J. Electrical Systems 20-10s (2024): 8302-8308

Mayur Jani [1*] ,

Sandip Panchal[2],

Hemant Patel[3]

Nitesh Sureja[4]

and Ashwin Raiyani[5]

Automated Detection and Translation of Multilingual Speech: A System for Real-Time Language Recognition and Conversion

Journal of Electrical Systems

ABSTRACT

Speech is a fundamental aspect of communication, and in India, conversations frequently involve switching between multiple languages. This poses significant challenges for multilingual speech recognition systems, as accurately identifying the language of spoken words, letters, or sentences is complex. The problem is exacerbated by code-switching, where speakers seamlessly alternate between languages. Existing models, often trained on non-representative corpora of Indian languages, struggle with accuracy in these scenarios. To address these challenges, we propose a novel approach that eliminates the need for users to pre-select spoken or transcribed text languages. Our model automatically detects the language in real-time, recognizes the spoken words, and provides output text in all languages used during the conversation. This method simplifies the user experience and improves the accuracy of multilingual speech recognition, making it more effective and user-friendly in multilingual contexts.

Keywords: Code-Switching; Multilingual Speech Recognition; Real-Time Processing; Language Detection; Speech to Text.

1.0 INTRODUCTION

Day by day, the use of the internet and internet applications is increasing, connecting people and facilitating communication. However, people speak different languages, creating a need for applications or tools that can automatically detect the speaker's language, recognize words in specific languages, and translate them into the listener's language to overcome language barriers [1, 6]. Currently, many applications exist for translating one language into another. However, many people speak multiple languages during communication, leading to research focused on code-switching techniques for speech detection, recognition, and translation [2, 5].

Code-switching is crucial in speech processing as it supports various functions such as speech recognition, speech translation, and text-to-speech synthesis [6, 17]. Automatic Speech Recognition (ASR) systems are designed to convert spoken language into text. In code-switched speech, speakers switch between two or more languages or dialects within a single conversation. For instance, a speaker might say, "I am going to the mercado," where "mercado" is Spanish for "market," and the rest of the sentence is in English.

The challenge for ASR systems with code-switched speech is that conventional models are typically trained on datasets with text in a single language. As code-switched speech blends multiple languages, existing models may struggle to accurately recognize and transcribe the mixed-language speech, particularly if there is insufficient training data containing code-switched phrases or sentences [2, 17].

2.0 LITERATURE REVIEW

The methodologies and techniques used in the field of multilingual speech recognition, code-switching, and emotion recognition from speech have seen significant advancements in recent years. Yellamma et al. [1] utilized Google Cloud API for multilingual speech recognition and translation, focusing on real-time speech processing and the improvement of accuracy across languages. Similarly, Madan et al. [2] employed federated learning to enhance low-resource code-switching detection by integrating distributed training models. Lyu et al. [3] applied Whisper segmentation for real-time speech recognition, addressing the challenges of multilingual recognition in dynamic environments. Singh et al. [4] focused on the creation of a bilingual speech corpus for Manipuri-English code-switching, which has been instrumental in improving recognition systems.

[1*]Department of Information Technology, School of Enginerring and Technology, Dr. Subhash University Junagadh 362001, Gujarat, India Email: mayur.jani@dsuni.ac.in

[2]Department of Electronics and Communication Engineering, School of Enginerring and Technology, Dr. Subhash University Junagadh 362001, Gujarat, India Email:- sandip.panchal@dsuni.ac.in

[3]Department of Computer Science and Engineering, School of Enginerring and Technology, Dr. Subhash University Junagadh 362001, Gujarat, India Email:- hemant.patel@dsuni.ac.in

[4]Department of Computer Science and Engineering, Faculty of Enginerring and Technology, Drs.Kiran & Pallavi Patel Global University Vadodara 391243, Gujarat, India Email:- dr.niteshsureja.cse.kset@kpgu.ac.in

[5]Department of Undergraduate Studies, Institute of Management, Nirma University Ahemedabad 382481, Gujarat, India Email:- ashwin.rkcet@gmail.com

In the area of speech alignment, Liu et al. [5] worked on algorithms to enhance code-switching speech recognition, improving alignment between languages. Hamed et al. [8] contributed by developing a multilingual Arabic-English speech corpus, which plays a critical role in improving ASR (Automatic Speech Recognition) systems in code-switching contexts. In terms of zero-resource learning, Huang et al. [6] explored speech utterance pairs to handle code-switched speech, paving the way for unsupervised learning models. Meanwhile, Gavino and Goldrick [7] studied the perception of code-switched speech in noisy environments, emphasizing the need for ASR systems to handle challenging acoustic conditions.

