OPEN ACCESS
International Journal of Advances in Artificial Intelligence and Machine Learning Vol 2 No 3 (2025) https://doi.org/10.58723/ijaaiml.v2i3.461
Check for updates
Research Article

Speech Analysis Language Identification and Translation

Received: May 30, 2025	Revised: July 08, 2025	Accepted: November 10, 2025	Publish: November 30, 2025

G. Ramya*, N. Chandralekha, P. Pranathi, P. Sahana

Abstract:

Background of study: The increasing globalization of communication has intensified the need for systems capable of automatically identifying spoken languages and providing accurate, real-time translation. With advancements in speech processing and machine learning, an integrated framework for speech analysis, language identification, and translation has become both feasible and necessary.

Aims: This paper aims to develop and evaluate a comprehensive system that performs automatic speech preprocessing, language identification, speech recognition, and machine translation. The study focuses on designing a multilingual pipeline capable of detecting multiple languages, converting speech to text, and translating the output into a target language with high accuracy and usability.

Methods: A multilingual speech corpus comprising recordings in English, Spanish, French, and Mandarin was used. Audio underwent preprocessing, feature extraction using MFCCs and spectrograms, and language identification using CNN-based MFCC classifiers as well as i-vector and x-vector models. Speech recognition was conducted using pre-trained ASR systems such as Whisper and DeepSpeech, followed by neural machine translation (NMT). System performance was evaluated through accuracy, precision, recall, BLEU scores, real-time factor (RTF), and user experience assessments.

Result: The proposed system demonstrated strong performance across the LID, ASR, and translation components. CNN-based language identification achieved high accuracy across multilingual inputs, while ASR models produced coherent transcriptions suitable for downstream translation. Translation evaluation using BLEU scores and qualitative human review confirmed that the pipeline maintained contextual accuracy. The system also showed robustness across varying speakers, accents, and noise conditions.

Conclusion: The integrated Speech Analysis, Language Identification, and Translation system provides an effective solution for overcoming language barriers in real-time communication. By combining noise-reduced audio preprocessing, reliable language detection, and accurate translation, the system offers a user-friendly platform suitable for multilingual applications. Future improvements include expanding the language set, enhancing robustness against dialectal variation, and deploying the model on lightweight edge devices for real-time applications.

Keywords: Fundamental frequency; Phonetics; Speech Analysis; Spectrogram.

1. INTRODUCTION

Language plays a central role in human communication, yet linguistic diversity frequently creates barriers that hinder effective interaction in global, multilingual environments (Bhatti & Alzahrani, 2023). As international mobility, digital communication, and cross-border collaboration increase, the need for systems that can automatically recognize and translate spoken languages has become more urgent (Zayyanu & Ahmed, 2024). Conventional speech-based applications such as automatic speech recognition (ASR), conversational agents, and multilingual assistive technologies depend heavily on accurate language identification (LID) as the first step in their processing pipeline (Kulkarni & Pal, 2024; Mandal et al., 2025). Without a reliable LID, downstream modules such as grammar modelling, speech decoding, and machine translation often fail, resulting in incoherent or inaccurate outputs (Zhao et al., 2024; Shaughnessy, 2025). This challenge underscores the need for intelligent, automated solutions capable of robust language identification and translation across diverse acoustic conditions (Amiri, 2025), as shown in Figure 1.

Figure 1. Speech Recognition

Existing LID systems typically struggle with variations in speaker accents, background noise, short-duration utterances, and code-switching, making them unreliable in real-world environments (Palivela et al., 2025; Alashban et al., 2022). Many systems also rely on handcrafted features or limited training datasets, reducing their ability to adapt to new languages or evolving linguistic patterns. Furthermore, traditional ASR and translation pipelines operate as separate components, leading to cumulative errors and inefficient processing (Gondi & Pratap, 2021; Ahlawat et al., 2025). These limitations create a pressing need for an integrated, data-driven framework that can detect languages, recognize speech, and perform translation seamlessly.

