Global Scientific JOURNALS

GSJ: Volume 13, Issue 8, August 2025, Online: ISSN 2320-9186
www.globalscientificjournal.com

SPEECH TO TEXT AUDIO LECTURE TRANSCRIPTION USING DEEP LEARNING

Ogbeide Oluwatosin Lara

Ogbeide Oluwatosin Lara ,Ph.D student in Computer Science Federal University of Technology, Akure,Nigeria
E-mail: oluwatosin.ogbeide@aaua.edu.ng

KeyWords
Accents, deep learning, impairments, pre-trained, speech recognition, transcription*, *transformer

ABSTRACT

This research work focuses on the development of a robust system for transcribing audio lectures into text using advanced deep learning techniques. This research work addresses this need by leveraging state-of-the-art of neural networks to convert spoken language into written text, facilitating better access to educational resources for all learners, including those with hearing impairments. The system is built upon a deep learning framework Whisper, the pre-trained audio-to-text transcription model developed by OpenAI. It utilizes the transformer architecture which are particularly effective for sequence-to-sequence tasks like speech recognition due to their ability to handling range dependencies and parallel processing. The model is trained on large datasets of audio lectures, enabling it to learn and generalize across various accents, speaking styles, and academic terminologies. To enhance transcription accuracy, the system incorporates techniques like data augmentation, noise reduction, and attention mechanisms. Evaluation of the system demonstrates high accuracy in transcribing lectures, making it a valuable tool for educational institutions and platforms that offer online courses.

INTRODUCTION

Education is a cornerstone of societal development, enabling individuals to acquire knowledge, skills, and perspectives necessary for personal growth and societal progress. However, this essential facet of human development is not universally accessible. Among the marginalized groups are individuals with hearing impairments, specifically deaf and dumb students, who face considerable challenges in engaging with traditional educational methods that predominantly rely on auditory communication. This backdrop underscores the urgency of exploring innovative approaches that facilitate their educational inclusion.

Often student with no hearing impairments also tend to depend more on the note taken during the lecturer's explanation to enable them understand the topic being taught with the help of any other course material given by the lecturer. Having the lecturers' explanation given in text format will be an added advantage for students in their study period. Automated speech to text transcription of lecture given by the lecturers could help the students study and understand the topic being taught.

Speech2Text is a speech model that accepts a float tensor of log-mel filter-bank features extracted from the speech signal. It's a transformer-based seq2seq model, so the transcripts are generated autoregressively. In recent years,

there has been significant progress in speech translation technology driven by advances in deep learning and spoken language processing[24]. With the development of machine learning and deep learning algorithms, automated voice recognition has become a major study area using pre-trained language models, like Hugging Face, for fine-tuning is one such method [17], Whisper and HuBERT to mention a few.

For the purpose of this research, Whisper pre-trained audio-to-text transcription model developed by OpenAI is utilized. It is based on the Transformer architecture. Transformers are particularly effective for sequence-to-sequence tasks like speech recognition due to their ability to handle long-range dependencies and parallel processing. Whisper uses the Transformer to convert audio waveforms into text, leveraging attention mechanisms to capture the nuances of speech across different languages and accents. This architecture allows Whisper to achieve high accuracy in transcription tasks.

Related works

This section provides an overview of existing research and projects related to the development and implementation of real-time speech transcription systems for individuals with hearing and speech impairments, with a specific focus on deaf and dumb students. The review aims to contextualize the current project within the broader landscape of assistive technologies and educational accessibility.

Several traditional methods have been adopted speech transcription for ease of communication with deaf and dumb students, For example, [2] developed Speech Recognition Application for the Speech Impaired people using the Android-based Google Cloud Speech API. Using the Google Cloud Speech Application Programming Interface (API), this allows converting audio to text, and it is user-friendly to use such APIs. The Google Cloud Speech API integrates with Google Cloud Storage for data storage. Although research into speech recognition to text has been widely practiced, this research try to develop speech recognition, specially for speech impaired's speech, as well as perform a likelihood calculation to see the factor of tone, pronunciation, and speech speed in speech recognition.

[11] developed a model for automatic transcription of lecture speech using topic-independent language and modeling with a vocabulary selection mechanism based on a mutual information criterion. The developed baseline model was adapted to specific lectures using preprint texts.

