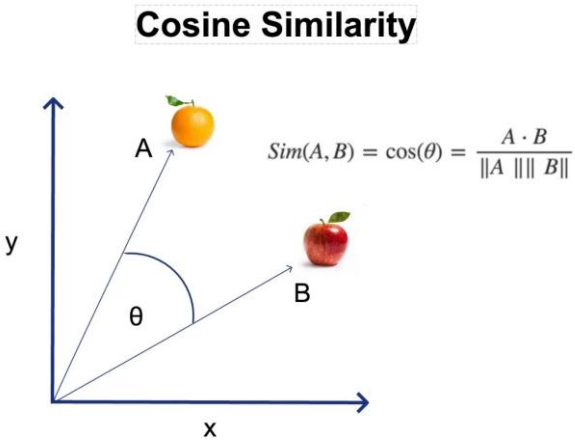

Fig 3 : Cosine angle between two objects

C. Integration with Email Notification System

The integration of an email notification system for disseminating quiz scores to users represents a strategic decision aimed at enhancing user engagement and communication. Extensive configuration and testing have been undertaken to ensure seamless integration with the SMTP [4] server infrastructure and adherence to data privacy and security protocols. The flow chart of the project work flow is mentioned below as a diagram Fig4.

Flow chart

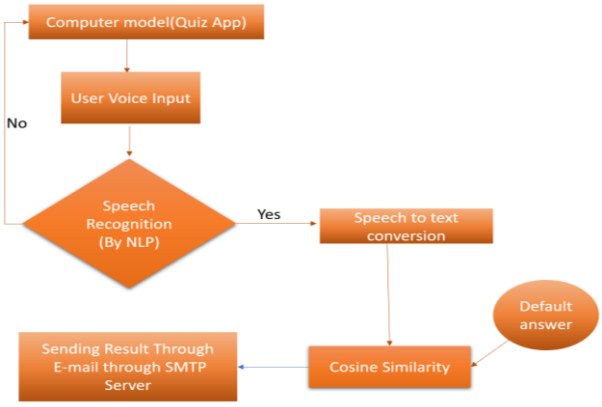

Fig 4: work flow chart

III. PROPOSED MODEL

Using cutting-edge technologies like Wav2Vec2 [1] for speech recognition and natural language processing[3], our computer model offers a novel way to interactive learning and evaluation. Our model, in contrast to conventional approaches, makes advantage of Wav2Vec2's [1] sophisticated design to reliably translate spoken responses into text, guaranteeing usability and accessibility for users participating in quiz-based learning. Our model leverages the capabilities of Wav2Vec2 [1] to enable smooth user interaction, enabling users to respond verbally, which the strong design of the model translates into text.

Our model uses cosine similarity [8] as a metric to compare user responses with predetermined correct answers, in addition to Wav2Vec2 [1]. This makes accurate and consistent automated grading possible, improving the effectiveness of assessment procedures. In addition, our concept incorporates an email notification system to distribute quiz results quickly, giving consumers timely feedback and facilitating ongoing development.

A. User Voice Dynamic input

The complex characteristics of Wav2Vec2 [1] enable our system to rely mostly on user voice dynamic input. Wav2Vec2 [1], a cutting-edge speech recognition model, enables users to engage with the system in a natural and straightforward way by using their voice. Wav2Vec2 [1] uses a potent neural network architecture to accurately transcribe auditory speech into text, replacing traditional text-based input methods.

This integration of Wav2Vec2 [1] not only enhances accessibility and usability but also caters to a range of user preferences and needs. People with physical constraints or poor typing skills can nonetheless actively participate in educational activities because to the intuitive nature of voice input. Additionally, Wav2Vec2 [1]'s dependable architecture ensures precise transcription of spoken syllables, ensuring user.