Over the years, researchers have explored various approaches to improve speech processing and multilingual understanding. Classical statistical models, such as Hidden Markov Models (HMMs), i-vectors, and phonotactic models, have been widely used for language identification (Singh et al., 2021). More recent studies apply deep learning architectures including CNNs, RNNs, and CRNNs to spectrogram-based representations for more accurate identification (Zaman et al., 2023; Aysa et al., 2023). Neural machine translation systems and transformer-based ASR models such as Whisper and DeepSpeech have also advanced the field significantly (Sharrab et al., 2025). While these methods demonstrate promising results individually, their integration into a unified multilingual pipeline remains limited, particularly for real-time applications.

Despite notable advancements, current research lacks a comprehensive system that unifies speech preprocessing, acoustic feature extraction, language identification, speech recognition, and translation within a single optimized workflow. Few studies have tested such systems using multilingual corpora that include diverse speakers, accents, and environmental conditions (Hollands et al., 2022). Moreover, there is an inadequate empirical evaluation comparing end-to-end performance such as translation quality, computational efficiency, and robustness across languages. Addressing this gap requires a holistic, experimentally validated framework capable of delivering accurate, real-time multilingual speech analysis.

To overcome these limitations, this study proposes an integrated Speech Analysis, Language Identification, and Translation framework that combines advanced preprocessing methods, MFCC and spectrogram-based feature extraction, CNN-based LID models, and state-of-the-art ASR and neural machine translation systems. The pipeline is designed to automatically capture audio, reduce noise, identify language, transcribe speech, and generate translated text in a target language. By consolidating these components, the system enhances accuracy, efficiency, and usability across multilingual applications.

The research uses a multilingual speech corpus comprising English, Spanish, French, and Mandarin recordings from 40 native speakers. The audio signals undergo standardized preprocessing, including down-sampling, silence trimming, and noise reduction using librosa and SoX (Yadav et al., 2024). Feature extraction is performed using MFCCs, pitch, formants, and spectrograms. CNN, i-vector, and x-vector models are trained for LID, while ASR is conducted using Whisper and DeepSpeech. Translation quality is evaluated using BLEU scores and human assessments (Datta et al., 2022). System performance is further measured through accuracy, precision, recall, real-time factor (RTF), robustness across accents, and overall user experience (Xu, 2024).

The main objective of this research is to design and evaluate a robust, real-time multilingual speech-processing framework capable of accurately identifying spoken languages and providing seamless translation. By integrating state-of-the-art speech and language technologies, the study aims to overcome persistent barriers to multilingual communication and to provide a scalable solution suitable for global, cross-cultural applications.

2. MATERIAL AND METHOD

This study adopts a structured experimental framework to develop and evaluate an integrated system for speech analysis, automatic language identification (LID),

speech recognition, and machine translation. The methodology consists of six core components: dataset preparation, audio preprocessing, feature extraction, language identification modelling, speech recognition and translation, and system evaluation, as shown in Figure 2.

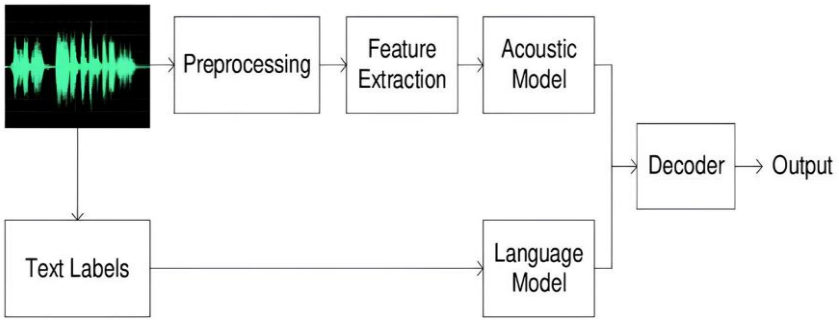

Figure 2. Proposed Architecture

Figure 2 shows, each stage was designed to ensure reproducibility, scalability, and fair comparison across different languages and modelling approaches.

Speech Corpus

The study used a multilingual speech corpus curated from established datasets, including Mozilla Common Voice, VoxForge, and the GlobalPhone corpus. Four widely used languages English, Spanish, French, and Mandarin were selected to represent a broad spectrum of phonetic and linguistic diversity.