Several scholars have also delved into the application of pre-trained model in speech to text transcription. [16] investigated efficient strategies to build cascaded and end-to-end speech translation systems based on pre-trained models. Using this strategy, we can train and apply the models on a single GPU. While the end-to-end models show superior translation performance to cascaded ones, the application of this technology has a limitation on the need for additional end-to-end training data. The authors developed an additional similarity loss to encourage the model to generate similar hidden representations for speech and transcript. It is challenging for media officers, secretaries or attendees to listen and take written accounts of what is said at the same time. [18] developed a speech-to-text conversion model that can help media officers overcome these challenges using full verbatim and clean verbatim. This model has ability to transcribe the recorded audio and then convert it to texts.

This research seeks to bridge this gap by designing and implementing an automated speech transcription system uniquely adapted to the needs of students to allow access for students to revisit the lecture which have been taught earlier by the lecturer. The goal is to empower these students with the ability to access spoken content of lecturers in class, thereby enhancing their engagement, participation, and academic success. The system will address challenges related to speech recognition accuracy, varying accents, classroom acoustics, and technical implementation.

Overview of the Speech2text Audio Transcription Model

Install the necessary libraries: The transformers library and the torch library are installed, this is done using pip install.

The pre trained model and tokenizer are loaded.

This is followed by the preprocessing the raw audio files**.** The audio data from lectures must be preprocessed before it can be fed into the model. This includes loading the audio file, resampling, and normalizing it. The model takes in the raw audio file and the files are preprocessed. the audio sample rate is set to 16,000 Hz. The processed audio is then saved to a new file with processed added to the name and the path is return to processed audio file. The Speech to text transcription model.

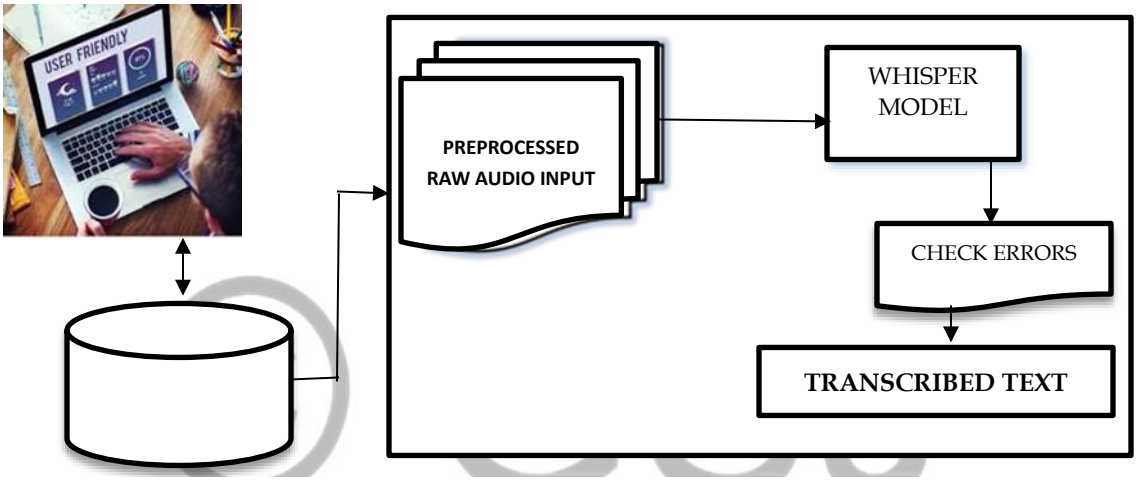

Figure 1: Archictecture of the speech to text audio lecture transcription

This encoder is critical for the processing of raw audio data and converting it into a form that the subsequent transformer-based model can handle effectively. The pretrained model is composed of a multi-layer convolutional feature encoder $f : X \rightarrow Z$ which takes as input raw audio X. The convolutional layer extract low level audio features from the raw waveform. This involves detecting basic patterns like phonemes, tones and other audio signals. The encoder typically consists of multiple convolutional layers. Each layer progressively extracts more complex features from the audio signal. The initial layer focus on capturing local patterns, such a pitch and short-term frequency components.

As the audio file moves through the layers, the receptive fields (the area of the input that the layer can see) expands enabling the model to capture broader patterns and contextual information.

The encoder consists of several blocks containing a temporal convolution followed by layer normalization and a GELU activation function. The raw waveform input to the encoder is normalized to zero mean and unit variance. After processing by the convolutional layers, the audio features are transformed into a sequence of feature vectors. The sequence is much shorter than the original waveform, making it easier for the transformer to process. The total stride of the encoder determines the number of time-steps T which are input to the Transformer

THE TRANSFORMER MODEL ARCHICTECTURE

The transformer consists of multiple layers that apply self-attention mechanisms to weigh the importance of different parts of the input sequence relative to each other, allowing the model to understand the context and relationships across the entire sequence. The outputs latent speech representations $z_1, \ldots, z_T$ for T time-steps. They are then fed to a Transformer $g : Z \rightarrow C$ to build representations $c_1, \ldots, c_T$ capturing information from the entire sequence. The output of the transformer is a sequence of tokens(text) that represent the transcription of the input audio. The final step involves converting the sequence of tokens into human readable text. This typically involves techniques like beam search to find the most probable transcription given the learned language model.

