Automatic Speech Grading using a Multimodal Deep Learning Framework using Bert and Whisper

M. Hemantheswar Reddy[1*], K. Rishitha[2], P. Bharath Raj[3], D N Kiran Pandiri[4] and U. Thulasi Srinivas[5]
{hemanth14082004@gmail.com[1], rishithakattamudi998@gmail.com[2], bharathpechetti@gmail.com[3], pdnkiran@gmail.com[4], uppalapatisrinivas26@gmail.com[5]}

Department of ACSE, VFSTR Deemed to be University, Guntur -Tenali Rd, Vadlamudi, Andhra Pradesh, India[1, 2, 3, 4, 5]

Abstract. This paper proposes a Natural Language Processing (NLP-based) program of speech grading for not only the audio but also the video portion that quantitatively evaluates speech in terms of grammar, vocabulary, pronunciation, fluency and accuracy. These conventional speech evaluation methods are prone to be subjective, inefficient, low feedback, and thus limit their application in overall assessment. The proposed system is a system that combines Automatic Speech Recognition (ASR) models such as Whisper that transcribe speech to text and then Natural Language Processing (NLP) technologies that analyze and score them in a standardized way. By providing plentiful and actionable feedback, the system has the potential to improve the reliability and consistency in assessment of speech. This technique has broad uses in education, recruitment, and communication training, provides a scalable and objective approach towards speech measurement.

Keywords: Speech Grading, Automatic Speech Recognition, Whisper, NLP, Pronunciation Scoring, Fluency Measurement.

1 Introduction

Measurement of spoken communication is needed in order to give an effective score in areas like language instruction, education, or hiring. Effective discourse relies on fluency, coherence, pronunciation, grammar and the use of a rich vocabulary. Traditional assessments commonly-produce subjectivity, variation in results and limited capacity for expansion as they rely on human raters. While such manual evaluations can provide palatable insights, they are inherently susceptible to rater bias and fatigue and might become prohibitively slow in a large-scale setting.

Besides, the subjective nature of traditional techniques restrain the scope of input and call for automatic systems ensuring consistent, unbiased and full evaluations. Recent advances in Automatic voice Recognition (ASR) technology and Natural Language Processing (NLP) have enabled intelligent voice scoring systems. Whereas NLP methods asses quality in terms of structure, lexical diversity, and fluency ASR models like Whisper output transcriptions that are accurate and consistent in varying acoustic conditions. When combined with deep learning models, such systems can score and describe accurately. The system architecture at work here validates pronunciation, fluency, and content via NLP analysis and broadcasts this to Whisper for transcription. The integration, for both the assessors and the students, offers consistent scoring, scalability, and descriptive feedback. The characteristics make it more efficient, less biased, and easy to be placed into a large-scale setup, comparing with other manner to conduct the experiment.

ICITSM-Part I 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357788

Despite this, challenges such as computing loads, speech fluctuation, and dataset diversity still exist. Domain adaptation techniques and model architecture tuning are required to solve this. For convenience of real-time evaluation in educational and working environments, the paper here introduces a scalable, accurate, and objective speech grading system using ASR and NLP. The method, dataset, and comparative analysis employed to gauge its effectiveness are elaborated in detail in the subsequent sections. The structure of this paper is as follows. Section 2 is a comprehensive review of literature related to automated speech grading, with an emphasis on modern methods, evaluation metrics, and their drawbacks. Section 3 is the methodology used in this work, including the model architectures of the grading models (BILSTM, FCNN, and BERT with Whisper), the audio preprocessing pipeline, and the Whisper-based transcribing process. Comparative experimental data analysis is discussed in Section 4, comparing the performance of each model based on pronunciation and fluency metrics. The important findings are discussed in detail in Section 5, which also gives the strengths, weaknesses, and suitability of each model for deployment. The important findings and implications for speech grading are concluded in Section 6.