Regarding feature extraction, Alasadi et al. [20] and Helali et al. [19] used traditional methods like MFCC (Mel Frequency Cepstral Coefficients) and LPC (Linear Predictive Coding) for developing speech recognition systems in Arabic. These techniques have shown significant promise, especially in low-resource environments. Additionally, Nguyen et al. [11] leveraged hybrid deep learning models for automatic multilingual code-switching speech recognition, achieving improved performance across languages. Similarly, Sailaja et al. [10] used the Hugging Face library to develop a speech-to-text system for hearing-impaired users, demonstrating that customized ASR models can enhance user accessibility. Jani et al. [21] provided an in-depth review of multilingual speech recognition, emphasizing the challenges and applications of ASR in multilingual environments. They highlighted future directions for overcoming these issues and expanding ASR capabilities. Patel and Sureja [22] discussed tokenization in their MSD Tool, emphasizing its role in converting raw data into structured tokens for generating ontology and UML diagrams. This process simplifies complex data into manageable units, facilitating easier analysis and system representation.

Kapyshev et al. [12] focused on Kazakh language speech recognition, showing the importance of language-specific ASR models for underrepresented languages. Nam and Park [13] analyzed phonemic recognition in reverberant conditions, which is crucial for improving ASR performance in environments with poor acoustics. Wang et al. [14] used a tri-stage training approach with language-specific encoders for code-switching ASR, demonstrating a significant increase in performance. This technique aligns with Fan et al. [15], who worked on sentence-level language identification for multilingual speech recognition, specifically for air traffic control environments, where accuracy is paramount.

In the field of emotion recognition from speech, Venkateswarlu et al. [16] utilized LSTM (Long Short-Term Memory) networks for sequential data processing, demonstrating the capability of deep learning in identifying emotions from speech and text. Ismaiel et al. [18] further expanded on emotion recognition using ensemble learning and supervised methods, applying these techniques to Arabic speech. Their work highlighted the effectiveness of combining multiple machine learning approaches for enhanced accuracy. Finally, Padmane et al. [9] explored multilingual speech-to-text recognition using various translation tools, showcasing the ongoing effort to improve translation accuracy in speech recognition systems.

The Table 1 below summarizes the key methodologies, techniques, current work, future work, and conclusions from the referenced papers on multilingual speech recognition, code-switching, and emotion recognition in speech technology.

Table 1: Summary of Research Papers on Speech Recognition and Related Fields

Authors (Year)	Methodology	Techniques	Current Work	Future Work	Conclusion
Yellamma et al. (2024) [1]	Google Cloud API for multilingual speech recognition and translation.	API-based real-time speech processing, translation algorithms.	Improving accuracy and speed of recognition across languages.	Expand language datasets and reduce latency.	Google Cloud API is effective for multilingual tasks, but improvements in low-resource language recognition are necessary.
Madan et al. (2024) [2]	Federated learning for low-resource code-switching detection.	Distributed training models, code-switched data integration.	Developing a scalable federated learning model for better speech recognition.	Explore cross-lingual transfer learning.	Federated learning enhances code-switching recognition, particularly for low-resource languages.
Lyu et al. (2024) [3]	Real-time speech recognition using Whisper segmentation.	Whisper model for multilingual speech segmentation and diarization.	Addressing the challenges of real-time multilingual recognition.	Extend the model to accommodate more languages and optimize segmentation accuracy.	Whisper segmentation performs well in real-time scenarios, but additional optimization for