This process involves the combining of convolutional feature extraction with transformer-based sequence processing. This is what enables Whisper to perform well on various speech recognition tasks especially in handling complex and noisy audio data.Once the audio lecture file is preprocessed, you can pass it through the model to get the transcription. The audio file was loaded on the Whisper Library

USER INTERFACE DESIGN

The user interface (UI) design of the automated real-time speech transcription system is meticulously crafted to ensure accessibility, intuitiveness, and an overall positive user experience. This section provides an in-depth exploration of the key features incorporated into the UI.

LOGIN PAGE

The login page serves as the initial interaction point for students and provides a secure gateway to access the system's features. It incorporates essential elements such as:

i. Username and Password Fields: Where the users will input their unique credentials (username and password) to gain access. This ensures authentication and restricts unauthorized entry.

Figure 1: Login page

HOMEPAGE

Once successfully logged in, users are directed to the student dashboard. On this page the important information about the lecture audio file like Course title, course code and lecturer name are submitted for storage in the database.The audio file upload of the recorded lecture will be done on this page. Prominent call-to-action buttons for upload is done enabling users to upload audio or video files.

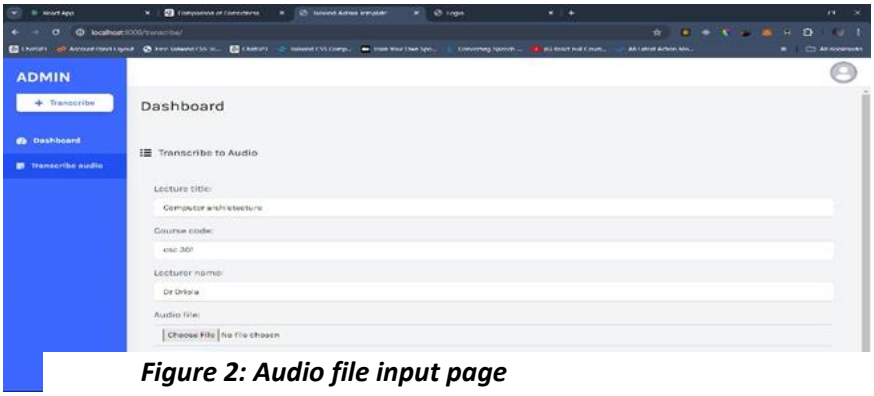

Figure 2: Audio file input page

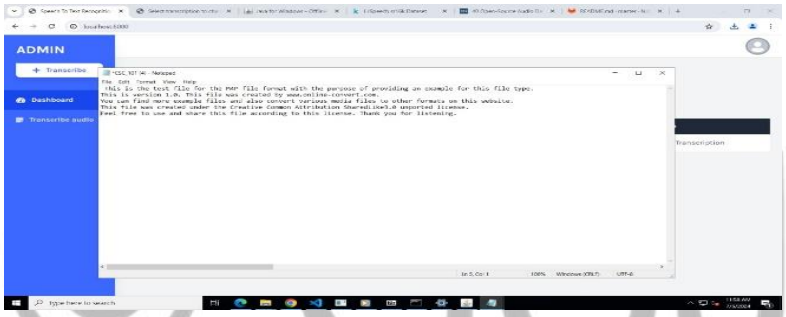

Figure 3: Transcription Results page

Transcription results: the generated text is presented clearly on this page, allowing users to review and edit.

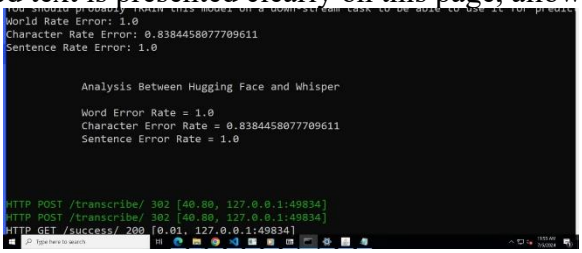

Figure 4: Error rate of the speech to text lecture transcription model

Conclusion

This project not only contributes to the field of speech recognition but also paves the way for future innovations in making educational content more inclusive and widely accessible.

Acknowledgment

The author wish to thank everyone that contributed to the success of this research work.