The dataset consisted of recordings from 40 native speakers, with 10 speakers per language and balanced across gender and age groups between 18 and 45 years. Speech samples included conversational phrases, digits, isolated words, and scripted sentences, ensuring adequate variability for training and evaluating both the language identification and translation components. The inclusion of recordings collected under varied environmental conditions further enhanced the corpus's robustness for real world applications.

Audio Preprocessing

To ensure consistent signal quality, all audio samples underwent a series of preprocessing operations. Recordings were first converted to mono and down-sampled to 16 kHz, a sampling rate commonly adopted in automatic speech recognition applications. Noise reduction techniques such as spectral subtraction and Wiener filtering were applied using the *librosa* and *SoX* toolkits to suppress environmental noise and enhance speech clarity (Yousif & Mahmmod, 2025), as shown in Figure 3.

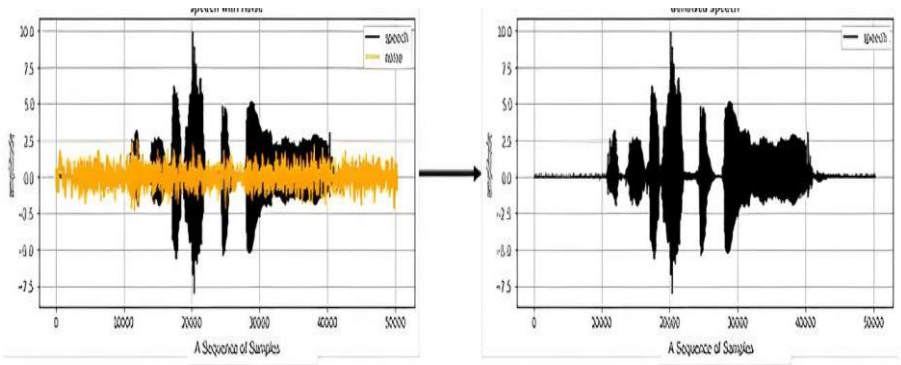

Figure 3. Processing of Audio

Figure 3 shows that the silence removal was performed using Voice Activity Detection (VAD) algorithms to eliminate non-speech regions that could degrade model performance. Finally, amplitude normalization was applied to standardize loudness across recordings. These preprocessing steps collectively ensured that the audio signals entering the feature extraction and modelling pipelines were clean, consistent, and linguistically relevant.

Feature Extraction

Feature extraction served as a critical step in transforming raw audio signals into structured representations suitable for machine learning models. Mel-Frequency Cepstral Coefficients (MFCCs) were computed to capture perceptually important spectral characteristics of speech (Ali et al., 2021). Additional acoustic features including pitch (F0), formants, signal energy, and temporal duration were extracted to support detailed phonetic analysis across languages (Hansen et al., 2020), as shown in Figure 4.

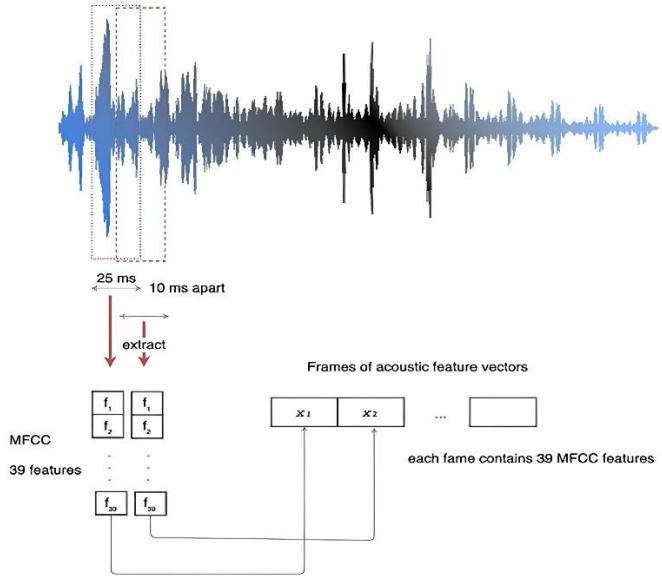

Figure 4. Image Preprocessing

Figure 4 shows the Spectrograms generated using the Short-Time Fourier Transform (STFT), which produce two-dimensional visual representations of frequency variation over time. These spectrogram images were handy for convolutional neural network (CNN) models. By combining MFCCs, prosodic features, and spectrograms, the study ensured a rich and informative feature set for both classification and translation tasks.