The key contributions of this paper are outlined as follows:

- This work integrates advanced pretrained language models to enable deeper semantic and syntactic analysis, enhancing the overall quality and accuracy of automated essay evaluation.
- A comprehensive, multi-dimensional scoring mechanism is introduced, which evaluates various aspects of written responses including content relevance, grammatical correctness, fluency, and vocabulary usage ensuring a holistic assessment approach.
- The system generates detailed, feedback-oriented outputs designed to provide meaningful insights for learners to improve their speaking, as well as actionable information for educators to support instructional decision-making.

2 Literature Review

Recent developments in speech processing have created an increasing amount of work on automatic assessment and transcribing accuracy with the aid of ASR and NLP methods. Ramacharla et al. [1] created an internet-based transcription system with the aid of PyDub, MoviePy, and the Google Speech API with the users' emphasis on accessibility. However, in noisy and uncontrolled acoustic conditions, the system demonstrated only limited robustness. In a related work, Kobylvukh et al. [2] compared a number of ASR systems for the Ukrainian language and concluded that grammatical complexity and accent variation in low-resource languages pose serious challenges to transcription accuracy.

To improve reliability, Ziman et al. [3] used the Google Cloud Speech API to transcribe speech from psychological tests with confidence-based grading. However, their results 'wider relevance is limited by the fact that there are no comparisons with other engines. Agre et al. [4] used the MSER algorithm to text extraction from video, but the process had difficulty separating text from complex visual backgrounds reliably. Saraswathi et al. [5] proposed a multimodal transcription system, which used Pytesseract for OCR and OpenCV for frame extraction.

When the system worked well under perfect visual conditions, low-resolution frames and visually intensive pictures made it suffer with a drastic fall in performance. CNN and RNN models using MFCC features were utilized in Dhale et al. [6] to test deep learning-based transcription. Their technology was limited by background noise and speaker characteristics despite its promising results. In a comparison of Whisper ASR models with Javanese speech materials, Pratama and Amrullah [7] showed that the models worked well in linguistic environments with few resources.

Future work is needed because their results didn't generalize to multimodal or audiovisual speech recognition. Aswin et al. [8] suggested an ensemble NLP pipeline for subtitle generation and video summarization using Luhn's algorithm, LSA, TextRank, and LexRank. Although the architecture is new, scalability in unannotated domains is limited by reliance on existing subtitles. Kavitha et al. [9] suggested a framework for Tamil speech recognition using MFCC and DTW for automatic text and speech grading. Variation in performance was controlled mainly by speaker-specific features like pitch and accent. Suresh et al. [10] used the ASAP dataset to automate essay grading using LSTM and DNN architecture but faced deployment consistency and scalability challenges. Based on features like pitch, rhythm, and intonation, Xu [11] describes a DNN-based model for English speech recognition and pronunciation scoring. The model is better than traditional methods, provides unbiased and real-time feedback, and can be used in real-world learning systems. However, it does not have the ability to score semantics and is limited by acoustic-level analysis, which is not sufficiently holistic. Jones et al. [12] provide a comprehensive description of automated writing and speaking score methods. Although they describe relevant issues with fairness, model architecture, and combining AI and human judgment into evaluation pipelines, they mention that deep learning significantly enhances the accuracy and efficiency of scoring. Although deep learning provides much superior efficiency and accuracy, the authors identify limitations in terms of generalizability, interpretability, and transparency. For deployment in schools to be successful, the study identifies the need for hybrid approaches, domain-level adaptation of models, and having continuous validation for ensuring fairness and pedagogic alignment.

With Word2Vec embeddings and Naïve Bayes classification over the Twitter corpus, Sadanand et al.(2022) investigate essay grading through combining sentiment analysis and natural language processing methods. Syntactic issues, coherence, argument structure, and logical coherence are all overlooked by the focus of their system on syntactic issues [13]. Ghanta (2019) applies automated essay grading with the Hewlett dataset, Coh-Metrix tool, and various regression methods, i.e., Random Forest, Linear Regression, and Support Vector Regression (SVR). The strength of assessments is hampered by the model's incapability in efficiently handling stylistic and rhetorical variety, whichhinders its performance [14].