References

[1] D.Amodei,, C.Olah,, J., Steinhardt, V. Christiano, V., Schulman, D. Mané, (2016). A survey of machine learning for speech recognition. *AI Magazine,* 37(4), 93-107.
[2] N., Anggraini, A., Kurniawan, L., Wardhani & N., Hakiem (2018). Speech Recgnition Application for the Speech Impaired using the Android-based Google Cloud Speech API,TELKOMNIKA(Telecomunication Computing Electronics and Control),16(6)1693-6930
[3] N., Chomsky,(1957). *Syntactic structures.* The Hague: Mouton.
[4] F., García , F., Pérez-Lancho, , B. González-Díaz, L., Casillas (2019). Towards real-time translation systems for educational contexts. *Machine Translation,* 33(3), 428-462.
[5] A., Graves,, A. Mohamed, G. Hinton, (2013). Speech recognition with deep recurrent neural networks. In *NIPS deep learning and unsupervised feature learning workshop* (pp. 1-10).
[6] G., Hinton, L.,Deng, D.,Yu,, G.,Dahl, A., Mohamed, N. Jaitly & B., Kingsbury, (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE signal processing magazine,* 29(6), 82-97.
[7] D., Intan, S., Saputra, S., Handani, & V.,Diniary, (2017). Utilization of Cloud Speech API for the Development of English Language Learning Media using Speech Recognition Technology. TELEMATIKA; 10(2): 92–105.
[8] A., James, D., Thomas, & J., Newton, (2020). Enhancing inclusivity through technological solutions for deaf students. *Education and Technology*, 18(5), 52-66.
[9] S., Johnson, & T., Brown, (2021). "I felt more confident speaking out loud in class": a critical examination of speech recognition and its impact on the learning experiences of deaf college students. *American Annals of the Deaf*, 166(2), 107-123.
[10] S., Kawas, M., Dennis, & Z., Liu, (2020). Investigating the impact of automated transcription systems on deaf students' learning experiences. *Computer Assisted Language Learning,* 33(3), 289-307.
[11] K., Kazuomi, N., Hiroaki & K., Tatsuya,(2020). Automatic Transcription of Lecture Speech using Topic-Independent language modeling,International Conference on Spoken Language Processing INTERSPEECH
[12] A., Baevski, S. Schneider, & M. Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. In Proc. of ICLR, 2020.
[13] R., Kheir, D., Galbraith, T., Krakowiak, M., Gonzalez, A., Westerveld, & S., Alsubaie,,(2018). Improving speech recognition to assist real-time classroom note-taking. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 209-212).
[14] , Y., LeCun, Y.,Bengio, G., & Hinton, (2015). Deep learning. *Nature*, 521(7553), 436-444.
[15] R. Liyanagunawardena, H., Adams, & J. Williams, (2019). A comparison of automatic transcription software tools for closed captions of video lectures for students with hearing disabilities. *Journal of Postsecondary Education and Disability*, 32(1).
[16] Z., Li, & J., Niehues(2022). Efficient Speech Translation with Pre-trained Models, Conference on Neural Information Processing Systems (NeurIPS 2022)
[17] M. Salleh, S., Zulaiha, & N., Anuar,(2020). Developing sign language recognition system for educational setting. In *International Conference on Education Technology and Computer* (pp. 600-604). Springer, Singapore.
[18] O., Isiaka, A., Ibraheem, ,D., Bolaji-Adetoro & T., Saka,(2023) Speech-to-text conversion for effective audio transcription in mass media industry operations, Fedpolad Journal Of Management, 3(1),2786-9644
[19] H., Jegou, M. Douze, & C. Schmid.(2011) Product quantization for nearest neighbor search. IEEE Trans. Pattern Anal. Mach. Intell., 33(1):117–128.

[20] N., Papastratis, D., Tang, S.,Dhakal, J., Taylor, B., Janssen, & Walter, C. (2021). Real-time speech transcription in educational settings: A scoping review. *Educational Technology & Society*, 24(3), 62-88.
[21] P., Pandey, S., Awasthi, & S., Goyal,(2020). Academic performance and challenges faced by deaf and dumb students in the context of Kashmir. *Social Sciences and Humanities Letters*, 8(4), 164-173.
[22] R., Rabiner. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257-286.
[23] N., Rao, M., Sudeep, & S., Kariyam, Thota, A., Chauhan, D., Reddy, R. (2020). Real-time multilingual speech-to-text translation system for assistive technologies. *IETE Journal of Research*, 66(1), 27-34.
[24] N.,Yuta & N., Satoshi.(2023). Inter-connection: Effetive Connection between Pre-trained encoder and Decoder for Speech Translation, INTERSPEECH 20-24