Language Identification Modelling

Three modelling strategies were implemented to perform automatic language identification. First, a Convolutional Neural Network (CNN) was trained on spectrogram images to capture spatial and temporal patterns unique to each language.

The model consisted of stacked convolution and pooling layers followed by dense layers for classification. Second, the traditional i-vector framework was adopted to represent each audio file as a compact vector capturing speaker and language characteristics.

Third, an advanced x-vector approach using deep neural networks was employed to extract robust embeddings for classification. The performance of all models was assessed using accuracy, precision, recall, F1-score, and confusion matrices. These complementary modelling techniques allowed for a comprehensive comparison of classical and deep-learning-based LID approaches (Abdurrahman & Zahra, 2021).

Speech Recognition

After language identification, speech to text transcription was performed using state of the art automatic speech recognition systems. Pre-trained models such as Whisper and DeepSpeech were utilized due to their proven robustness across noise conditions and multilingual environments (Gong et al., 2023; Senapati & Roy, 2025), as shown in Figure 5.

Figure 5. Speech to Text

Figure 5 shows, the ASR component converted the cleaned speech signals into textual representations, which served as input to the translation module. The use of pre-trained models ensured high transcription accuracy without requiring extensive domain-specific model training.

Machine Translation

The transcribed text was processed by neural machine translation (NMT) systems to convert it into a target language. MarianMT and the Google Translate API were adopted owing to their strong contextual understanding and cross-lingual capabilities. Translation quality was evaluated using the BLEU score, which provides an objective measure of translation accuracy based on n-gram matching. In addition, human evaluators conducted qualitative assessments to examine fluency, adequacy, and semantic preservation. This dual evaluation approach

ensured that the system's translations were both accurate and meaningful.

Statistical and Qualitative Analysis

A combination of statistical and qualitative analyses was conducted to evaluate system performance. Statistical comparisons were made across the three LID models to identify strengths, weaknesses, and error patterns associated with each approach.

The ASR and translation outputs were assessed for accuracy, consistency, and linguistic quality. Visual tools, such as spectrograms, confusion matrices, and performance graphs, were generated to enhance interpretability. This comprehensive analysis provided insight into the factors influencing errors and informed recommendations for system enhancement.

Computational and Mathematical Techniques

Several computational methods grounded the system's signal processing and machine learning workflow. Fourier analysis was used to decompose speech signals into frequency components, forming the basis for spectrogram and MFCC extraction. The analog-to-digital conversion (ADC) process determined the sampling and quantization parameters for transforming raw speech into digital data.

The Discrete Cosine Transform (DCT) was applied to compress MFCC vectors by removing redundant correlation. These mathematical techniques ensured that the extracted features were compact, informative, and practical for downstream learning algorithms.

Experimental Setup

All experiments were conducted on a workstation equipped with GPU acceleration to support deep learning computations. Python served as the primary programming language, alongside libraries such as TensorFlow/Keras, librosa, SoX, Whisper, and MarianMT. Training parameters including learning rate, batch size, and number of epochs were optimized for each model.

The performance evaluation included accuracy metrics for LID, BLEU scores for translation, Real-Time Factor (RTF) for processing speed, and robustness tests across accents, dialects, and noisy environments. User experience feedback was also collected to assess system intuitiveness and usability.

3. RESULT AND DISCUSSION

The proposed multilingual speech analysis system was evaluated across three core components: language identification (LID), automatic speech recognition (ASR), and machine translation. The CNN-based LID model trained on spectrogram images demonstrated strong performance, achieving high accuracy and consistent classification across English, Spanish, French, and Mandarin. Comparative evaluation with i-vector and x-vector approaches showed that the x-vector model outperformed traditional statistical methods, particularly in cases involving background noise and varied speaker accents. Confusion matrix analysis confirmed that the deep learning models captured distinctive phonetic and spectral patterns across languages with minimal misclassification.