Even with advancements in automated assessment systems, certain challenges continue to exist in current methodologies. They are mostly rule-based or fuzzy logic-based, which, although interpretable, are not adaptive enough to accommodate linguistic variability and open-ended, complex responses. This reduces their generalizability and effectiveness in actual classrooms. Excessive reliance on commercial NLP APIs also limits transparency, flexibility, and privacy of data. These black-box systems prevent model interpretability and domain-specific adaptation. Additionally, current research focuses on shallow aspects like grammar and basic coherence at the expense of underlying semantic processing and discourse-level analysis. Moreover, insufficient feedback mechanisms and assessment measures often limited to accuracy or

correlation fail to report fairness, robustness, or educational efficacy. These limitations point towards the necessity of a multimodal, pedagogically aligned, and flexible assessment system. Moreover, insufficient feedback mechanisms and assessment measures often limited to accuracy or correlation fail to report fairness, robustness, or educational efficacy. These limitations point towards the necessity of a multimodal, pedagogically aligned, and flexible assessment system.

3 Methodology

3.1 Dataset Overview

This study makes use of the Speech Content, Fluency, and Pronunciation Scores dataset [15]as shown in fig.1. The following standards are used to evaluate each sample:

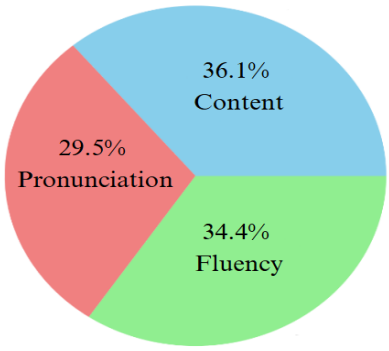

Fig. 1. Evaluation standards for speech sample.

- **Content:** Evaluates the spoken response's coherence, comprehensiveness, and thematic relevancy.
- **Pronunciation:** Assesses the speaker's articulation and intelligibility by evaluating phonetic clarity and accuracy.
- **Fluency:** Assesses the speech's natural flow, taking into account disfluencies, rhythm, and pause patterns.

The corpus is appropriate for training and testing models of automatic speech evaluation based on its multi-aspect scoring method, which conforms to standard language testing practice. It is organized as an Excel file cross-referencing the scores to individual audio samples that are located in a designated directory. By facilitating generalization over a range of accents and speech patterns, the range of speakers adds to the robustness of the model.

3.2 Data Preprocessing and Feature Engineering

To obtain linguistic and auditory features necessary for the evaluation of content, fluency, and pronunciation, a two-stream preprocessing pipeline was implemented. Whisper, a state-of-the-art ASR model developed by OpenAI, was employed to produce the transcriptions because it was selected for its resistance to background noise, accent changes, and spontaneous speech.

Its ability to produce punctuated and structured text enabled the provision of accurate input for linguistic analysis later on. A BERT-based model was employed to process these transcriptions in an effort to evaluate grammatical correctness, lexical variety, and semantic coherence.

At the same time, phonetic and prosodic characteristics were modeled through the extraction of acoustic information. Spectrograms allowed easier extraction of spatial features using the FCNN, while MFCCs extracted spectral information related to articulation. Pitch and energy contours that extracted prosodic cues like intonation and rhythm were used to extract temporal fluency patterns and were passed to the BiLSTM model. After normalization of all features to unit variance and zero mean, input distributions were min-max scaled to standardize. Pitch shifting and speed perturbation were used as augmentation methods to promote generalization and minimize overfitting, particularly considering the limited amount of labeled data.By using this feature engineering, each model is guaranteed input optimized to its architecture: While BiLSTM and FCNN work upon complementing acoustic features, BERT works upon linguistically rich text, making for a coherent and comprehensible speech evaluation framework.

3.3 Proposed Methodology

This article illustrates an AI-driven speech grading system based on deep learning with a focus on content analysis, fluency, and pronunciation of spoken language. It features the application of the Whisper speech-to-text model integrated with a BERT semantic grading model.

3.3.1 BERT+Whisper Model

The core model consists of two integrated stages:

Transcription using the Whisper model and semantic evaluation using a BERT-based architecture.