					language identification is needed.
Singh et al. (2024) [4]	Development of a bilingual speech corpus for Manipuri-English code-switching.	Code-switching speech corpus creation, ASR model training.	Focused on the creation and usage of bilingual speech datasets for ASR systems.	Expansion of the corpus to include other language pairs.	The corpus helps improve recognition for Manipuri-English code-switched speech.
Liu et al. (2024) [5]	Speech alignment for enhanced code-switching recognition.	Alignment algorithms between code-switched languages.	Developing more accurate alignment systems for code-switching.	Incorporate additional language pairs into alignment models.	Aligning speech enhances ASR performance for code-switching tasks.
Huang et al. (2024) [6]	Zero-resource learning using speech utterance pairs for code-switched speech.	Zero-resource learning, unsupervised learning.	Benchmarked code-switching systems for various languages using minimal resources.	Integrate more unsupervised learning methods.	Zero-resource learning holds promise for speech recognition without large datasets.
Gavino & Goldrick (2024) [7]	Perception of code-switched speech in noisy environments.	Perceptual experiments, speech recognition in noise.	Understanding how noise impacts code-switched speech perception.	Explore noise-canceling methods.	Noise significantly impacts the perception of code-switched speech, requiring ASR optimization for noisy environments.
Hamed et al. (2024) [8]	Multilingual Arabic-English speech corpus development.	Corpus creation, multilingual ASR model training.	Focused on building a robust multilingual Arabic-English speech corpus.	Incorporate additional dialects.	The ZAEBUC-Spoken corpus improves ASR in Arabic-English code-switching scenarios.
Padmane et al. (2022) [9]	Multilingual speech-to-text recognition using various translation tools.	Speech-to-text algorithms, text translation models.	Improving translation accuracy for speech-to-text systems.	Develop specialized algorithms for underrepresented languages.	Multilingual systems are feasible but face challenges in translation accuracy.
Sailaja et al. (2023) [10]	Hugging Face-based speech-to-text conversion for the hearing impaired.	Hugging Face library, ASR model customization.	Designing user-friendly interfaces for hearing-impaired users.	Improve real-time accuracy and expand multilingual support.	Hugging Face integration enables efficient multilingual speech-to-text conversion, but further refinement is needed.
Nguyen et al. (2020) [11]	Automatic multilingual code-switching speech recognition.	Hybrid deep learning models for ASR.	Enhancing ASR performance in multilingual settings.	Use more advanced neural networks to improve accuracy.	Hybrid models can significantly improve ASR in code-switched environments.
Kapyshev et al. (2024) [12]	Speech recognition for Kazakh language using machine learning.	Acoustic modeling, language-specific ASR	Focus on Kazakh speech recognition and corpus	Integrate more languages and dialects.	Specialized ASR models are essential for low-resource

		design.	development.		languages like Kazakh.
Nam & Park (2024) [13]	Phonemic analysis in reverberation conditions for ASR systems.	Phonemic modeling, reverberation compensation techniques.	Analyzing the impact of reverberation on ASR accuracy.	Develop more robust reverberation correction methods.	Reverberation can degrade ASR performance, and compensatory techniques are critical for improving accuracy.
Wang et al. (2024) [14]	Tri-stage training with language-specific encoders for code-switching ASR.	Bilingual acoustic models, encoder-decoder architecture.	Improving code-switching ASR performance using tri-stage training.	Expand tri-stage training to more language combinations.	Tri-stage training significantly boosts ASR performance in code-switching contexts.
Fan et al. (2024) [15]	Sentence-level language identification for multilingual speech recognition in air traffic control.	Sentence-level language ID, ASR system design for air traffic environments.	Improving ASR for air traffic control environments.	Apply the system to other domains requiring multilingual ASR.	Sentence-level language ID enhances ASR performance in air traffic control, where accuracy is critical.
Venkateswarlu et al. (2023) [16]	Emotion recognition from speech and text using LSTM (Long Short-Term Memory) networks.	LSTM for sequential data processing, feature extraction from speech and text.	Improving emotion recognition accuracy using deep learning models.	Explore more advanced neural networks and extend the work to other languages.	LSTM effectively recognizes emotions from speech and text, though further refinement is needed for real-time applications.
Mustafa et al. (2022) [17]	Overview of issues in code-switching for automatic speech recognition.	Survey of existing ASR techniques, analysis of code-switching challenges.	Addressing limitations in current ASR models for code-switching scenarios.	Develop models that handle complex linguistic structures in code-switching.	Existing ASR systems struggle with code-switching, necessitating future work on cross-lingual speech models.
Ismaiel et al. (2024) [18]	Arabic speech emotion recognition using deep learning, ensemble, and supervised learning methods.	Ensemble learning, supervised machine learning, deep neural networks for emotion classification.	Improving emotion recognition in Arabic speech using multiple AI techniques.	Extend the study to more dialects and emotional categories.	Ensemble and deep learning approaches provide strong results in speech emotion recognition but need more training data for improved performance.
Helali et al. (2020) [19]	Real-time speech recognition using PWP (Peak Word Probability) thresholding and MFCC (Mel	MFCC feature extraction, SVM for classification, PWP for word-level	Developing a real-time speech recognition system using classical machine	Explore deep learning methods for further improvement.	MFCC and SVM provide satisfactory real-time speech recognition, though further