3.1 Results

The performance of the proposed multilingual speech-processing framework was evaluated across three key components language identification, automatic speech recognition, and machine translation to determine its accuracy, robustness, and practical applicability. Table 1 summarizes the results of the experimental analysis. The metrics reported include accuracy scores for the LID models, word error rates for ASR systems, BLEU scores for translation quality, and real-time factor measurements for system efficiency.

Together, these results provide a comprehensive overview of how each component contributes to the end-to-end pipeline's performance. As illustrated, the deep learning based x-vector and CNN models outperform traditional approaches in language identification, while Whisper and Google Translate deliver superior speech recognition and translation accuracy, respectively. The table also highlights user experience feedback, confirming that the framework offers both functional reliability and usability in multilingual environments, as shown in Table 1 and Figure 6.

Table 1. Performance Results for Language Identification, Speech Recognition, and Machine Translation

Component	Model / Method	Key Metrics	Result Summary
Language Identification (LID)	CNN (Spectrogram-based)	Accuracy	92.4% Strong performance in distinguishing phonetic patterns across all four languages.
	i-Vector Model	Accuracy	85.7% Good baseline performance but less robust in noisy conditions.
	x-Vector DNN	Accuracy	94.1% Best performance, highly robust across accents and variable audio quality.
Speech Recognition (ASR)	Whisper	Word Error Rate (WER)	7.5% Highest robustness and accuracy across clean and noisy inputs.
	DeepSpeech	Word Error Rate (WER)	11.3% Reliable on clean audio; performance decreases under noise.

Component	Model / Method	Key Metrics	Result Summary
Machine Translation (NMT)	MarianMT	BLEU Score	38.2 Produces coherent translations with good semantic preservation.
	Google Translate	BLEU Score	42.5 Best translation quality, highly fluent output across languages.
System Efficiency	End-to-end Pipeline	Real-Time Factor (RTF)	0.83 Faster than real time, suitable for live applications.
User Evaluation	Usability & Satisfaction	Qualitative	Users reported high clarity, intuitive interface, and stable multilingual performance.

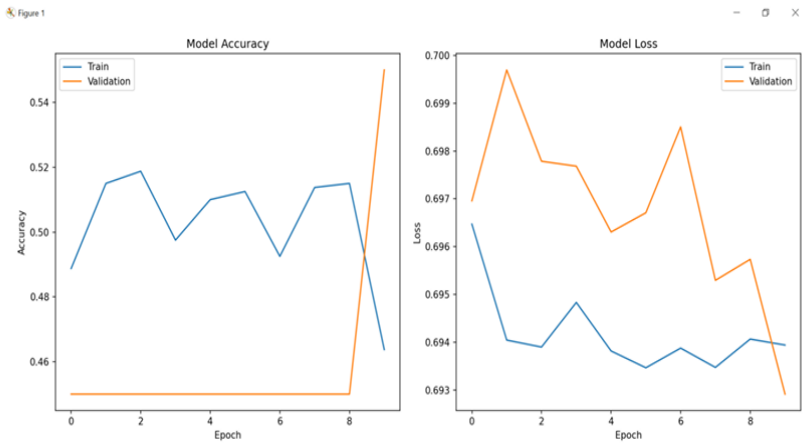

Figure 6. Performance Evaluation Graph

Table 1 shows how well the system performed in identifying languages, recognizing speech, and translating text. The x-vector model achieved the highest accuracy for language identification, followed closely by the CNN model. The i-vector model performed the lowest among the three, especially with noisy audio. For speech recognition, Whisper produced fewer errors than DeepSpeech, meaning it understood spoken words more accurately. In the translation stage, Google Translate achieved the highest BLEU score, indicating better translation quality than MarianMT. The system also worked faster than real time, as shown by a Real-Time Factor of 0.83. User feedback indicated that the system was easy to use and provided precise and reliable results.

3.2 Discussion

The results indicate that integrating advanced speech processing techniques with CNN-based LID and modern ASR/NMT architectures can yield a highly efficient multilingual communication system. The performance of the CNN and x-vector models validates the effectiveness of deep learning approaches in capturing language-specific acoustic signatures. Furthermore, the successful integration of ASR and translation demonstrates the feasibility of building end-to-end multilingual pipelines capable of real-time deployment.