3.3.1.1 Whisper-Based Transcription

The system processes raw audio input, which is transcribed by the Whisper-small model. Whisper, being a multilingual automatic speech recognition model, is most appropriate to handle noisy conditions and varying accents. The model produces a robust and accurate textual form of the audio input, which is the basis for subsequent judgment in the pipeline.

- A full transcription of the audio.

The transcription serves as the primary feature input for further content evaluation.

3.3.1.2 Whisper Architecture for Speech Recognition

The fig. 2 below illustrates the internal workflow of the Whisper model. The input audio is first transformed into a log-Mel spectrogram, capturing time-frequency patterns. This is processed by an encoder that learns abstract representations from the audio signal.

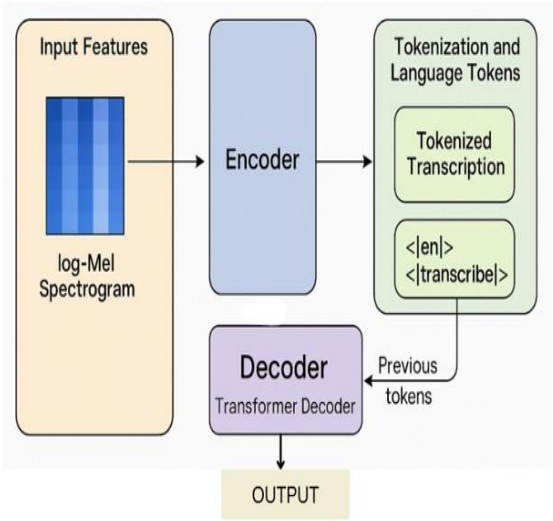

Fig. 2. Whisper Model Architecture.

Encoder output, language tokens, and task tokens (i.e., en|><|transcribe|>) are fed as input to a transformer decoder. The decoder generates the transcribed output autoregressively. Whisper is least affected by noise environments and by different speaker accents and is therefore best suited to real-world use in speech grading.

3.3.1.3 BERT-Based Semantic Evaluation

After the transcription process, the created text is refined with a pre-trained BERT model to judge the coherence and relevance of the spoken words. Using BERT's contextualized language representations, the model analyzes the inherent semantic structure and meaning of the answer. This enables the system to judge not only grammatical correctness but also logical flow and subject consistency of the spoken words. In order to combine acoustic and semantic knowledge, the contextual embedding for the [CLS] token for the aggregated sentence-level semantics is obtained. Optionally, an average confidence of the transcription is included to help capture any potential ASR variability. These entities are combined into a rich hybrid feature vector, which is then fed into a regression layer charged with the responsibility of predicting separate scores for fluency, pronunciation, and content quality.

3.3.1.4 BERT Architecture for Semantic Evaluation

The fig. 3 below depicts the BERT-based evaluation module. After transcription by Whisper, the text undergoes tokenization and is passed through a pre-trained BERT encoder to extract deep semantic embeddings.

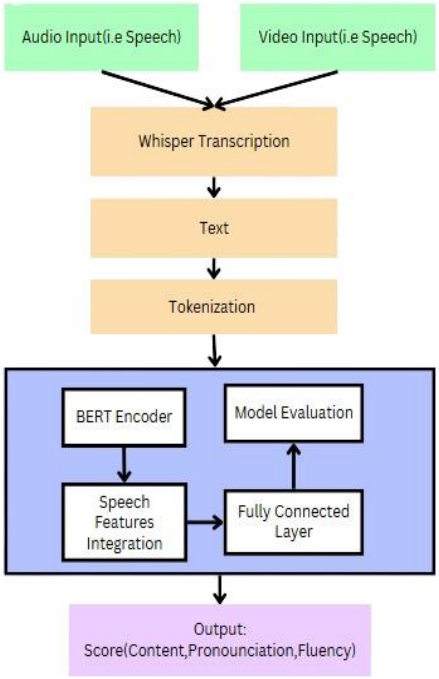

Fig. 3. BERT + Whisper Architecture.

The token output serves as a summary of the sentence which can reflect relevance, coherence, and grammaticality. This representation is combined with prosody speech features and then fed into a fully connected layer.