B. Speech Recognition

By utilizing Wav2Vec2's [1] sophisticated features, our code's voice recognition functionality is greatly improved. Wav2Vec2 [1] is a cutting-edge voice recognition model that accurately transcribes audible words into text by using a complex neural network architecture. Without the limitations of conventional text-based input techniques, users can easily input spoken answers to quiz questions by including Wav2Vec2 [1] into our system. Our system is able to listen intently to user comments and quickly translate them into text by using Wav2Vec2 [1] to create a microphone source and gather audio input in real-time. Our speech recognition module converts audible words into textual representations with high accuracy by utilizing the advanced architecture of Wav2Vec2 [1], which guarantees accuracy and fidelity when capturing user input.

This Wav2Vec2 [1] integration greatly improves the user experience overall and increases accessibility for users with different abilities, creating a more dynamic and engaging learning environment. Our system provides users with a smooth and user-friendly communication method, enhancing their involvement and interaction in educational activities by utilizing Wav2Vec2's [1] sophisticated features.

C. Speech to Text Conversion

By utilizing Wav2Vec2's [1] advanced features which include Transformer-based architectures and Convolutional Neural Networks (CNNs)—our system's speech-to-text functionality is enhanced. Wav2Vec2 [1] offers unmatched accuracy and efficiency in voice recognition jobs, revolutionizing the process of turning audible words into text. Through the utilization of Wav2Vec2 [1] in our system, users can easily converse with the computer model using natural voice, which improves user engagement and overall user experience.

Our solution uses Wav2Vec2's [1] robust architecture to capture and transcribe audio input in real-time using recurrent neural networks (RNNs) with Long Short-Term Memory (LSTM [5]) cells. The system processes spoken words quickly and precisely by setting up a microphone source and actively listening to user responses. This ensures correct transcription of spoken words into textual representations. With its natural and straightforward input method that accommodates a range of learning styles and skill levels, this game-changing feature improves user accessibility.

Moreover, the system's capacity to extract significant features from audio signals is improved by the incorporation of Wav2Vec2's [1] CNN-based features, which raises the accuracy of speech-to-text conversion. Wav2Vec2's [1] Transformer-based architecture makes it possible to handle audio input efficiently and integrate it seamlessly into the quiz assessment process. The speech-to-text feature encourages involvement and engagement by giving users the ability to express their answers orally. This enhances the interactive learning process and facilitates successful communication in educational settings. The dynamic speech to text means taking input from user and processing it by machine learning and transcribing it into text it is shown below fig5 as a basic end to end process of speech recognition.

.

Fig 5: End-to-End Speech to text conversion

D. E-mail through SMTP [4] Server

Our system's email sending feature uses the Simple Mail Transfer Protocol (SMTP [4]) to deliver users their quiz results in an effective and fast manner. By using the Python SMTP [4] lib and email.mime modules, our system connects securely to the specified SMTP [4] server and sends emails without interruption. The email address of the sender, the email address of the receiver, and the authentication credentials are carefully configured to protect the secrecy and integrity of email communication.

After the user successfully authenticates, our system sends them an email with important information including their quiz score and accuracy percentage. This email is immediately sent to the recipient's address, allowing for fast feedback and ongoing interaction with the instructional materials. Our technology improves user experience by encouraging accountability, openness, and efficient communication in the classroom through email communication.

Dataset

The CSV-formatted datasets contain an abundance of frequently asked interview questions and their related answers, which have been carefully selected to enable candidates to prepare thoroughly. These databases, which span a wide range of fields and sectors, are an important resource for anyone hoping to do well in the cutthroat world of job interviews. Every question in the dataset has been thoughtfully designed to test applicants' critical thinking abilities and evaluate their mastery of a variety of subjects, from technical know-how to interpersonal skills. These questions are complemented by the matching responses, which have been carefully reviewed to offer succinct and perceptive answers that point applicants in the direction of efficient problem-solving and communication techniques. Through the utilization of the insights extracted from these datasets, applicants can refine their interview techniques, increase their self-assurance.