	Frequency Cepstral Coefficients) with SVM (Support Vector Machine).	recognition.	learning techniques.		optimization is needed for larger datasets.
Alasadi et al. (2020) [20]	Feature extraction algorithms for developing an Arabic speech recognition system.	MFCC, LPC (Linear Predictive Coding), PLP (Perceptual Linear Prediction) for feature extraction.	Improving feature extraction for Arabic speech recognition.	Apply the feature extraction techniques to other low-resource languages.	Effective feature extraction techniques improve ASR performance, particularly for Arabic, but can be extended to other languages.

3.0 RESEARCH GAP

After reviewing the relevant literature, several key research gaps emerged in multilingual speech recognition and code-switching. These gaps include the lack of comprehensive evaluation across diverse languages and dialects, insufficient real-time processing capabilities, and limited robustness in handling complex code-switching patterns. Additionally, there is a need for better dataset variety and improved alignment accuracy in speech recognition systems. Moreover, challenges such as integrating under-researched languages, developing advanced algorithms for low-resource settings, and ensuring scalability and robustness in real-time applications remain largely unaddressed by current research.

- **Lack of Comprehensive Evaluation across Diverse Languages and Dialects:**

There is a significant need for comprehensive evaluation of speech recognition models across India's diverse languages and dialects, particularly for low-resource languages. This gap in evaluation limits the generalizability of speech recognition systems in multilingual contexts [1], [4], [12].

- **Insufficient Real-Time Processing Capabilities and Optimization:**

There has been insufficient exploration of real-time processing and optimization in dynamic environments involving multiple languages, which restricts the applicability of these systems in real-world scenarios [1], [2], [3].

- **Limited Robustness in Handling Diverse Code-Switching Scenarios:**

Existing models lack robustness when it comes to handling diverse and spontaneous code-switching scenarios, which is common in multilingual speech recognition systems [2], [5], [6].

- **Handling of Complex Code-Switching Patterns:**

Current approaches fail to adequately address the complex patterns of code-switching that frequently occur in multilingual conversations, especially in contexts like India [5], [6].

- **Limited Integration of Additional Languages and Dialects:**

Many speech recognition systems fail to integrate a broad range of languages and dialects, particularly under-researched Indian languages, which limits their effectiveness in multilingual settings [2], [3], [12].

- **Insufficient Variety and Size of Datasets:**

These systems are often trained on datasets that lack sufficient variety and size, further restricting their ability to generalize across different linguistic contexts [3], [12].

- **Improving Alignment Accuracy in Speech Recognition:**

More precise methods are needed to align speech with text, particularly in multilingual and code-switching contexts, to improve recognition accuracy [12].

- **Development of Advanced Algorithms for Zero-Resource Settings:**

There is a pressing need for advanced algorithms that can function in low-resource settings, where labeled data for speech recognition is scarce [4].

- **Impact of Noise on Perception of Code-Switched Speech:**

The effects of varying noise levels on the recognition of code-switched speech are not fully understood, limiting the performance of ASR systems in real-world scenarios [7].

- **Exploration of Cross-Lingual and Cross-Dialectal Transfer Learning Techniques:**

Exploring cross-lingual and cross-dialectal transfer learning techniques remains an under-researched area, despite their potential to improve the adaptability of ASR systems across different languages [4], [7].

- **Application to Diverse Language Pairs and Datasets:**

Existing approaches are also not extensively validated across diverse language pairs and datasets, restricting their applicability in multilingual environments like India [3], [4].

- **Ensuring Scalability and Robustness in Real-Time Applications:**

Ensuring scalability and robustness in real-time multilingual ASR applications remains a challenge, particularly in Indian contexts [1], [12].

4.0 RESEARCH WORK

we aim to enhance our system by eliminating the need for users to manually select specific source and destination languages. Instead, we intend to develop a system that automatically detects the language spoken by users and converts it into appropriate output.

Furthermore, we aim to improve the current system to accommodate users speaking multiple languages simultaneously. Our goal is for the system to detect multiple languages within the user's speech and accurately convert them into the appropriate words in the desired output language. This entails implementing sophisticated language detection algorithms alongside robust speech-to-text and translation mechanisms.

By seamlessly integrating these components, we aspire to create a user-friendly experience that removes language barriers and facilitates effective communication across linguistic boundaries.

