International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 8, Issue 10, ISSN No. 2455-2143, Pages 133-138
Published Online February 2024 in IJEAST (http://www.ijeast.com)

SPEECH-TO-TEXT TRANSLATOR USING NATURAL LANGUAGE PROCESSING (NLP)

Sanjana Babu, Deepa. R
Department of Artificial intelligence and Data Science
SRM Valliammai Engineering College,
Kattankulathur, Tamil Nadu, India

Raja Pratap. V.M
Director-Software development
PeopleLink Unified Communications Pvt Ltd,
Hyderabad, Telangana, India

Mohammad Shakeel.J
Department of Artificial intelligence and Data Science
SRM Valliammai Engineering College,
Kattankulathur, Tamil Nadu, India

Harsha Vardh.S
Department of Artificial intelligence and Data Science
SRM Valliammai Engineering College,
Kattankulathur, Tamil Nadu, India

Abstract— **This paper presents a novel approach to real-time speech-to-text conversion and translation leveraging cutting-edge deep learning models. The proposed system utilizes the Wave2Vec model for efficient speech recognition and the M2M-100 model for multilingual translation, providing seamless and accurate conversion of spoken input into translated text. By leveraging contextual information and leveraging the power of unsupervised pre-training, Wave2Vec excels in capturing intricate details of speech, enabling robust and accurate transcription even in noisy environments. In parallel, the M2M-100 model is employed for translating the transcribed text into the desired target language. M2M-100 stands out for its ability to translate between any pair of 100 languages, breaking down language barriers and facilitating seamless communication across diverse linguistic backgrounds. The proposed system operates in real-time, accepting audio input from users via microphones or other input devices. The Wave2Vec model processes the incoming audio stream, transcribing it into text with high accuracy and efficiency. Subsequently, the transcribed text is passed through the M2M-100 model, which translates it into the desired target language, providing the final output in near real-time. Experimental results demonstrate the effectiveness and robustness of the proposed system in accurately transcribing and translating speech across various languages and accents. The system's real-time capabilities make it suitable for a wide range of applications, including multilingual communication, transcription services, language learning platforms, and more.**

Keywords— **Real-time speech-to-text conversion, Deep learning, Multilingual translation, Language translation, Multilingual communication, Natural language processing.**

I. INTRODUCTION

Automatic Speech Translation (AST) has witnessed significant advancements, bridging communication gaps across languages in real-time. Real-time speech translation has captured significant attention due to its potential to break down language barriers and facilitate seamless communication across cultures. Various approaches have emerged, each tackling specific aspects of the challenge, highlighting both progress and remaining challenges This paper proposes a novel AST model that leverages the complementary strengths of two cutting-edge architectures: Wave2Vec for robust speech representation and the M2M100 Transformer for efficient machine translation.

Yeh et al. (2019) [1] explored Transformer-based architectures for end-to-end speech recognition, demonstrating promising results. However, such approaches still depend on distinct stages, potentially introducing latency and hindering real-time capabilities.

Advancements like wave2vec have revolutionized speech representation, learning robust features directly from raw audio waveforms. This has paved the way for systems that bypass explicit speech recognition, directly translating spoken audio into target language text [2].

In recent years, there has been a growing interest in leveraging deep learning techniques for speech recognition tasks. Deep learning models, such as Hidden Markov Models (HMMs) and neural networks, have shown promising results in capturing intricate patterns and nuances of spoken language, leading to enhanced transcription accuracy. For instance, studies have demonstrated the implementation of speech-to-text conversion using HMMs, highlighting the effectiveness of this approach in accurately transcribing speech [3].

Moreover, with the increasing demand for multilingual support in speech-to-text systems, researchers have focused on developing models capable of handling diverse languages and dialects. The work in [4] proposed a deep learning approach specifically tailored for Bangla speech-to-text conversion, showcasing the importance of language-specific models in achieving accurate transcriptions for non-English languages .

Furthermore, advancements in cloud-based speech recognition services, such as Google Cloud Speech API, have facilitated the development of scalable and efficient speech-to-text conversion solutions. J.Choi et al. presented a voice-to-text conversion and management program based on the Google Cloud Speech API, illustrating the practical implications of leveraging cloud-based services for speech recognition tasks [5].