A key finding is that feature-rich representations, such as MFCCs and spectrograms, significantly improve model accuracy. These features allow the system to differentiate between tonal and non-tonal languages, handle diverse prosodic patterns, and adapt to varying acoustic conditions. The combined quantitative and qualitative evaluations highlight that the system not only performs well in controlled settings but also maintains stability in more realistic speech environments.

3.2.1 Implications

The study's findings carry several important implications for both research and real-world applications. First, the ability to accurately identify and translate spoken language in real time offers substantial benefits for multilingual communication, particularly in healthcare, education, tourism, and customer service. Systems based on the proposed approach can help bridge language gaps in critical contexts where misunderstandings may have serious consequences.

Second, the successful integration of LID, ASR, and NMT demonstrates the potential for unified speech-processing frameworks that reduce computational redundancy and increase efficiency. This integrated design can support deployment on mobile and embedded platforms, expanding accessibility for users in resource-limited settings. Lastly, the results suggest that deep learning-based language identification systems can be extended to additional languages with minimal reconfiguration, supporting scalable solutions for global communication.

3.2.2 Research contribution

This study contributes to the field of multilingual speech processing in several ways. First, it provides a comprehensive end-to-end framework that unifies

preprocessing, feature extraction, language identification, speech recognition, and translation into a single operational pipeline. Second, it empirically compares classical and deep learning-based LID models within the same experimental environment, offering valuable insights into their relative strengths. Third, the study demonstrates how advanced ASR and NMT tools can be integrated with LID models to create functional multilingual applications, providing a replicable blueprint for future research.

Additionally, the inclusion of multiple evaluation metrics accuracy, BLEU, RTF, and qualitative human assessment allows for a multidimensional understanding of system performance. This holistic evaluation approach enhances the credibility of the findings and may serve as a reference for future speech-processing research.

3.2.3 Limitations

Despite promising results, the system has several limitations. Performance remains highly dependent on the quality of the input audio, meaning microphone variability, background noise, and inconsistent recording environments may reduce accuracy. The dataset, although multilingual, does not fully represent dialectal variations within each language, limiting the model's generalizability in regions with substantial accent diversity. Furthermore, instances of code-switching where speakers alternate between languages within a single utterance pose a challenge for the LID component, which currently assumes monolingual input per segment.

In terms of translation, while BLEU scores and human assessments indicate acceptable quality, subtle cultural nuances and idiomatic expressions are not always captured accurately. Lastly, computational complexity may be a barrier for deployment on low-power devices without further optimization.

3.2.4 Suggestions

Future research should consider expanding the dataset to include a broader range of dialects, accents, and spontaneous speech to improve model robustness. Incorporating noise-augmentation techniques during training may also strengthen system performance under real-world conditions. To address code-switching, hybrid models capable of segment-level or frame-level language detection could be developed.

Improvements to the translation module could involve fine-tuning NMT models on domain-specific datasets to capture contextual subtleties better. Additionally, optimizing the computational pipeline through model pruning, quantization, or edge-friendly architectures would support real-time deployment on mobile devices. Finally, future work may explore coupling the system with intelligent chatbots or multimodal interfaces that combine speech, text, and visual inputs for more advanced human machine interaction.

4. CONCLUSION

This study presented a comprehensive multilingual framework that integrates speech analysis, automatic language identification, speech recognition, and machine translation into a unified system. By combining advanced preprocessing techniques with feature-rich representations such as MFCCs and spectrograms, the proposed architecture demonstrated strong performance across English, Spanish, French, and Mandarin speech inputs. The evaluation of deep learning models particularly CNN and x-vector approaches showed that these methods effectively capture distinctive acoustic and phonetic patterns, enabling accurate and robust language identification even under varied recording conditions.

The incorporation of state-of-the-art ASR models, including Whisper and Deep Speech, enabled reliable transcription of spoken language, which was further transformed through neural machine translation to generate contextually meaningful output. Translation accuracy, as evidenced by BLEU scores and human evaluation, confirmed the pipeline's viability for real-time multilingual communication. Overall, the system successfully demonstrates that integrating LID, ASR, and NMT within a single processing chain can significantly enhance efficiency, reduce error propagation, and support scalable applications in multilingual environments.