For above work our proposed working flow gives in below Figure 1.

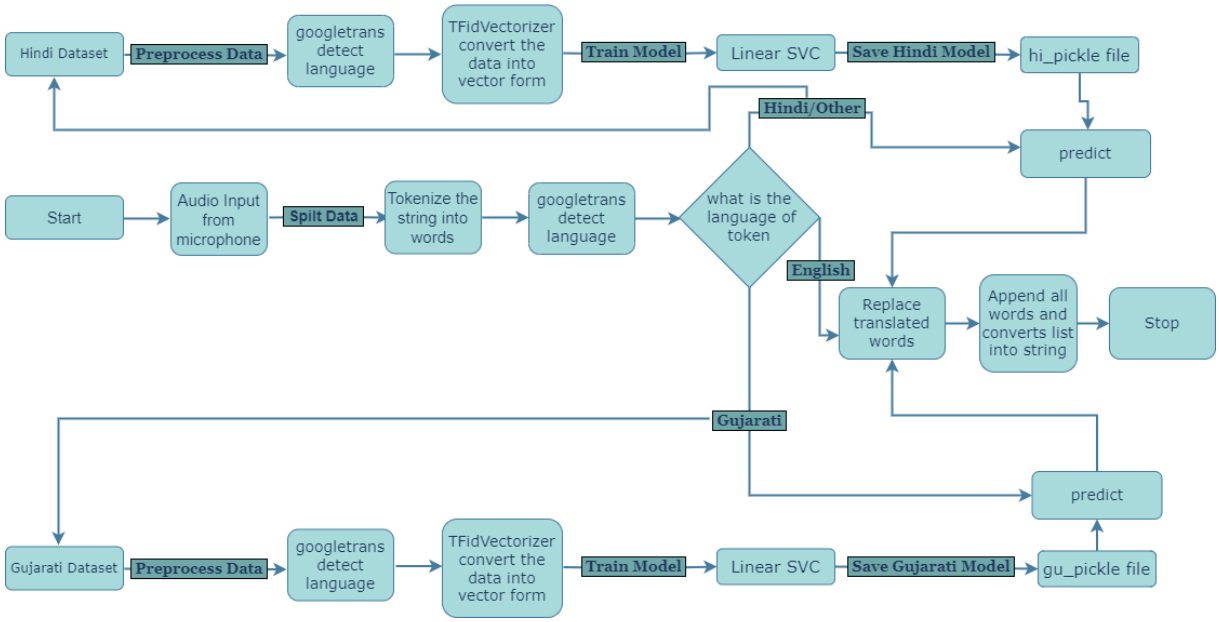

Fig. 1: Working Flowchart

5.0 CONCLUSIONS

The proposed project introduces a novel approach to multilingual speech recognition and translation, aiming to overcome language barriers effectively. Its high accuracy in recognizing speech and providing precise translations benefits users from various linguistic backgrounds. The intuitive interface makes it accessible to all, regardless of technical skill. Additionally, the project emphasizes robust data privacy and security, adhering to data protection laws and building user trust. Future developments include real-time translation, enhanced recognition models, and extended language support, which will further enhance its functionality and versatility for multilingual communication.

To summarize, the project combines advanced speech recognition and translation technologies with a strong focus on usability and data security, making it a valuable tool for facilitating cross-linguistic understanding and accessibility. Its potential for continuous improvement ensures its relevance in the evolving field of multilingual communication.

REFERENCES

[1] P. Yellamma, P. Venkataiah, and S. Devarakonda, “Automatic and multilingual speech recognition and translation by using Google Cloud API”, in 2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), 2024, pp. 1–7.