And the scores are predicted on three important aspects: Content, Pronunciation and Fluency. The model's hybrid architecture integrates acoustic and semantic features in an end-to-end manner and offers an efficient and robust solution to the automatic spoken language assessment.

3.3.2 Baseline Models for Speech Grading

We used two baseline models (FCNN and BiLSTM) to compare with the performance of the proposed Whisper + BERT. Bidirectional Long Short-Term Memory network (BiLSTM) model is used to model the two-way temporal coherence through Mel-Frequency Cepstral Coefficients (MFCCs) as the input. Phoneme, fluency and prosody are all well represented by the BiLSTM model with sequences in both directions. The continuous scores are obtained from the stacked BiLSTM layers, dense layers, and regression layers of the architecture.

But we could have used a lighter and faster model, the FCNN. It uses a few fully connected layers with ReLU activation to process fixed-length vector MFCC features. It is able to learn meaningful audio representations and produce fluency, pronunciation and content scores and those without preservation of temporal context. Baselines are established for the advantages of BERT in semantic evaluation and Whisper in transcription in the proposed system.

4 Results and Comparative Analysis

A comparative study of the three deep learning models applied to the task of automatic grading of speech; Bidirectional Long Short-Term Memory (BiLSTM), Fully Connected Neural Network (FCNN), and BERT model with Whisper Automatic ASR; appears next in this section. All models were evaluated using the content-specific, pronunciation, and fluency components. For a detailed study on how the method perform, we used classification accuracy and regression measures (Mean Absolute Error, Mean Squared Error).

4.1 Performance Evaluation Metrics

To assess model performance quantitatively, the following metrics were utilized:

- **Mean Absolute Error (MAE):** Reflects the average absolute difference between predicted and actual values.

$$MAE = (1/n) \sum |y_i - \hat{y}_i| \quad (1)$$

- **Mean Squared Error (MSE):** Emphasizes larger errors by squaring the difference between predicted and actual values.

$$MSE = (1/n) \sum (y_i - \hat{y}_i)^2 \quad (2)$$

- **Accuracy:** Adapted for regression by considering a prediction accurate if it falls within a specified threshold (δ) of the true score.

A comprehensive knowledge of each model's accuracy, consistency, and generalizability in actual grading situations is made possible by these criteria. The overall performance of each model across accuracy, mean squared error (MSE), and mean absolute error (MAE) is summarized in Table 1.

Table 1. Model Performance Metrics.

Model	Accuracy	MSE	MAE
BiLSTM	81.55	4827.49	11.77
FCNN	80.46	4709.66	19.54
BERT + Whisper	87.00	520.48	19.22

The Mean Absolute Error (MAE) and Mean Squared Error (MSE) were shown over training epochs in order to further examine the training dynamics of the suggested models. These charts demonstrate how well the corresponding models perform in terms of generalization and convergence.

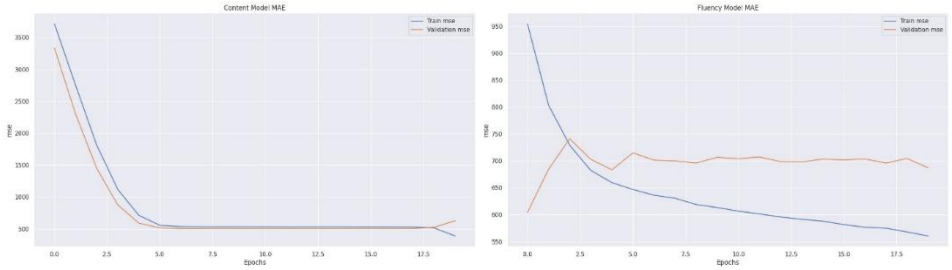

Fig. 4. BiLSTM: Training and validation MAE.

The BiLSTM model exhibits a consistent drop in MAE, as seen in Fig 4, with little difference between the training and validation curves, suggesting efficient learning and little overfitting.