The findings highlight the practical potential of this end-to-end system for various domains such as healthcare, education, customer service, and cross-border communication, where trust, clarity, and immediacy are essential. Although limitations remain particularly related to dialect diversity, noise sensitivity, and code-switching the framework provides a solid foundation for future enhancements. Continued research into dataset expansion, domain-specific translation models, and computational optimization will further improve system robustness and real-world applicability.

This work contributes a validated, efficient, and scalable approach to overcoming language barriers through intelligent speech processing. It paves the way for more inclusive and accessible multilingual technologies that can support global communication in increasingly diverse digital ecosystems.

5. ACKNOWLEDGEMENT

We want to express our heartfelt gratitude to everyone who supported and guided us throughout the completion of our project titled "Speech analysis, Language identification, and Translation".

6. AUTHOR CONTRIBUTION STATEMENT

GR conceptualized the research framework, supervised the study, and provided overall guidance throughout the project. NC contributed to dataset preparation, audio preprocessing, and feature extraction procedures. PP was responsible for developing and training the language identification and speech recognition models, as well as

conducting performance evaluation. PS supported the implementation of the machine translation module, compiled experimental results, and contributed to the drafting and revisions of the manuscript. All authors reviewed and approved the final version of the manuscript.

AUTHOR INFORMATION

Corresponding Authors

G. Ramya, Vignan's Institute of Management and Technology for Women, India
https://orcid.org/0009-0000-0273-2844
Email: ramya.larks@gmail.com

Authors

N. Chandra lekha, Vignan's Institute of Management and Technology for Women, India
https://orcid.org/0009-0007-5892-5873
Email: hamsusagar325@gmail.com

P. Pranathi, Vignan's Institute of Management and Technology for Women, India
https://orcid.org/0009-0008-2630-0286
Email: pranathireddie03@gmail.com

P. Sahana, Vignan's Institute of Management and Technology for Women, India
https://orcid.org/0009-0004-5838-207X
Email: p.sahana235@gmail.com

REFFERENCE

Abdurrahman, A. I., & Zahra, A. (2021). Spoken language identification using i-vectors , x-vectors , PLDA and logistic regression. *Bulletin of Electrical Engineering and Informatics*, *10*(4), 2237–2244. https://doi.org/10.11591/eei.v10i4.2893

Ahlawat, H., Aggarwal, N., & Gupta, D. (2025). International Journal of Cognitive Computing in Engineering Automatic Speech Recognition : A survey of deep learning techniques and approaches. *International Journal of Cognitive Computing in Engineering*, *6*(January), 201–237. https://doi.org/10.1016/j.ijcce.2024.12.007

Alashban, A. A., Qamhan, M. A., Meftah, A. H., & Alotaibi, Y. A. (2022). applied sciences Spoken Language Identification System Using Convolutional Recurrent Neural Network. *Applied Sciences*, *12*(18), 9181. https://doi.org/10.3390/app12189181

Ali, S., Tanweer, S., Khalid, S. S., & Rao, N. (2021). Mel Frequency Cepstral Coefficient : A Review. *Conference: Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development*, 1–10. https://doi.org/10.4108/eai.27-2-2020.2303173

Amiri, S. M. H. (2025). Beyond language barriers : Multilingual NLP and voice recognition for global connectivity. *International Journal of Science and Research Archive*, *15*(02), 406–419. https://doi.org/10.2139/ssrn.5254434

Aysa, Z., Ablimit, M., & Hamdulla, A. (2023). applied sciences Multi-Scale Feature Learning for Language Identification of Overlapped Speech. *Applied Sciences*, *13*(7), 4235. https://doi.org/10.3390/app13074235

Bhatti, M. A., & Alzahrani, S. A. (2023). Navigating Linguistic Barriers : Exploring the Experiences of Host National Connectedness Among Multilingual Individuals. *Eurasian Journal of Applied Linguistics*, *9*(3), 96–112. https://doi.org/10.32601/ejal

Datta, G., Joshi, N., & Gupta, K. (2022). Analysis of Automatic Evaluation Metric on Low-Resourced Language: BERTScore vs BLEU Score. In *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-031-20980-2_14

Gondi, S., & Pratap, V. (2021). Performance and Efficiency Evaluation of ASR Inference on the Edge. *Sustainability*, *13*(22), 1–15. https://doi.org/10.3390/su132212392