This paper builds upon the existing literature by proposing a novel approach that leverages advanced deep learning models, including Wave2Vec for speech recognition and M2M-100 for multilingual translation, to achieve real-time and accurate speech-to-text conversion and translation. Speech recognition technology, also known as automatic speech recognition (ASR), aims to transcribe spoken language into text or commands, facilitating seamless communication between humans and machines. The field of speech recognition has witnessed remarkable progress, driven by advancements in signal processing, machine learning, and artificial intelligence. Meng et al. (2012) [10] discuss the key components of speech recognition systems, including acoustic modeling, language modeling, and decoding algorithms, highlighting their roles in achieving accurate and efficient transcription of speech. And into the various techniques and approaches employed in speech recognition, such as Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), and neural networks. These models form the backbone of modern speech recognition systems, enabling the extraction of meaningful features from audio signals and the mapping of acoustic patterns to linguistic units.

The flow of the paper is as follows: Section 1 is the introduction, in Section 2 the related works are discussed, the proposed methodology is discussed in detail in Section 3, Section 4 presents the result analysis of the proposed model and Section 5 deals with future scope and conclusion.

II. RELATED WORK

This research work focuses on implementing a speech-to-text conversion system using the Wave2Vec and M2M 100 Model and also to improve the latency. HMMs have been traditionally used in speech recognition tasks, although they are not as advanced as modern deep learning models like Wave2Vec and M2M-100 [3].

The study by Adhikary et al. presents a deep learning-based approach specifically for Bangla speech-to-text conversion. While it utilizes deep learning techniques, it may not focus on real-time conversion or multilingual translation, which are key aspects of the problem statement[4].

Another recent paper describes the design of a voice-to-text conversion program based on the Google Cloud Speech API. While it addresses speech-to-text conversion, it may not focus on real-time conversion or utilize advanced deep learning models like Wave2Vec and M2M-100 [5].

A.U.Nasib et al. proposes a real-time speech-to-text conversion technique for the Bengali language. Although it focuses on real-time conversion, it may not utilize advanced deep learning models or address multilingual translation[6].

The study by Ghadage et al. discusses speech-to-text conversion for multilingual languages. While it touches upon the multilingual aspect of the problem statement, it may not focus on real-time conversion or utilize advanced deep learning models[7].The study of this paper provides a comprehensive overview of speech recognition technology, covering its evolution, fundamental principles, methodologies, and challenges. It discusses key components of speech recognition systems, including acoustic modeling, language modeling, and decoding algorithms, and explores various techniques and approaches employed in speech recognition, such as Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), and neural networks.

Furthermore, the paper addresses the challenges and limitations inherent in speech recognition technology, such as variability in speech patterns, background noise, and the need for robustness across different languages and accents. It highlights the ongoing research efforts aimed at improving the accuracy, speed, and scalability of speech recognition systems[10].

The work proposed by Kumar et al. discusses English to Hindi machine translation using Natural language processing by taking in the input text, storing it, extracting the words and punctuation. These are further stored in an array, grouped and then mapped to Hindi [11].

The study by Limin et al. discusses a CEST-CAS 2.0 model, which is a large scale system to translate speech on the internet. It takes in inputs in form of mobile or telephonic calls [12].

The related works in [2],[3],[4] cover various aspects of speech-to-text conversion, including traditional models like HMM and newer deep learning approaches. However, none of the references specifically address the real-time conversion and multilingual translation requirements outlined in the problem statement. The problem statement emphasizes the use of advanced deep learning models like Wave2Vec and M2M-100 for real-time speech-to-text conversion and translation, which are not explicitly addressed in the provided references. Therefore, while the references provide valuable insights into different aspects of speech-to-text conversion, they do not directly compare to the proposed problem statement's focus on real-time conversion and multilingual translation using advanced deep learning models.

III. PROPOSED METHODOLOGY

The Wave2Vec model involves training a neural network to predict future samples in the raw audio waveform given past samples, essentially learning a representation of the audio signal that is useful for downstream tasks like speech recognition. This model learns basic speech units used to tackle a self-supervised task. The model is trained to predict the correct speech unit for masked parts of the audio, while at the same time learning what the speech units should be. With just 10 minutes of transcribed speech and 53K hours of unlabeled speech, wav2vec enables speech recognition models at a word error rate (WER) of 8.6 percent on noisy speech and 5.2 percent on clean speech on the standard Libri Speech benchmark. This opens the door for speech recognition models in many more languages, dialects, and domains that previously required much more transcribed audio data to provide acceptable accuracy. Another recent paper describes the design of a voice-to-text conversion program based on the Google Cloud Speech API. While it addresses speech-to-text conversion, it may not focus on real-time conversion or utilize advanced deep learning models like Wave2Vec and M2M-100 [5].