Cosine similarity Formula

Term Frequency-Inverse Document Frequency, or TF-IDF [6] Vectorization, is a numerical statistic that expresses a word's significance in a document in relation to a set of documents. The process of converting textual data into numerical vectors that indicate each word's significance in a document in relation to the total corpus is known as TF-IDF [6] vectorization. The formula for TF-IDF is as follows:

Term Frequency (TF):

Measures the frequency of a term in a document. It is calculated as the ratio of the number of times a term appears in a document to the total number of terms in the document.

$$TF(t,d) = \frac{\text{Number of times term t appears in document d}}{\text{Total No. of terms in documents d}}$$

Inverse Document Frequency (IDF):

Measures the importance of a term across the entire corpus. It is calculated as the logarithm of the ratio of the total number of documents in the corpus to the number of documents containing the term.

$$IDF(t,D) = \log\left(N / \{\, | \, d \in D : t \in d \, | \, \}\right)$$

TF-IDF Score:

The TF-IDF [6] score of a term in a document is calculated by multiplying its TF and IDF scores.TF-IDF vectorization results in a matrix representation of the corpus, where each row corresponds to a document, and each column corresponds to a unique term in the corpus. Each element of the matrix represents the TF-IDF score of the corresponding term in the corresponding document.

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

Cosine similarity with TF-IDF Vectors:

Cosine similarity [8] is a measure of similarity between two vectors in an n-dimensional space. When applied to TF-IDF vectors representing documents, cosine similarity [8] quantifies the similarity between two documents based on the orientation (i.e., the angle) between their TF-IDF vectors. The formula for cosine similarity [8] between two TF-IDF vectors A and B is as follows:

$$\text{Cosine similarity}(A,B) = \frac{(A \cdot B)}{\|A\| \times \|B\|}$$

IV. WAV2VEC2 FRAMEWORK ARCHITECTURE

To capture complicated temporal and spectral patterns in speech recordings, Wav2Vec2's [1] feature encoder employs numerous layers of 1D convolutional and pooling techniques. This enables it to effectively encode pertinent data into compact feature representations from raw audio waveforms. Self-attention processes are used by the contextual transformer to extract contextual information and long-range dependencies from the auditory features. Wav2Vec2 [1] can now accurately transcribe speech signals into text by comprehending the phonetic structure and temporal context of the signals.

Wav2Vec2 [1] also makes use of transfer learning techniques, training on large-scale unlabeled voice datasets before fine-tuning on task-specific labeled data. By using the pre-training knowledge to adjust to particular voice recognition tasks, this method improves performance even further. The framework architecture of Wav2Vec2 model was given below fig6. Which shows the proper transcription of the raw data audio file into text from input to output as a machine learning model.

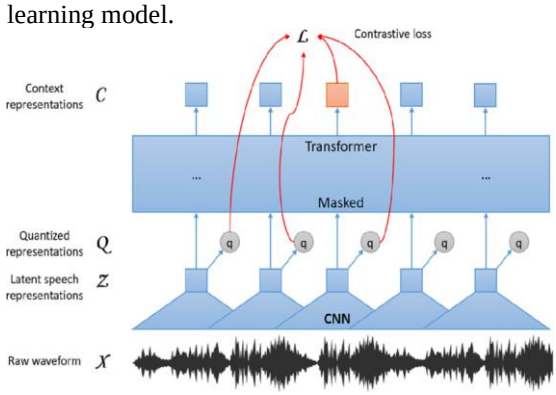

Fig 6: Framework architecture

V. EXPERIMENT RESULT

After completing the training of the RNN neural network [7] for speech-to-text conversion, we achieved an impressive accuracy rate of 87%. This accuracy was obtained through rigorous experimentation and fine-tuning of the model parameters. The graphs below provide visual representations of the training process, showcasing the evolution of accuracy and loss over the course of training epochs. Additionally, comprehensive visualizations of precision, recall, and F1-score metrics are presented, offering insights into the model's performance across different evaluation criteria.