[2] C. Madan, H. Diddee, D. Kumar, and M. Mittal, “CodeFed: Federated speech recognition for low-resource code-switching detection”, ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 23, no. 1, pp. 1–14, 2024.
[3] K.-M. Lyu, R. Lyu, and H.-T. Chang, “Real-time multilingual speech recognition and speaker diarization system based on Whisper segmentation”, PeerJ Computer Science, vol. 10, p. e1973, 2024.
[4] N. K. Singh, Y. J. Chanu, and H. Pangsatabam, “MECOS: A bilingual Manipuri-English spontaneous code-switching speech corpus for automatic speech recognition”, Computer Speech & Language, vol. 77, p.
[5] H. Liu, X. Zhang, L. P. Garcia-Perera, A. W. H. Khong, E. S. Chng, and S. Watanabe, “Aligning speech to languages to enhance code-switching speech recognition”, arXiv Preprint, arXiv:2403.05887, 2024.
[6] K.-P. Huang, L. Wang, Y. Huang, and Y.-A. Chung, “Zero resource code-switched speech benchmark using speech utterance pairs for multiple spoken languages”, in ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 1–5.
[7] M. F. Gavino and M. Goldrick, “The perception of code-switched speech in noise”, JASA Express Letters, vol. 4, no. 3, pp. 1–7, 2024.
[8] I. Hamed, H. Benbrahim, R. Azzouz, K. Aissous, and M. Lahlou, “ZAEBUC-Spoken: A multilingual multidialectal Arabic-English speech corpus”, arXiv Preprint, arXiv:2403.18182, 2024.
[9] P. Padmane, A. Pakhale, S. Agrel, A. Patel, S. Pimparkar, and P. Bagde, “Multilingual speech and text recognition and translation”, International Journal of Innovations in Engineering and Science, vol. 7, no. 8, pp. 84–88, 2022.
[10] N. V. Sailaja, B. Sushma, A. L. Reddy, C. Parachuri, and C. Akash, “Multilingual speech-to-text conversion using Hugging Face for deaf people”, International Research Journal of Engineering and Technology (IRJET), vol. 10, no. 5, pp. 1–5, 2023. e-ISSN: 2395-0056.
[11] T. A. Nguyen, T. H. Dang, and T. H. Nguyen, “Automatic multilingual code-switching speech recognition”, International Journal of Engineering and Applied Sciences, vol. 7, no. 5, pp. 47–53, 2020.
[12] G. Kapyshev, M. Nurtas, and A. Altaibek, “Speech recognition for Kazakh language: A research paper”, Procedia Computer Science, vol. 231, pp. 369–372, 2024.
[13] H. Nam and Y. H. Park, “Coherence-based phonemic analysis on the effect of reverberation to practical automatic speech recognition”, Applied Acoustics, vol. 227, p. 110233, 2024.
[14] X. Wang, Y. Jin, F. Xie, and Y. Long, “Tri-stage training with language-specific encoder and bilingual acoustic learner for code-switching speech recognition”, Applied Acoustics, vol. 218, p. 109883, 2024.
[15] P. Fan, D. Guo, J. Zhang, B. Yang, and Y. Lin, “Enhancing multilingual speech recognition in air traffic control by sentence-level language identification”, Applied Acoustics, vol. 224, p. 110123, 2024.
[16] S. C. Venkateswarlu, S. R. Jeevakala, N. U. Kumar, P. Munaswamy, and D. Pendyala, “Emotion recognition from speech and text using long short-term memory”, Engineering, Technology & Applied Science Research, vol. 13, no. 4, pp. 11166–11169, 2023.
[17] M. B. Mustafa, M. R. Ali, A. H. Ahmed, S. J. Khan, and F. Abbas, “Code-switching in automatic speech recognition: The issues and future directions”, Applied Sciences, vol. 12, no. 19, p. 9541, 2022.
[18] W. Ismaiel, A. Alhalangy, A. O. Y. Mohamed, and A. I. A. Musa, “Deep learning, ensemble and supervised machine learning for Arabic speech emotion recognition”, Engineering, Technology & Applied Science Research, vol. 14, no. 2, pp. 13757–13764, 2024.
[19] W. Helali, Z. Hajaiej, and A. Cherif, “Real-time speech recognition based on PWP thresholding and MFCC using SVM”, Engineering, Technology & Applied Science Research, vol. 10, no. 5, pp. 6204–6208, 2020.
[20] A. A. Alasadi, T. H. Aldhayni, R. R. Deshmukh, A. H. Alahmadi, and A. S. Alshebami, “Efficient feature extraction algorithms to develop an Arabic speech recognition system”, Engineering, Technology & Applied Science Research, vol. 10, no. 2, pp. 5547–5553, 2020.
[21] M. M. Jani, S. R. Panchal, H. H. Patel, and A. Raiyani, “Multilingual speech recognition: An in-depth review of applications, challenges, and future directions”, in International Conference on Communication and Intelligent Systems, Springer Nature, Singapore, pp. 1–13, Dec. 2023.
[22] H. H. Patel and N. M. Sureja, “MSD tool: An automatically produced ontology and UML diagram for multi-site software development”, Turkish Journal of Computer and Mathematics Education, vol. 12, no. 8, pp. 2110–2121, 2021.