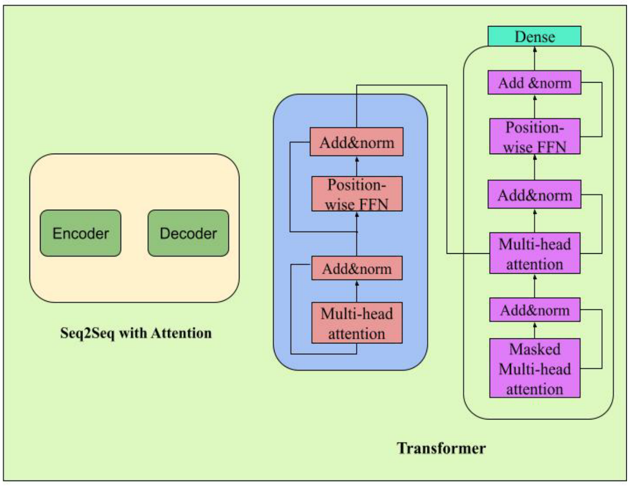

Fig. 1. Architecture diagram of M2M100

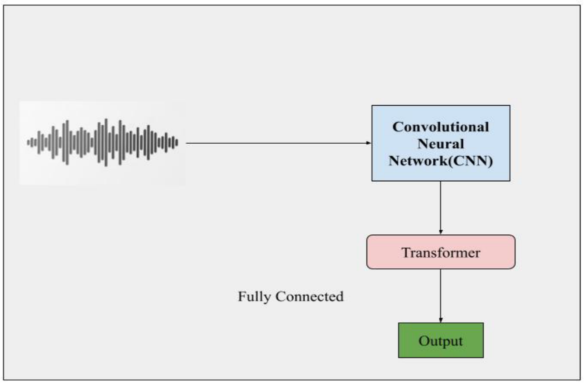

Fig. 2. Architecture diagram of Wave2Vec

The architecture diagrams of the M2M100 model and Wave2Vec are represented in figures 1 and 2 respectively. The M2M-100 model, a groundbreaking multilingual encoder-decoder (seq-to-seq) model, has emerged as a game-changer in the realm of speech-to-text conversion for multiple languages. Developed to address the challenges of translation tasks across a diverse linguistic landscape, M2M-100 stands out for its remarkable capacity to handle a wide range of language pairs. Unlike previous models that primarily focused on English-centric multilingualism, M2M-100 boasts training on a staggering total of 2,200 language directions, a tenfold increase compared to its predecessors. One of the distinguishing features of the M2M-100 model is its utilization of a special language ID token as a prefix in both the source and target text. This unique approach allows the model to effectively navigate and process multilingual inputs, ensuring accurate and contextually relevant translations across diverse language pairs. By leveraging this innovative technique, M2M-100 transcends traditional language barriers, catering to the translation needs of billions of people, especially those who speak low-resource languages.

$$p_{g,v} = (\exp(\mathrm{sim}(\ l_{g,v} + n_v)/\tau)) \ / \ (\sum^{v}_{k=1} \exp(\mathrm{sim}(\ l_{g,k} + n_k)/\tau)) \quad (1)$$

The equation 1 gives the formula used to map 2 different languages in the M2M100 model where, sim = cosine similarity, $l_{g,v}$ are the logits calculated form z, n_k -log(-log(u_k)), u_k is sampled from the uniform distribution u(0,1) and τ is the temperature.