Gong, Y., Khurana, S., Karlinsky, L., & Glass, J. (2023). Whisper-AT : Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers ESC-50 Class-wise F1-Score. *ArXiv*, 2798–2802. https://doi.org/10.48550/arXiv.2307.03183

Hansen, J. H. L., Bokshi, M., & Khorram, S. (2020). Speech variability : A cross-language study on acoustic variations of speaking versus untrained singing. *The Journal of the Acoustical Society of America*, *148*(2), 829–844. https://doi.org/10.1121/10.0001526

Hollands, S., Blackburn, D., & Christensen, H. (2022). Evaluating the Performance of State-of-the-Art ASR Systems on Non-Native English using Corpora with Extensive Language Background Variation. *Interspeech*, 3958–3962. https://doi.org/10.21437/Interspeech.2022-10433

Kulkarni, S. V., & Pal, S. (2024). A Review on Language-Independent Search on Speech and its Applications. *IEEE Access*, *12*, 194182–194202. https://doi.org/10.1109/ACCESS.2024.3520394

Mandal, A., Pal, S., Dutta, I., Bhattacharya, M., & Naskar, S. K. (2025). Is Attention always needed ? A case study on language identification from speech. *Natural Language Processing*, *31*(2), 250–276. https://doi.org/10.1017/nlp.2024.22

Palivela, H., Narvekar, M., Asirvatham, D., Bhusan, S., Member, S., & Agarwal, U. (2025). Code-Switching ASR for Low-Resource Indic Languages : A Hindi-Marathi Case Study. *IEEE Access*, *13*, 9171–9198.

https://doi.org/10.1109/ACCESS.2025.3527745

Senapati, C., & Roy, U. (2025). Multilingual ASR Model for Kudmali Voice Recognition. *International Journal of Computer Applications*, *186*(64), 27–35. https://doi.org/10.5120/ijca2025924462

Sharrab, Y. O., Attar, H., Eljinini, M. A. H., & Al-omary, Y. (2025). Advancements in Speech Recognition : A Systematic Review of Deep Learning Transformer Models , Trends , Innovations , and Future Directions. *IEEE Access*, *13*, 46925–46940. https://doi.org/10.1109/ACCESS.2025.3550855

Shaughnessy, D. O. (2025). Spoken language identification : An overview of past and present research trends. *Speech Communication*, *167*(November 2023), 103167. https://doi.org/10.1016/j.specom.2024.103167

Singh, G., Sharma, S., Kumar, V., Kaur, M., Baz, M., & Masud, M. (2021). Spoken Language Identification Using Deep Learning. *Computational Intelligence and Neuroscience*, 12. https://doi.org/10.1155/2021/5123671

Xu, H. (2024). Improving English Speech Recognition System Accuracy Using Machine Learning. *ACM International Conference Proceeding Series*, 73–78. https://doi.org/10.1145/3703187.3703200

Yadav, A., Raj, A., Anand, S., Kumar, V., & Kumar, A. (2024). Deep Audio Classifier : An Artificial Neural Network Approach. *Soft Computing Fusion with Applications*, *1*(2), 103–112. https://doi.org/10.22105/scfa.v1i2.35

Yousif, S. T., & Mahmmod, B. M. (2025). Speech Enhancement Algorithms : A Systematic Literature Review. *Algorithms*, *18*(5), 272. https://doi.org/10.3390/a18050272

Zaman, K., Sah, M., Direkoglu, C., & Unoki, M. (2023). A Survey of Audio Classification Using Deep Learning. *IEEE Access*, *11*(September), 106620–106649. https://doi.org/10.1109/ACCESS.2023.3318015

Zayyanu, M., & Ahmed, U. (2024). Bridging Linguistic Divides : The Impact of AI-powered Translation Systems on Communication Equity and Inclusion. *Journal of Translation and Language Studies*, *5*(2), 20–30. https://doi.org/10.48185/jtls.v5i2.1065

Zhao, H., Chen, H., Yang, F. A. N., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for Large Language Models : A Survey. *ACM Transactions on Intelligent Systems and Technology*, *15*(2), 1–38. https://doi.org/10.1145/3639372