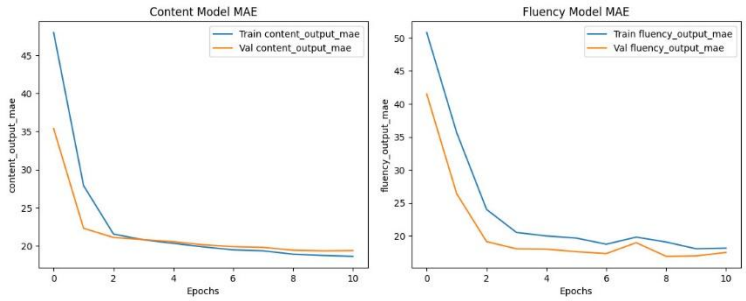

Fig. 5. BERT: Training and validation MAE.

Fig 5 demonstrates a consistent reduction in MAE for the BERT model, suggesting improved prediction accuracy over time.

4.2 Comparative Analysis

A thorough evaluation of the model performances is presented in Table 1 and Table 2, highlighting their respective strengths and limitations across classification and regression tasks.

The BiLSTM model achieved a classification accuracy of 81.55% and the lowest Mean Absolute Error (MAE) of 11.77, indicating accurate individual predictions. However, its Mean Squared Error (MSE) was significantly higher at 4827.49, particularly impacted by the Pronunciation component. This discrepancy, along with insights from training curves (see Fig 4), reveals a tendency toward overfitting despite favorable MAE values limiting the model's generalization capacity.

In contrast, the FCNN model achieved a slightly lower accuracy of 80.46% and the highest MAE at 19.54. Despite this, it showed marginally better MSE performance (4709.66) compared to BiLSTM and maintained a low risk of overfitting. Its lightweight and straightforward architecture makes it suitable for low-resource environments, albeit at the cost of capturing temporal or contextual information critical in speech grading.

The BERT-Whisper model delivered the best overall performance, achieving the highest classification accuracy (87.00%) and the lowest MSE (520.48), indicating superior stability and generalization. Although its MAE (19.22) was slightly higher than that of BiLSTM, the model consistently demonstrated lower error variance across training and validation sets, as shown in Fig 5. This affirms the model's robustness, largely due to its transformer-based contextual understanding and high-quality transcription from Whisper.

Table 2. Model Characteristics Comparison.

Model	Overfitting Risk	Computational Cost	Reliability
BiLSTM	High	Medium–High	Moderate
FCNN	Low	Low	Limited
BERT + Whisper	Low	High	High

While the BiLSTM model achieves strong MAE performance, its susceptibility to overfitting and high MSE undermines its reliability. FCNN offers simplicity and resource efficiency but lacks the representational depth for complex speech assessment. The BERT-Whisper model, despite a slightly higher MAE, delivers the most balanced and generalizable performance, making it the preferred choice for precision-driven and scalable speech grading systems.

5 Discussion and Key Observations

The Whisper-BERT approach presented above suggests a unified approach towards automatic speech evaluation through the combination of accurate transcription and deep contextual comprehension. Unlike previous work such as Ramacharla et al. [1] and Ziman et al. [3], which was ASR-based, our approach consists of semantic and acoustic feature-based evaluation, allowing for multi-dimensional evaluation. While Pratama and Amrullah [7] evaluated Whisper for low-resource ASR, they did not utilize it for content evaluation. Our use of BERT addresses this by enabling contextual speech comprehension, an advancement over sentiment- or keyword-based evaluation by Suresh et al. [10] and Sadanand et al. [13].

BiLSTM and FCNN baselines, inspired by previous acoustic modeling [6], validate the strength of our design. BiLSTM worked extremely well in examining pronunciation patterns, and FCNN introduced efficiency at the cost of diminished temporal awareness. Whisper-BERT tends to work better than traditional ASR or NLP-only pipeline [14], and offers an interpretable and scalable solution to speech grading.

6 Conclusion

BiLSTM, FCNN, and the novel BERT-Whisper model were tested for automated fluency, pronunciation, and content grading of speech. Although BiLSTM is able to learn bidirectional temporal features, it is plagued with generalization and increased computing overhead. Although FCNN was economical, it didn't have sequential modeling capability essential for a valid evaluation. Conversely, the BERT-Whisper architecture combined contextual embeddings from BERT with high-quality Whisper ASR transcripts to provide high accuracy and reliability in speech scoring tasks. The findings show how novel and cost-effective it is to combine pretrained ASR with language models to provide an integrated speech evaluation. The approach makes scalable, multilingual, and intelligent speech evaluation systems feasible.