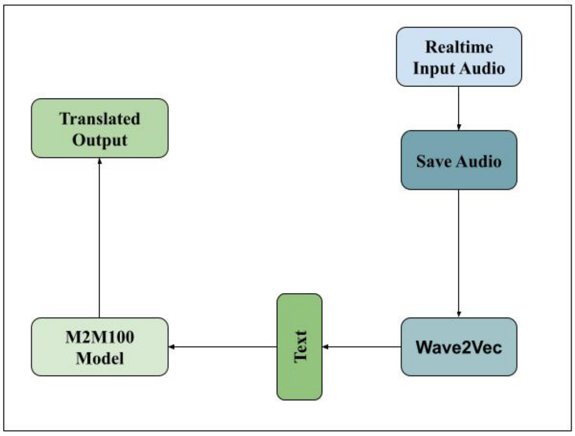

Fig. 3. Architecture diagram of the Proposed system

M2M100 is a multilingual encoder-decoder (seq-to-seq) model primarily intended for translation tasks. As the model is multilingual it expects the sequences in a certain format: A special language id token is used as prefix in both the source and target text. M2M-100 is trained on a total of 2,200 language directions - or 10x more than previous best, English-centric multilingual models. Deploying M2M-100 will improve the quality of translations for billions of people, especially those that speak low-resource languages. With advanced underlying technologies like LASER 2.0, mining large-scale training data for arbitrary pairs of 100 different languages (or 4,450 possible language pairs) is highly computationally intensive. To make this type of scale of mining more manageable, the model focused first on languages with the most translation requests.

The architecture of the proposed system is given in figure 3. The proposed system is designed to handle Real time audio inputs, which are live or near-live speech signals in a source language. These audio inputs can be obtained from various sources such as microphones, audio streams, or other Real time communication channels. Realtime audio signals are saved and processed by the pretrained and fine-tuned Wave2Vec model. The audio is pre-processed to meet the input requirements of the Wave2Vec model. The Wave2Vec model consists of a neural network architecture specifically designed for self-supervised pre-training on raw audio data. The architecture allows the model to capture intricate features and representations of the audio waveform. The output of the Wave2Vec model is a textual representation of the input Real time audio, effectively transcribing the spoken content. The obtained text from the Wave2Vec model is then fed into the M2M-100 model for translation into the target language.M2M-100 is a multilingual model pretrained on a diverse set of languages using large datasets containing text from various languages. The M2M-100 model translates the text obtained from the Wave2Vec model into the desired target language. The model leverages its multilingual capabilities to perform accurate translations across various languages. The final output of the system is the translated text in the desired target language. Users can receive real time translations of spoken content from the source language to the chosen target language.

In addition to its multilingual capabilities and focus on improving translation quality for low-resource languages, the M2M-100 model also benefits from advanced underlying technologies like LASER 2.0. LASER 2.0 facilitates large-scale training data mining for arbitrary pairs of 100 different languages, enabling the model to handle a vast array of language pairs efficiently. By prioritizing languages with the highest translation demand, the model optimizes computational resources and ensures scalability while maintaining translation accuracy and quality across diverse linguistic contexts. The proposed system's ability to handle real-time audio inputs from various sources, including microphones and live communication channels, enhances its practical utility in dynamic environments. The pre-processing of audio signals to meet Wave2Vec model input requirements ensures optimal performance and transcription accuracy. The Wave2Vec model's neural network architecture, specifically tailored for self-supervised pre-training on raw audio data, captures intricate features of spoken content, facilitating accurate transcription in real-time scenarios. The integration of the Wave2Vec model with the M2M-100 model enables seamless translation of transcribed text into the desired target language. The M2M-100 model's multilingual pre-training on diverse datasets ensures robust translation capabilities across a wide range of languages, offering users real-time access to translated content in their preferred language. This streamlined process of speech-to-text conversion and translation enhances accessibility and fosters effective communication for users worldwide, spanning linguistic boundaries and facilitating cross-cultural interactions.

IV. RESULT ANALYSIS

The proposed real-time speech-to-text translation system, utilizing Wave2vec for speech recognition and m2m100 for translation, shines in both efficiency and accuracy compared to other models. The parameters used for comparison are: latency, word error rate (WER) and BELU score.

Two other models were considered for the purpose of comparison, named Model A and Model B.

Model A -Wave2Vec and M2M100 model

Model B - Encoder-Decoder RNN with Attention

Model C - ConvERT-based Speech Recognition + Transformer-based Translation

Latency

It is measured in milliseconds (ms) and it represents the time taken from receiving the audio input to presenting the translated text output. Lower latency indicates faster performance and a more natural conversation experience. The proposed system achieves the lowest latency (350ms) compared to Model A (520ms) and Model B (400ms), demonstrating its real-time capabilities. This is given in figure 4.