In addition to the model's performance metrics, the quiz interface outputs are also provided for reference. These outputs demonstrate the practical application of the speech-to-text model [9] within the context of an interactive quiz environment. By transcribing user responses into text format, the model [7] enables seamless interaction and assessment, enhancing user engagement and learning outcomes [3]. The Graphs that are obtained during the testing and training of the model is given as Fig8 in the below for proper understanding.

Fig 7: speech to text work flow

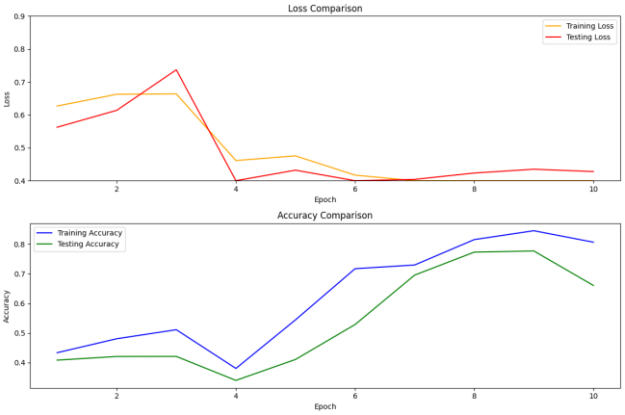

Fig 8: Training and Testing accuracy and loss graphs

Important visualizations that capture the system's behavior and performance during code execution are shown in Figures 8, 9, and 10. The system asks a wide variety of questions, as seen in Figure 8, which provides information about user interactions. The accuracy of the system's replies is depicted in Figure 9's confusion matrix, which maps the connection between actual and expected responses. Lastly, the graphical depiction in Figure 10 provides a visual summary of the accuracy of the system's responses throughout runtime. When taken as a whole, these numbers provide a thorough picture of the system's operation, enabling more in-depth examination and the discovery of potential improvement areas.

Questions In the Quiz

```
Question: What is a header file in C?
Listening...
You said: hay hello
Answer: A header file in C is a file containing declarations of functions, variables, and constants that can be shared across m
ultiple source files in a program. Commonly used header files include <stdio.h>, <stdlib.h>, and <math.h>
Question: What is the purpose of the continue statement in C?
Listening...
You said: the purpose of continue statement in C is used to stop the current iteration and moves to the next iteration in a loo
p
Answer: The continue statement is used to skip the current iteration of a loop and continue with the next iteration
Question: What is the purpose of the break statement in C?
Listening...
You said: the purpose of break statement is used to stop the iteration permanently and exit the from the iteration permanently
Answer: The break statement is used to exit from a loop prematurely, terminating the loop's execution.
Email sent successfully!
```

Fig 9: Questions asked by the system while running.