References

[1] R. Adolph, “Speech to Text Transcription Sai,” Int. J. Res. Appl. Sci. Eng. Technol., vol. 12, no. April, pp. 1–23, 2016.

[2] L. Kobylyukh, Z. Rybchak, and O. Basystiuk, “Analyzing the Accuracy of Speech-to-Text APIs in Transcribing the Ukrainian Language,” CEUR Workshop Proc., vol. 3396, pp. 217–227, 2023.

[3] K. Ziman, A. C. Heusser, P. C. Fitzpatrick, C. E. Field, and J. R. Manning, “Is automatic speech-to-text transcription ready for use in psychological experiments?,” Behav. Res. Methods, vol. 50, no. 6, pp. 2597–2605, 2018, doi: 10.3758/s13428-018-1037-4.

[4] B. Chheda and M. Patil, “Text Recognition and Extraction from Video Kiran Agre Sairaj Gaonkar,” vol. 5, no. 01, pp. 1–4, 2017, [Online]. Available: www.ijert.org

[5] M. Saraswathi et al., “International Journal of Research Publication and Reviews Implementation of Video and Audio to Text Converter,” vol. 4, no. 5, pp. 1204–1208, 2023.

[6] C. Dhale, R. Gajare, A. Babar, G. Pise, and P. R. Yevale, “Transcribe Audio / Video Into Text Using Deep Learning,” vol. 12, no. 4, pp. 187–197, 2024.

[7] R. S. A. Pratama and A. Amrullah, “Analysis of Whisper Automatic Speech Recognition Performance on Low Resource Language,” J. Pilar Nusa Mandiri, vol. 20, no. 1, pp. 1–8, 2024, doi: 10.33480/pilar.v20i1.4633.

[8] V. Aswin, M. Javed, P. Parihar, K. Aswanth, C. Druval, A. Dagar, and A. C. V. Aravinda,(2021). "NLP-Driven Ensemble-Based Automatic Subtitle Generation and Semantic Video Summarization Technique", Springer, 2021, doi: 10.1007/978-981-15-3514-7_1.

[9] R. Kavitha, N. Nachammai, R. Ranjani, and J. Shifali, “Speech Based Voice Recognition System for Natural Language Processing,” vol. 5, no. 4, pp. 5301–5305, 2014.

[10] V. Suresh, R. Agasthiya, J. Ajay, A. A. Gold and D. Chandru, "AI based Automated Essay Grading System using NLP," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023, pp. 547-552, doi: 10.1109/ICICCS56967.2023.10142822.

[11] Y. Xu, “English Speech Recognition and Evaluation of Pronunciation Quality Using Deep Learning,” Mob. Inf. Syst., vol. 2022, 2022, doi: 10.1155/2022/7186375.

[12] D. M. Jones, L. Cheng, and M. G. Tweedie, “Automated Scoring of Speaking and Writing: Starting to Hit its Stride,” Can. J. Learn. Technol., vol. 48, no. 3, 2022, doi: 10.21432/cjlt28241.

[13] V. S. Sadanand, K. R. R. Guruvyas, P. P. Patil, J. J. Acharya, and S. G. Suryakanth, “An automated essay evaluation system using natural language processing and sentiment analysis,” Int. J. Electr. Comput. Eng., vol. 12, no. 6, pp. 6585–6593, 2022, doi: 10.11591/ijece.v12i6.pp6585-6593.

[14] H. Ghanta, “Automated Essay Evaluation Using Natural Language Processing and Machine Learning,” Theses Diss., pp. 1–46, 2019, [Online]. Available: https://csuepress.columbusstate.edu/theses dissertations/327/.

[15] https://www.kaggle.com/datasets/kartikeybartwal/speech-content-fluency-and-pronunciation-scores.

[16] **16.** G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012, doi: 10.1109/MSP.2012.2205597.