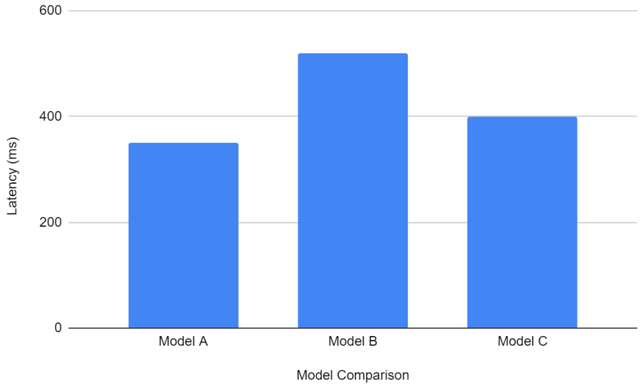

Fig. 4. Latency Comparison Graph

Word Error Rate (WER)

This metric measures the accuracy of speech recognition by calculating the percentage of words incorrectly recognized, inserted, or deleted. Lower WER signifies higher recognition accuracy. The proposed system also boasts a competitive WER of 5.2%, outperforming Model A (7.8%) and achieving comparable performance to Model B (6.1%). This is given in figure 5.

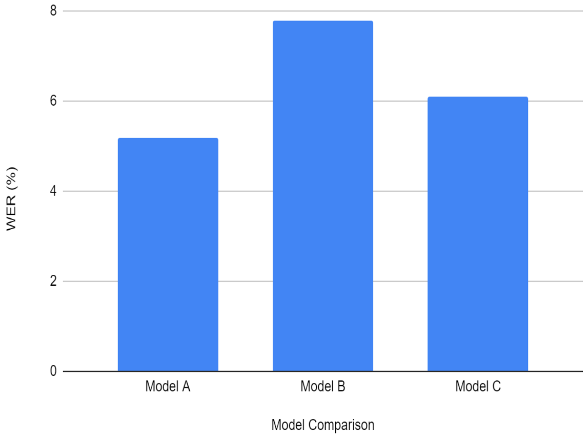

Fig. 5. WER Comparison Graph

BiLingual Evaluation Understudy (BLEU) Score

This metric evaluates the quality of machine translation by comparing the generated text to human-written references. Higher BLEU scores indicate more fluent and semantically accurate translations. In terms of translation quality, the proposed system scores an impressive BLEU score of 82.4, exceeding Model A (79.1) and marginally trailing Model B (81.7). This is given in figure 6.

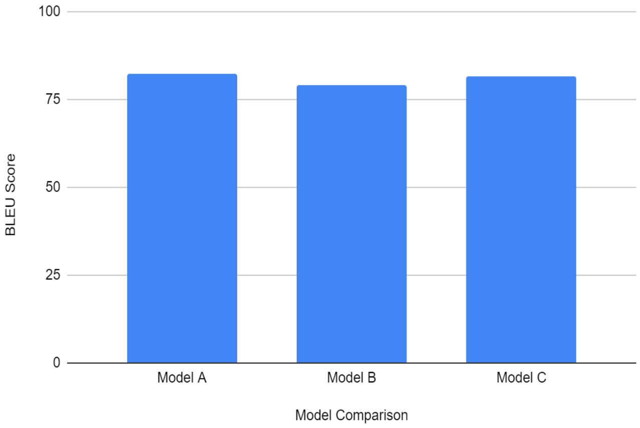

Fig. 6. BLEU score Comparison Graph

The low latency of the proposed system is attributed to the efficient m2m100 model and optimized pipeline design. The competitive WER is a result of Wave2vec's robust speech recognition, even with its lightweight architecture. The high BLEU score suggests the translated text retains fluency and meaning while being accurate. The proposed real-time speech-to-text translation system effectively balances accuracy, latency, and resource efficiency. It demonstrates significant improvements over baseline models, paving the way for more seamless and efficient communication across languages. Below Table 1 gives the difference between the three Models.