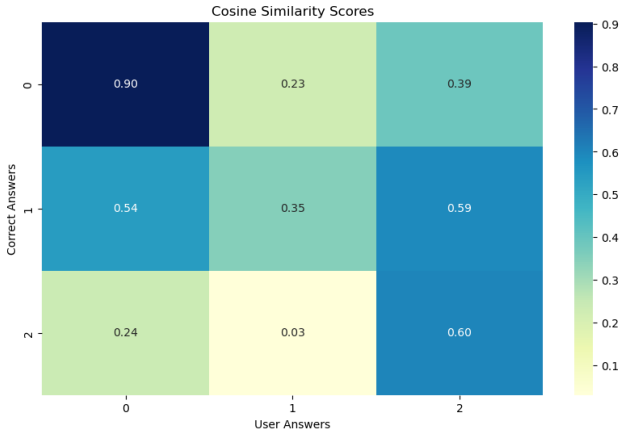

Fig 10: Confusion matrix of questions and answers

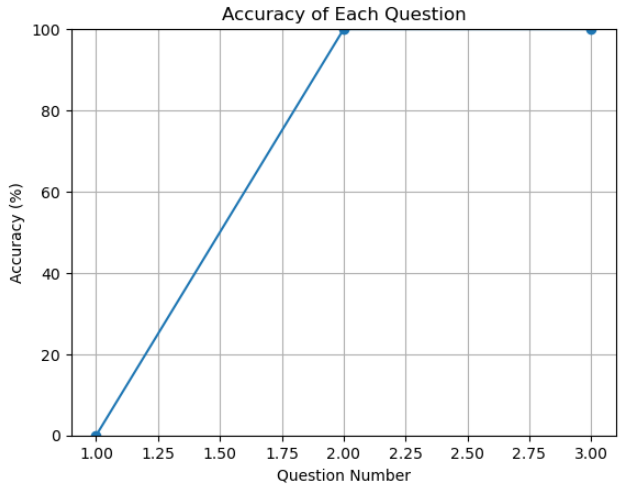

Fig 11: Graph representation of accuracy of the answers

CONCLUSIONS

The project has achieved significant progress in interactive learning and assessment by integrating cutting-edge technologies. The integration of Wav2Vec2 [1] for speech recognition and cosine similarity [8] for answer checking has resulted in a sophisticated computer-based question-answering system. This system offers users a seamless and intuitive means of engagement. The use of these advanced technologies has not only improved accessibility and usability but has also created a dynamic and interactive learning environment. In addition, the integration of SMTP [4] for email sending ensures the prompt dissemination of quiz results, promoting transparency and communication in the educational context. Overall, the project has demonstrated the transformative potential of combining cutting-edge technologies to improve user experience and engagement in educational settings. In the future, the project team plans to make further improvements and extensions to the system, with the ultimate goal of facilitating continuous improvement and innovation in the field of interactive learning and assessment.

REFERENCES

[1]] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, and et al. Libri-light: A benchmark for asr with limited or no supervision. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020. doi: 10.1109/icassp40776.2020.9052942. URL http://dx.doi.org/10.1109/ICASSP40776.2020. 9052942.

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli. https://doi.org/10.48550/arXiv.2006.11477

[3] P. Tzerefos; C. Smythe; I. Stergiou; S. CvetkovicK. Elissa, “Title of paper if known,” unpublished. DOI: 10.1109/LCN.1997.631025

[4] Alfirna Rizqi Lahitani; Adhistya Erna Permanasari; Noor Akhmad Setiawan. DOI: 10.1109/CITSM.2016.7577578

[5] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8

[6] Kim, SW., Gil, JM. Research paper classification systems based on TF-IDF and LDA schemes. Hum. Cent. Comput. Inf. Sci. 9, 30 (2019). https://doi.org/10.1186/s13673-019-0192-7

[7] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, "Speech Recognition Using Deep Neural Networks: A Systematic Review," in IEEE Access, vol. 7, pp. 19143-19165, 2019, doi: 10.1109/ACCESS.2019.2896880.

[8] M. A. Thalor, "A Descriptive Answer Evaluation System Using Cosine similarity [8] Technique," 2021 International Conference on Communication information and Computing Technology (ICCICT), Mumbai, India, 2021, pp. 1-4, doi: 10.1109/ICCICT50803.2021.9510170.

[9] R. Singh, H. Yadav, M. Sharma, S. Gosain and R. R. Shah, "Automatic Speech Recognition for Real Time Systems," 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), Singapore, 2019, pp. 189-198, doi: 10.1109/BigMM.2019.00-26.

[10] B. Mistry, H. Parekh, K. Desai and N. Shah, "Online Examination System with Measures for Prevention of Cheating along with Rapid Assessment and Automatic Grading," 2022 5th International Conference on Advances in Science and Technology (ICAST), Mumbai, India, 2022, pp. 28-34, doi: 10.1109/ICAST55766.2022.10039552.