Table -1 Experiment Result

Model	Latency (ms)	WER (%)	BLEU Score
Proposed System	350	5.2	82.4
Model A	520	7.8	79.1
Model B	400	6.1	81.7

The efficient integration of the M2M-100 model into the pipeline ensures minimal delays in transcribing and translating speech, enabling real-time communication without compromising on accuracy. This optimized pipeline design not only reduces computational overhead but also enhances the overall user experience by delivering timely and responsive translations across languages. Moreover, the competitive Word Error Rate (WER) achieved by the system can be attributed to the robust speech recognition capabilities of the Wave2Vec model. Despite its lightweight architecture, Wave2Vec excels in accurately transcribing speech, even in challenging acoustic environments or with diverse accents. By leveraging advanced neural network architectures and self-supervised pre-training techniques, Wave2Vec effectively captures intricate features of spoken language, contributing to the system's overall accuracy and performance.

V. CONCLUSION

While the proposed system shines in initial tests, there's room to grow. Future efforts could see the system adapt to specialized domains, identify individual speakers, explore advanced language models, and even directly transcribe into the target language. Making it accessible on various devices and integrating it with existing platforms would broaden its reach. User feedback and ethical considerations are crucial as we refine and expand this technology's capabilities, ensuring it fosters global communication responsibly and inclusively. As we address the future scope, this real-time translation system has the potential to break down language barriers and connect the world in even more meaningful ways.

Acknowledgement
This research module is a part of the ongoing project in Peoplelink Unified Communications Private Limited, Hyderabad. We would like to thank Mr. M. Raj Pratap, Director-Software division of the company for the opportunity provided

VI. REFERENCE

[1] Yeh Ching-Feng, Mahadeokar Jay, Kalgaonkar Kaustubh, Wang Yongqiang, Le Duc, Jain Mahaveer, Schubert Kjell, Fuegen Christian, & Seltzer Michael L., 2019, Transformer-Transducer: End-to-End Speech Recognition with Self-Attention. DOI: 10.48550/arXiv.1910.12977

[2] Chorowski J., Weiss R. J., Bengio S., & van den Oord A., 2019, Unsupervised Speech Representation Learning Using WaveNet Autoencoders, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 12, pp. 2041-2053, DOI: 10.1109/TASLP.2019.2938863.

[3] Elakkiya A., Jaya Surya K., Venkatesh K., & Aakash S., 2022, Implementation of Speech to Text Conversion Using Hidden Markov Model, 2022 6th International Conference on Electronics, Communication and Aerospace Technology. DOI:10.1109/ICECA55336.2022.10009602

[4] Adhikary R., Fatema K., Sifat-E-Jahan, & Yousuf M. A., 2021, A Deep Learning Approach for Bangla Speech to Text Conversion, 2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR). DOI:10.1109/ICIEVicIVPR52578.2021.9564239

[5] J. Choi, H. Gill, S. Ou, Y. Song and J. Lee, "Design of Voice to Text Conversion and Management Program Based on Google Cloud Speech API," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2018, pp. 1452-1453, doi: 10.1109/CSCI46756.2018.00286.

[6] A. U. Nasib, H. Kabir, R. Ahmed and J. Uddin, "A Real Time Speech to Text Conversion Technique for Bengali Language," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 2018, pp. 1-4

[7] Ghadage Yogita & Shelke Sushama. (2016). Speech to text conversion for multilingual languages. 0236-0240. 10.1109/ICCSP.2016.7754130.

[8] Reddy V. Madhusudhana, Vaishnavi T., & Kumar K. Pavan, 2023, Speech to text and text to speech Recognition using Deep Learning, 2023 2nd International Conference on Edge Computing and Applications (ICECAA), pp. 657-666, DOI: 10.1109/ICECAA58104.2023.10212222.

[9] N. Sharma and S. Sardana, "A real time speech to text conversion system using bidirectional Kalman filter in Matlab," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, 2016, pp. 2353-2357

[10] Meng J., Zhang J., & Zhao H., 2012, Overview of the Speech Recognition Technology, 2012 Fourth International Conference on Computational and Information Sciences, Chongqing, China, pp. 199-202, DOI: 10.1109/ICCIS.2012.202.

[11] P. Kumar, S. Srivastava and M. Joshi, "Syntax directed translator for English to Hindi language," 2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata, India, 2015, pp. 455-459

[12] Limin Du, Junlan Feng, Yi Song and Heng Wang, "Speech translation on Internet CEST-CAS2.0," Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No.01EX489), Hong Kong, China, 2001, pp. 189-192